VOLUME 2 □ NUMBER 5 MAY 1965

IEEE spectrum

features

Departments: *please turn to the next page*

the cover

This month's cover represents a portion of a typical 32-level binary-code raster, commonly employed in pulse code modulation systems. An article describing the PCM's history, with predictions for the future, appears on pages 56 to 63.

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

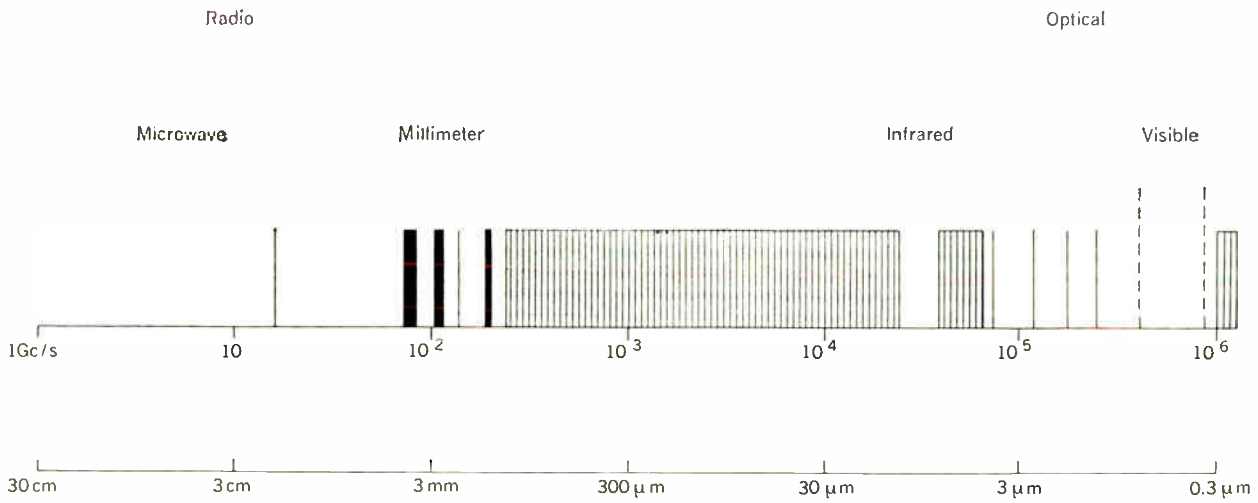


Fig. 1. Atmospheric transmission windows.

mension, $\eta \sim \lambda^{-2}$; hence very short or optical wavelengths seem preferable. In contrast, (2) shows that for a given range and efficiency, $\theta \sim \lambda^{1/2}$.

The angular tolerance in pointing the beam, therefore, becomes extremely critical. Here it should be noted that only heterodyne optical detectors are sensitive to the phase of the received signal. Where phase coherence is not required, a signal collector with relatively wide acceptance angle can be used. The critical pointing tolerance then applies only to the transmitting aperture.

Effective free-space transmission evidently requires increasingly precise position information at shorter wavelengths. In addition, for heterodyne detection, the velocity must be known accurately to compensate for Doppler frequency shifts, which are inversely proportional to the wavelength.

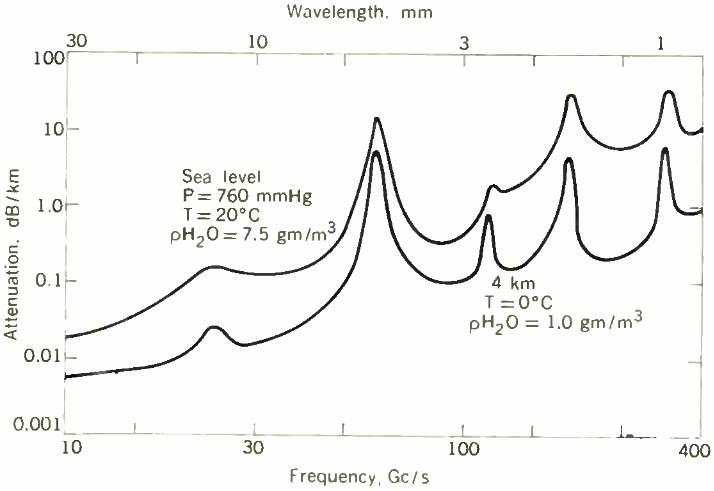
Sensitivity to angle and velocity coordinates is a burden for communications, particularly when one of the terminals is a rapidly moving vehicle such as an aircraft or satellite. Before communication can take place, the antenna must be precisely pointed and the frequency set to correct for Doppler shift. In essence, communication then requires tracking in angle and velocity. For radar and navigation, such tracking is a primary function rather than an added complication. However, even for this application, the sensitivity to angle and velocity coordinates must be kept within limits. In both cases, the acquisition time is proportional to the number of resolution cells to be searched. When there are too many such cells, it may not always be possible to find the correct pair in the time available.

For unguided transmission properties, the wavelength serves as the scale factor in both the solid angle and relative velocity domains. The desirable or tolerable precision in these quantities, as opposed to practical limits on aperture dimensions, determines the best choice of wavelength from a transmission point of view.

In the atmosphere, attenuation due to molecular absorption limits the spectrum available for transmission (Fig. 1). There are windows between oxygen and water vapor absorption lines at the radio end (Fig. 2), but these fill in and become insignificant for wavelengths below 1 mm. This periodic behavior is similar to the absorption characteristics of the atmosphere in the infrared end of the optical region. In both cases, good transmission takes place in several windows at the edge of the broad absorption region between 1 mm and 14 μm. There are approximately 150 water vapor absorption lines in the region between 1 mm and 50 μm whose overlapping wings contribute to the continuous absorption region. For signals passing through the entire atmosphere at near-zenith angles, the attenuation can be surprisingly low. In the windows shown in Fig. 2, this total vertical attenuation is equivalent to some 10 km of horizontal path at 4-km altitude. At 3.2 mm this corresponds to a 1-dB total absorption loss.

The effects of scattering from water droplets are superposed on the absorption. Here, the ratio of particle size to wavelength is the critical parameter in determining the scattered energy. To prevent total extinction over moderate path lengths, the particle diameters must be

Fig. 2. Horizontal propagation through the atmosphere at millimeter wavelengths, after Rosenblum.



much less than the wavelength. In this Rayleigh region, the scattered energy is proportional to λ^{-4}. Typical atmospheric particle sizes are: rain < 1 mm, fog < 10 μm, and aerosols < 1 μm. Therefore, in a qualitative sense, microwaves are relatively immune to rain, millimeter waves to fog, and infrared waves to haze.

Time variations in the attenuation due to absorption and scattering play a secondary role in determining propagation characteristics. Of primary importance are variations in the refractive index, which can cause fluctuations in the angle of arrival of the phase front to be received or partially destroy the coherence of the signal. The severity of this effect depends on the electrical or optical path length, and hence on the wavelength. In addition, the cross section of the beam relative to the dimensions of atmospheric inhomogeneities is an important parameter. The behavior of millimeter and optical beams under various atmospheric conditions is the subject of considerable current investigation. Because of the statistical nature of the problem, definitive conclusions may be long in coming, at least if the somewhat related phenomenon of microwave scatter communications is a reliable indicator. Based on present information, we may draw the general conclusion that millimeter-wave beams have propagation properties intermediate between their microwave and optical counterparts.

Guided transmission

Low-loss modes of propagation in any waveguide must have small coupling to the confining walls or collimating structures. This is necessary because all solid materials, both metallic and dielectric, are much more lossy than appropriate gas fillings. Increasing the cross section in terms of square wavelengths proportionately reduces the energy density at the walls, and hence the loss of any mode. For metal tubes the circular electric mode exhibits the least field intensity at the walls, and hence the lowest loss. In guides using periodic phase-correcting elements or lenses, diffraction losses at the periphery are similarly reduced for larger diameters.

To achieve the lowest loss within a convenient cross section, we naturally select the shortest wavelengths. In this sense, optical guides should excel. In practice, surface roughness and the increased conductor losses detract somewhat from the advantage of optical wavelengths and favor the millimeter region. Where the optimum lies has not yet been determined. However, it certainly is in the direction of wavelengths shorter than microwaves for long-distance guides. Within an enclosed waveguide, a suitable atmosphere can be provided, and hence absorption and scattering losses do not appear at any wavelength.

Generation

Generation of power at radio wavelengths takes place chiefly in free-electron devices. Crossed field or axial-type microwave traveling-wave amplifiers provide high power with stable amplitude and phase. At the optical end of the spectrum, lasers serve as efficient power sources whose potential performance approaches that of their microwave counterparts. Extension of either free-electron or quantum generators to the millimeter region leads to fundamental problems.

The scale factors relating to λ encountered previously reappear in a slightly different manner when we attempt to scale electron beam devices. First, the attainable Q is reduced in proportion to the skin depth, whence $Q \sim \lambda^{1/2}$. The reduced circuit efficiency at shorter wavelengths must be compensated by stronger interaction with the electron beam. This interaction takes place with field components that decay exponentially with the spacing d/λ from the interacting circuit structure. Accordingly, the beam diameter must be reduced in proportion to λ, merely to maintain the interaction strength. Electron guns with beam compression ratios of 100:1 or more are used to produce the requisite coupling factors. With such beams, outputs of over 100-watt average power have been obtained at 3.2 mm in linear traveling-wave amplifiers. The ultimate limit in power rests on our ability to pass very dense electron beams through small orifices in periodic structures without destructive effects. Fortunately, the relatively low interaction achievable at the shortest wavelengths is not necessarily detrimental to overall efficiency: a depressed beam collector can be effective in recovering the energy remaining in the beam.

Few stimulated emission or maser devices have so far been successful in the millimeter region. Since molecular transitions abound with a wide range of energy differences, such amplifiers may be developed for low-noise applications. However, the energy per emitted millimeter-wave photon is so low that high-power outputs cannot be expected.

Detection

The change from classical to quantum limits appears explicitly when we examine detectors. To account for quantum fluctuations in phase and amplitude, the expression for noise power per unit frequency must be modified to the following form:

$$W = \frac{hf}{\exp(hf/kT) - 1} + hf \qquad (3)$$

At radio frequencies, $hf \ll kT$, and the formula reverts to the familiar form, $W = kT$.

For optical frequencies, the converse is true; $hf \gg kT$ and $W = hf$. In convenient units, the ratio of terms is

$$\frac{hf}{kT} = \frac{1}{20}\,\frac{f(\text{Gc/s})}{T(^\circ\text{K})}$$

This places the transition at $\lambda \approx 3$ mm at liquid helium temperature and in the submillimeter region for ambient temperatures.

The maximum transmission range is significantly changed by the altered noise contribution at high frequencies. Thus, Eq. (1) expresses the range as

$$r = (A/\lambda)\eta^{-1/2} \qquad (4)$$

where $\eta = P_{\text{received}}/P_{\text{transmitted}}$.

For radio frequencies, unit signal to noise over a frequency band B yields $\eta = kTB/P_T$ while for the optical region, $\eta = hfB/P_T$. The corresponding maximum ranges are

$$r_{\text{radio}} = (A/\lambda)(P_T/kTB)^{1/2} \qquad (5a)$$

$$r_{\text{optical}} = (A/\lambda^{1/2})(P_T/hcB)^{1/2} \qquad (5b)$$

where $c = f\lambda$. Hence the effective aperture increase at shorter wavelengths is partially cancelled by the higher noise level.

In addition to greater transmission efficiency, higher frequencies offer greater channel capacity. However, the different character of noise contributions effectively raises the minimum power level at which information can be received. The classical information capacity of a channel in bits/second, as given by Shannon, is

$$C = B \log \left(1 + \frac{S}{N}\right) \tag{6}$$

where S, N are signal and noise power, respectively, and the logarithm is to the base 2. The corresponding quantum formulation given by Gordon is

$$C = B \log \left(1 + \frac{S}{N + hfB}\right) + \frac{S + N}{hf} \times \log \left(1 + \frac{hfB}{S + N}\right) - \frac{N}{hf} \log \left(1 + \frac{hfB}{N}\right) \tag{7}$$

A plot of this upper limit of information capacity for a light wave of frequency f is shown in Fig. 3 for a particular noise level and bandwidth. The knee of the curves occurs for $hf/kT \approx 10^3$, or well into the optical region at ambient temperatures. The intrinsic channel capacity shown is further degraded by the spontaneous emission noise in an ideal amplifier. The reduction factor after Gordon is

$$C_a = B \log 1 + \frac{S}{N + KhfB} \tag{8}$$

where S and N are incident signal and noise and K = 1 for an ideal amplifier. As indicated in Eq. (8), even in the absence of external noise N, the signal must exceed a threshold set by the spontaneous emission level, $S > hfB$.

The factor C_a is plotted in Fig. 4. For the 1-Gc/s band assumed, the intrinsic degradation becomes appreciable in the visible and near-infrared regions. The vastly greater bandwidths available in the optical region therefore are useful only at higher signal levels.

In terms of practical amplifiers and detectors, we find that the attainable optical quantum efficiencies and microwave noise temperatures are approaching theoretical limits far more closely than their millimeter-wave counterparts. Microwave noise temperatures in the order of 10°K are available, and photodetectors may approach unit quantum efficiency. The best equivalent noise temperature of millimeter mixers is near 25 000°K, or about four times as hot as the sun at these wavelengths. There is no basic obstacle to closing this performance gap, and new amplifiers are in prospect for doing so. Varactors for parametric amplifiers are being scaled from the microwave region with the aid of modern techniques in microminiaturization. Suitable molecular energy levels for possible maser action also exist.

Comparison

The interplay of the parameters in transmission, generation, and detection permits many choices of optimum operating wavelengths for various applications. To gain some perspective on the relative trade-offs, we make an elementary comparison between three hypothetical communications systems. For current interest, we select a link between a ground terminal and a space vehicle near the moon. For the ground terminals, we choose two of the most advanced antennas in their respective operating ranges: the 120-foot antenna at the M.I.T. Lincoln Laboratory operated at 3.2 cm and the 15-foot antenna at Aerospace Corporation at 3.2 mm. The corresponding beam widths, $\theta - \lambda/d$, are about 1 milliradian (3 arc minutes) and both systems have absolute pointing accuracies of about one tenth of their beam width. For the optical system we take a 3-inch telescope at 0.63-μm wavelength, giving a theoretical beam width of about 0.01 milliradian (2 arc seconds). We assume an attainable pointing accuracy of about 0.05 milliradian (10 arc seconds) and a proportionately broader beam for practical transmission purposes.

In the vehicle, we assume a 5-foot maximum antenna aperture and a 0.15-milliradian (30 arc seconds) pointing

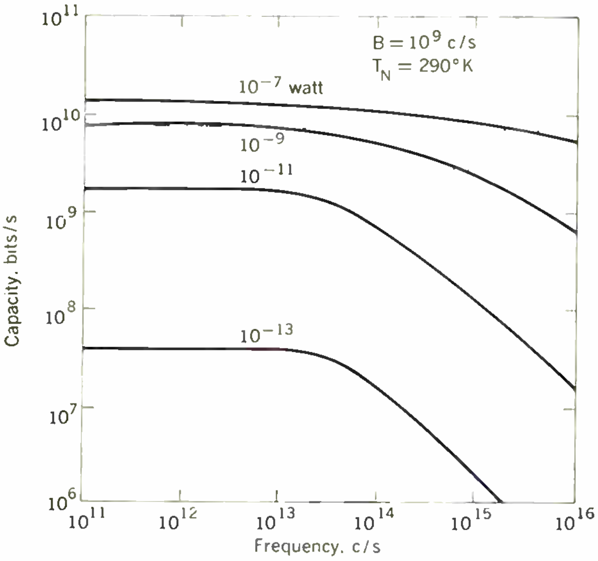
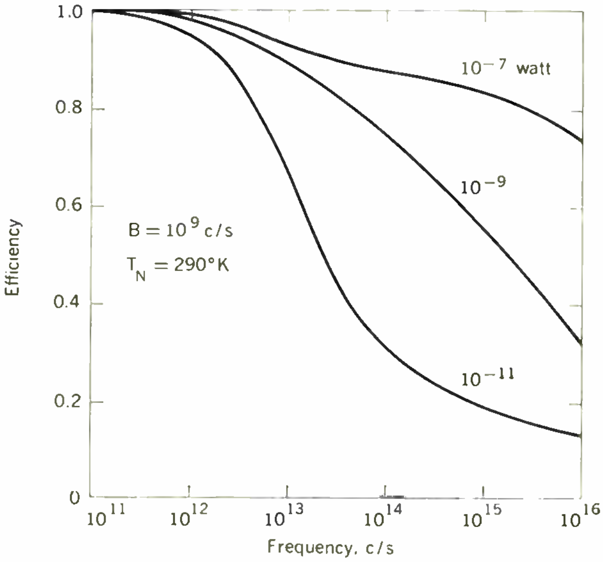


Fig. 3. Information capacity of an electromagnetic wave in a single mode as given by Gordon.

Fig. 4. Information efficiency of an ideal amplifier as limited by hfB, as given by Gordon.



accuracy derived from star tracking. The usable optical aperture is reduced by the beam width required for this pointing tolerance. A collector for envelope detection is not subject to this limitation and so is discussed separately. The resulting theoretical transmission efficiencies for 4 × 10^5 km are

	3.2 cm	3.2 mm	0.63 μm
$\eta = (\lambda/r)^2(\theta_1^{-2}\,\theta_2^{-2}) =$	1.9×10^{-11}	2×10^{-11}	6×10^{-14}

The two radio systems exhibit similar transmission characteristics. In contrast, the optical system requires orders of magnitude improvement in pointing. Propagation anomalies accentuate the pointing problem by their defocusing action. At 3.2 mm, this effect slightly broadens the beam of the 15-foot aperture at low elevation angles. For horizontal optical paths in the atmosphere, observations show much greater beam broadening as well as fluctuating deflections, often called "dancing," which effectively spread the beam over many times its theoretical width. Even for near-vertical paths, so-called "seeing" disturbances broaden star images to as much as 0.5 milliradian (10 arc seconds) in daytime. At night, "seeing" is better, but image broadening is still observed.

The power and sensitivity factors are difficult to assign without prejudice. We assume that an S/N ratio level of 10 dB is required over a 10-Mc/s acceptance band, which, in terms of Eq. (6), corresponds to 3.5 × 10^7 bits/second. Receiver noise temperatures of 30°K at 3.2 cm and 300°K at 3.2 mm appear to be reasonable performance limits. The necessary millimeter-wave low-noise amplifiers are still under development but are feasible in principle. The higher noise temperature expected at millimeter wavelength is due in part to atmospheric emission at ambient temperature. This emission is the inverse process to absorption, and is proportional to the air temperature and the absorption coefficient. For a coherent optical receiver, we take the signal level at 10 dB above hfB (B = 10^7). This gives us about the same channel capacity in bits/second for an ideal amplifier. The relatively small optical atmospheric absorption in clear air is neglected, as is the inevitable deviation from ideal amplifier performance. The required transmitter power for the system then becomes

	3.2 cm	3.2 mm	0.63 μm
$P_T = P_r/\eta$	2×10^{-3} W	2×10^{-2} W	5×10^2
P_T maximum available	10^5 W	10^2 W	1 W

The performance margins are evidently excellent at microwaves and adequate in future millimeter-wave systems, even if considerable atmospheric absorption is encountered. The optical performance factors appear least favorable, despite the ideal receiving amplifier assumed. Actually, the heterodyne optical receiver chosen for direct comparison purposes is less effective in the example than an envelope detector. In such a detector, the intensity or number of photons determines the output. These tolerances in the collecting aperture are therefore not critical, provided the incident energy is collimated on the sensitive area of the detector. A separate collector within the five-foot limit specified yields a transmission efficiency approaching that of the longer wave systems. The relatively wide acceptance angle of the collector, as compared with a precision reflector, is admissible where angular resolution is not required. The use of envelope detection also eliminates the troublesome Doppler correction. The resulting optical system is competitive with millimeter waves, except for weather effects. This has been achieved by avoiding linear scaling, and instead taking advantage of the high quantum efficiency of direct detection devices.

Conclusion

In choosing an operating wavelength, the designer of a radiating system is confronted with balancing the basic scale factor. Long range, high efficiency, and large antennas appear on the radio end of the spectrum, while extreme resolution and compact size are at the optical end. The high quantum efficiency of direct detection devices in the optical region makes possible receiving systems with less critical tolerances. Beam width and Doppler scale factors do not apply directly to such incoherent receivers. For many applications, the evolutionary improvement possible at millimeter wavelengths more nearly meets present needs than the quantum jump to optics. There are many inducements for system designers to consider such an alternative. As we have indicated, narrower, but not excessively critical, beam widths are available. Alternatively, smaller antenna apertures may be used. Broad and clear channels permit high information capacity with standard techniques, as compared with the more critical optical methods. For fair-weather or above-the-weather applications, high-resolution radars, communication links, navigation aids, and other systems now in the microwave spectrum may find a home in the millimeter region. Even within the restricted windows permitting propagation, the bandwidth available in the millimeter region exceeds that of the entire radio spectrum below. We may expect with some confidence that so much unused but potentially useful channel capacity will not long remain idle. The imminent improvements in millimeter power generation should provide the stimulus to rapid growth.

On the other hand, for guided or short-range applications, the optical scale factors appear to advantage, and waveguide technology shows signs of migrating directly to the optical bands. Certainly, the millimeter region can never become a dominant portion of the entire coherent electromagnetic spectrum, as microwaves have been. That significant use will soon be made of millimeter waves seems equally certain. The blending of classical and quantum techniques required promises an interesting convolution of old and new technology.

REFERENCES

1. Gordon, J. P., "Quantum Effects in Communication Systems," *Proc. IEEE*, vol. 50, Sept. 1962, pp. 1898–1908.

2. Megler, G. K., "Some New Aspects of Laser Communications," *Appl. Opt.*, vol. 2, March 1963, pp. 311 315.

3. Rosenblum, E. S., "Atmospheric Absorption of 10–400 K Mcps Radiation," *Microwave J.*, March 1961, pp. 91–96.

4. DuWaldt, B. J., and Hoffman, L. A., "Research on the Suitability of Millimeter Wavelength Systems for Space Applications," presented at WESCON Convention, San Francisco, Calif., Aug. 1964.

5. Forster, D. C., "Mid-1964 Review of Available Millimeter-Wave Sources," presented at WESCON Convention, 1964.

6. Johnson, C. M., "Injection-Laser Systems for Communications and Tracking," *Electronics*, Dec. 13, 1963, pp. 34–39.

7. Straiton, A. W., and Tolbert, C. W., "Millimeters—The New Radio Frontier," *Microwave J.*, Jan. 1960, pp. 37–39.

From control engineering to control science

Primitive and empiric up to the 1940s, control engineering is now entrenching itself in the domain of mathematical theory. Engineers, however, should continue to concentrate on their prime role—the application of theory to physical reality—rather than seek to be mathematicians

J. E. Gibson *Purdue University*

The young apprentice in control engineering today is learned in vector–matrix functions with differential side conditions. He is comfortable with the union of two convex sets and support hyperplanes and can discuss the variations on the problem of Bolza, which apply to the consideration of constrained state variables. But it sometimes appears that he lives in fear of being trapped—alone at night in the laboratory—without a matrix to invert or even a Jordan normal form for protection.

In this informal narrative, we will review for the nonspecialist in control how the control engineer arrived at this interesting condition. This is not an attempt at an annotated bibliography of the field, and the specialist in control will find few, if any, of the details he would require in a survey. Rather, this will be an attempt to classify trends and speculate on the future direction of automatic control. Inasmuch as this is a field in which almost all those who have ever worked are still active, these judgments, necessarily of a somewhat subjective nature, will draw a wide variety of reactions.

In the development of control engineering, four areas can be differentiated: primitive, classical, modern, and future control. In this breakdown, the time periods overlap somewhat but—and this perhaps illustrates the imitative, derivative nature of "control scientists"—all categories seem to be doing the same thing at the same time.

Primitive control

Prior to the 1940s, the field of control was characterized by a primitive, empiric nature. Isolated, self-regulating mechanisms such as the flyball governor for the steam engine and devices for orienting windmills were developed as long ago as the 18th century, but by the 1930s, with Black's patent on the feedback amplifier and the interest of chemical engineers in process control, a definite movement and body of technology may be discerned. An explanation for the operation of electronic oscillators and the regenerative receiver was needed and the value of the concept of positive feedback came to be appreciated. Tales are told, however, of the resistance encountered by Black when he proposed self-regulation, or negative feedback. The idea of sacrificing gain for stability of operation in the face of parameter drift was difficult to sell. Parenthetically, we might note that when the concept of positive feedback finally caught on, it was overworked and indeed improperly used, if early texts in the field are a reliable guide.

A few giants, such as Nyquist, Bode, and Minorsky, were working in the 1930s to lay the foundation for the science of control, but years passed before their work was appreciated by the average engineer. Minorsky's application of the mathematics of nonlinear mechanics to control systems was equally germinal but even less widely appreciated. In the early 1940s, Minorsky was responsible for bringing to the attention of a western audience the work of Andronow and Chaikin, in his David Taylor Model Basin report. In fact, this event helped to mark the opening of the classical era rather than being considered a part of the era we are now discussing.

Classical control

There seems to be almost unanimous agreement that a new era in control was opened by the work in the early 1940s at the M.I.T. Radiation Laboratories (documented by vol. 25 of the *Radiation Laboratories Series* by H. James, N. Nichols, and R. Phillips). This amazing book might be compared to the opening theme in a great symphony. It sounded the notes that have been echoing ever since. It is interesting that, exactly like educators in the arts and the classics, teachers in engineering have felt the need to reinterpret, explain, and water down James *et al.*, rather than to refer their students to the original source material, usually to the tune of: "Well, of course, it's a great book; I use it as a reference all the time. But you couldn't use it as a *text*."

In the classic period, system analysis was based on theory. The theory was based upon the solid rock of linearity. The frequency domain (James *et al.*, chap. 2) and the Laplace transform (James *et al.*, chap. 2) were used almost exclusively.

The frequency domain approach dominated the field of control through the 1950s. If a guess were to be made, we would say that more than 90 per cent of the systems in use today were designed with these concepts. While in some sense they were primitive and approximate, at the same time they were very practical. The frequency domain techniques are a complete, unified package. They constitute a "full line" because they allow the reduction of sinusoidal response data, taken from the fixed elements of the system, to manageable form by means of one or another of the graphical display methods. Next, we select the best of a number of controller designs worked out on paper, then build this element in the laboratory out of physically realizable components. When this controller is placed in the system, in an almost magical manner the system performance is improved by orders of magnitude.

One recalls some of the systems that were improperly designed in the early 1940s, such as the shipboard stable element, which was as big as two filing cabinets and reputedly cost more than $1 million each. If the gimbal system were slightly disturbed, the element would oscillate for 15 or 20 minutes because the system was on the verge of stability. The device was installed and used, however, because it was the best we could do. Yet, by the late 1940s a young student with three or four months of training could go into the laboratory, take a few open-loop frequency responses on this device, and design a simple network, worth perhaps 15 cents, consisting of two resistors and two capacitors. When the unit was turned on, it performed orders of magnitude better. This was an amazing performance.

James *et al.* also introduced the discussion of the effect of random noise on system performance. Their work was based on Wiener filtering theory and the rms error criterion. Control engineers have endeavored for 15 years to apply this approach to meaningful control system design but in the main the effort has failed. We can do a few trivial tricks with additive Gaussian noise, and every textbook author since Wiener has felt obliged to mention the subject, but control system design, unlike filter design and communication system design, has not been basically influenced by random noise theory. Of course, every Ph.D. candidate still is questioned on the analysis of his system in the presence of noise, and discussions in journals always comment on the fact that the author "considered only the trivial, noisefree case." But this is not to be taken seriously since the bare fact is that a meaningful noise analysis cannot be made; moreover, meaningful experimental design data cannot be obtained.

Through most of the 1940s M.I.T. maintained undisputed academic leadership in control under the direction of Gordon Brown. Brown and Campbell wrote the first widely accepted textbook on servomechanisms, a book that has been rewritten by different authors at least two dozen times since its publication in 1948. Later in the early 1950s, a supernova flared at Columbia under the brilliant drive of Ragazzini and Zadeh. They gathered several dozen outstanding students, and while the group contributed to a number of specialities, their major effort was in linear sampled-data systems. Chapter 5 of James *et al.* set the stage; the problem was introduced and the z transform, a special case of the Laplace transform, was developed by Hurewicz. But it remained for the group under Ragazzini and Zadeh to develop in a period of not more than five years the tremendous body of material that now exists on such systems. Even today the true believer or his academic descendent holds that a sampled system is simply a special case of continuous control. During the classical control period, some attention began to be paid to nonlinear systems and perhaps the describing function work by Kochenburger and bang-bang control work by Hopkins could be cited as two examples of this interest. But, generally speaking, the early 1950s was a period of broadening of interest in the linear frequency domain approach from a few leaders to the great mass of engineers and engineering educators throughout the country. It was felt by a number of academic leaders that control had now matured and that it was time for them to find something else to lead. In retrospect, this perhaps seems analogous to the 1920s in physics just before the quantum theory revolution.

We had a perfectly acceptable theory so long as we confined ourselves to linear systems and conventional performance criteria. We ignored the tough problems. We ignored nonlinear systems, systems with parameters that vary with time, systems which by their nature prevent complete a priori knowledge of their behavior, systems with many inputs and many outputs, systems that work in conjunction with human operators, etc. In short, we ignored much of the real world.

Modern control

While prenatal indications were perhaps discernible, it is convenient to date the birth of modern control in the United States as the first Joint Automatic Control Conference at Dallas in 1959. At this conference Kalman and Bertram gave their classic introduction to the rigorous discussion of stability of nonlinear systems by the second method of Liapunov. The concept of adaptive control was vigorously discussed. The state variable formulation of system equations was appearing in volume and the concept of optimum control loomed large on the horizon. Newton, Gould, and Kaiser had published their approach to analytic design in 1957, and by this time Bellman had suggested design criteria that separated the cost of error from the cost of correction. Kazda had used the state variable approach in 1954 and Hopkin had discussed time optimal control in 1953. As far back as the middle 1940s, McColl, Weiss, Goldfarb, Tustin, Opplet, and Kochenburger had used various approximation methods, such as harmonic balance, to analyze ON-OFF or relay control systems. In fact, in Hazen's classic paper in 1934 the ON-OFF controller was discussed. The tendency always exists to give history an organization and logic it does not really possess, but perhaps this serves a purpose in helping us to at least attempt the organization of our future. Thus, with this apology, and in spite of prior work, we will date modern control from 1959. By this date, it was obvious to almost all the leaders in the field that classical control was not enough.

The early workers in modern control had by now convinced the field that conventional engineering design by repeated trials until all system specifications were met could be replaced by exact optimization, provided the problem is properly put. The first type of optimal control discussed was time optimal control. The problem is simply stated: What form of input should be applied to a given plant in order to cause it to adjust its initial output condition to a desired output condition in minimum time? For the problem to be meaningful, certain constraints upon the acceptable control signal must be assumed, of course. For an engineer, the question that immediately follows is: "And how does one construct such an optimal controller?" From physical arguments it was made apparent that an optimum control built to these specifications will be bang-bang; that is, maximum control effort of various polarities will be used. The example most often cited to illustrate this is the "hot rod" problem. Assume that a hot rod exists and is unique in Euclidean 3-space and that the initial and terminal manifold are finite segments of a one-dimensional surface fixed with respect to an arbitrary basis which spans the whole space. In other words, the rod is sitting at the white line as the light turns green. The object is to get to the next red light in minimum time. The operator, who has received careful instruction in optimization theory by peer group interaction, immediately applies maximum torque, burns rubber, and accelerates in maximum fashion to a carefully chosen point, part way down the block. At this point, he instantaneously applies maximum braking torque. The car skids to a halt exactly at the next white line. He has reduced his error in position and velocity to zero simultaneously and in minimum time. (It may be noted that this solution involving a single switching is optimum only for a second-order system, in this case a pure mass. The error state variables of another mass nonrigidly coupled to the mass under propulsion may undergo excursions at the terminal manifold. In other words, "Hold tight, dad, or you'll bang your head on the dash when she grabs.")

By the Dallas meeting, however, a new theme was being sounded. It started with the realization by a number of mathematicians, both in the Soviet Union and in the United States, that engineers were working on some very interesting problems involving extremization of functions with mixed boundary conditions and differential side conditions. The so-called "automatic control problem" was soon thereafter rigorously stated (by mathematicians) in a form understandable to the mathematician. It gradually became apparent that this problem is a very complex problem in the calculus of variations called the problem of Bolza. It appears that we are now living through a fertile moment in scientific history, one in which mathematics draws nourishment from its roots in physical reality and returns to a problem bypassed previously because of its difficulty and lack of motivation. The dialogue between mathematics and systems engineering has been established to the mutual benefit of both fields. This is a heady experience for a number of young, scientifically oriented engineers. To work on the same problems as Bellman, Pontriagin, and Lefschetz is an exciting experience, and unfortunately a number have lacked the maturity to live through it unscathed. In effect, this is a problem in identity. The young, scientifically trained engineer must have a clear picture of himself and his professional function. He must face nature with the tools at hand. If a dialogue with mathematicians will help provide better tools, well and good. But the engineer must not be an egotist! He cannot be both engineer and mathematician. We have all seen examples of papers published in our professional journals that are unintelligible to any engineer. It is sad to relate that quite often when the confused engineer takes this sort of thing to a mathematician, he is told that it is nonsense to the mathematician as well! The control engineer must stop trying to supplant the mathematician. It is not the control engineer's job to generate theorems, lemmas, and proofs; it is sufficient to understand and apply them. The engineer supplies the motivation, the physical constraints, and the statement of the problem, and, most important, the interpretation of the results. Papers that illustrate solutions "in principle" or "the form of solution," or which show a black box labeled "7094" that is to furnish solutions, do not provide engineering solutions to engineering problems.

This is a dangerous doctrine. There will be those who disagree with me on the grounds that the mathematical theory must come first. This is an interesting intellectual argument, though difficult to prove by example. Nevertheless, I will accept it, *provided* we then move on to reality. There are, however, few signs of this in the literature.

The future of control engineering

It is interesting to note some stirrings of doubt among theorists that our direction for development of control engineering is correct. Apparently a number of other popular fields are serving to attract the bright young men who might otherwise work in control engineering. Our problems seem to have lost some of their luster. The prognosis is correct, but the prescription is not. "More of the same" is apparently the thought for the day: if our young engineers want to build things, give them mathematics; if they are confused about the reality of engineering, give them mathematics; and if they do not understand that an engineer's job is to build something that works regardless of whether the theory is complete or not, give them mathematics.

There is a great need now—and for at least the next five years—to learn how to apply some of the new theory. This means that engineers should concentrate on computational algorithms and means of solving some of the equations that modern theory shows us how to write. The development of such algorithms is no trivial task. The present methods of attack, such as dynamic programming or steep descent solution using the calculus of variations or the maximum principle, quickly swamp present-day computers. Modern optimum-control theory is critically dependent on the development of a meaningful index of performance. The theoretical methods all assume that such an index, or "measure of goodness of a solution," is given. It is an engineering problem to establish the relation between physically meaningful criteria and mathematical criteria meaningful in terms of theory.

But what of the future? Will control engineering for the next five years follow only those paths that are already well defined? My feeling is that if we restrict our attention to those problems that are "well defined" from the point of view of the mathematician, we are engaging in a process of self-sterilization. We should emphasize the newer areas, such as the study of self-organizing and learning systems, and move into the grey areas, such as biological control. We should attempt to generalize our field into general systems studies from an engineering point of view, not merely from a mathematical point of view. We must be ready to stumble, and even fall occasionally. For example, is it possible to formulate our foreign aid policy on a feedback control basis? We hear that South America is poor and that if we pour more money into it we will improve conditions. No control engineer would make this mistake. We know that an infinite source of energy (money) does not automatically guarantee stability in a physical system. It is not apparent that engineers have anything significant to say about such global problems, but we should try.

This article is an informal discussion on the development of automatic control over the past 20 years and is based on a talk given at the 16th Annual National Aerospace Electronics Conference, Dayton, Ohio, May 11–13, 1964.

Computer-controlled power systems

Part II—Area controls and load dispatch

All power systems must maintain a balance between the generation of electricity and the constantly varying consumer load, and obtain the maximum generating efficiency at a minimum possible cost. System-wide automation will greatly assist in optimizing this effort

Gordon D. Friedlander *Staff Writer*

It is not often possible to combine in one article the broad SPECTRUM, represented by power engineering and computer technology, that is implied by the name of this publication. But in a "marriage of the arts"—computer controls to power generation and distribution—we have a topic that should be of interest to our diverse readership.

In Part I of this series, boiler–turbine unit controls were described in considerable detail. In this installment we shall endeavor to explore the much broader area of the application of computer controls to the entire power system of a single utility company's operations, and also to interconnected grids that may cover a geographical area of several states.

Essentially, the prime functions of a computer used for power system automation are similar to those that are used for plant automation, but there is a considerable shift in program and operational emphasis. For example, in systemic operation, the computer is more loosely integrated into the system. There are fewer inputs and outputs, and the real-time requirements are less exacting.

Some historical background

Around 1940[1] utility companies were applying tie-line load control to their systems in an attempt to improve their performance efficiency through the maintenance of frequency and tie-line schedules. Even at that date, the term "economic dispatch" was used, and several utilities tried to reduce their fuel costs by frequent updating of the generator load setters.

About 15 years ago, analog equipment became available that could load units according to their incremental cost of generating power. Following this period in time, the state of the art advanced rapidly with many improvements in analog control systems. And, as breakthroughs were achieved, the control systems became more sophisticated, and included provisions for the evaluation of transmission losses, energy and time-zone interchange, etc. Thus, as a natural evolution, it became necessary to develop more reliable hardware.

To evaluate properly the various automated operating procedures and regulatory techniques, a general comprehension of the fundamental operation philosophy of our domestic power systems is essential.

Philosophy of operation—the overall scheme

In the United States, interconnected power systems[2] achieve their generating capabilities from many independent investor-owned utilities, public power developments, and rural electrification agencies. Participation in the interconnected grids is usually on a voluntary basis, and the overall coordination is attained by mutual and cooperative agreement. Often there is no centralized directorate or administrative nucleus to direct day-to-day operations. Therefore the association of adjacent power groups that comprise the pool is predicated upon informal understandings and procedural arrangements.

The physical geographic extent of the larger interconnections and the magnitude of their capabilities necessitate operational subdivisions, and these smaller areas frequently are served by a single utility interest—although, in some instances, a number of private interests may elect to form a pool by themselves, connected to

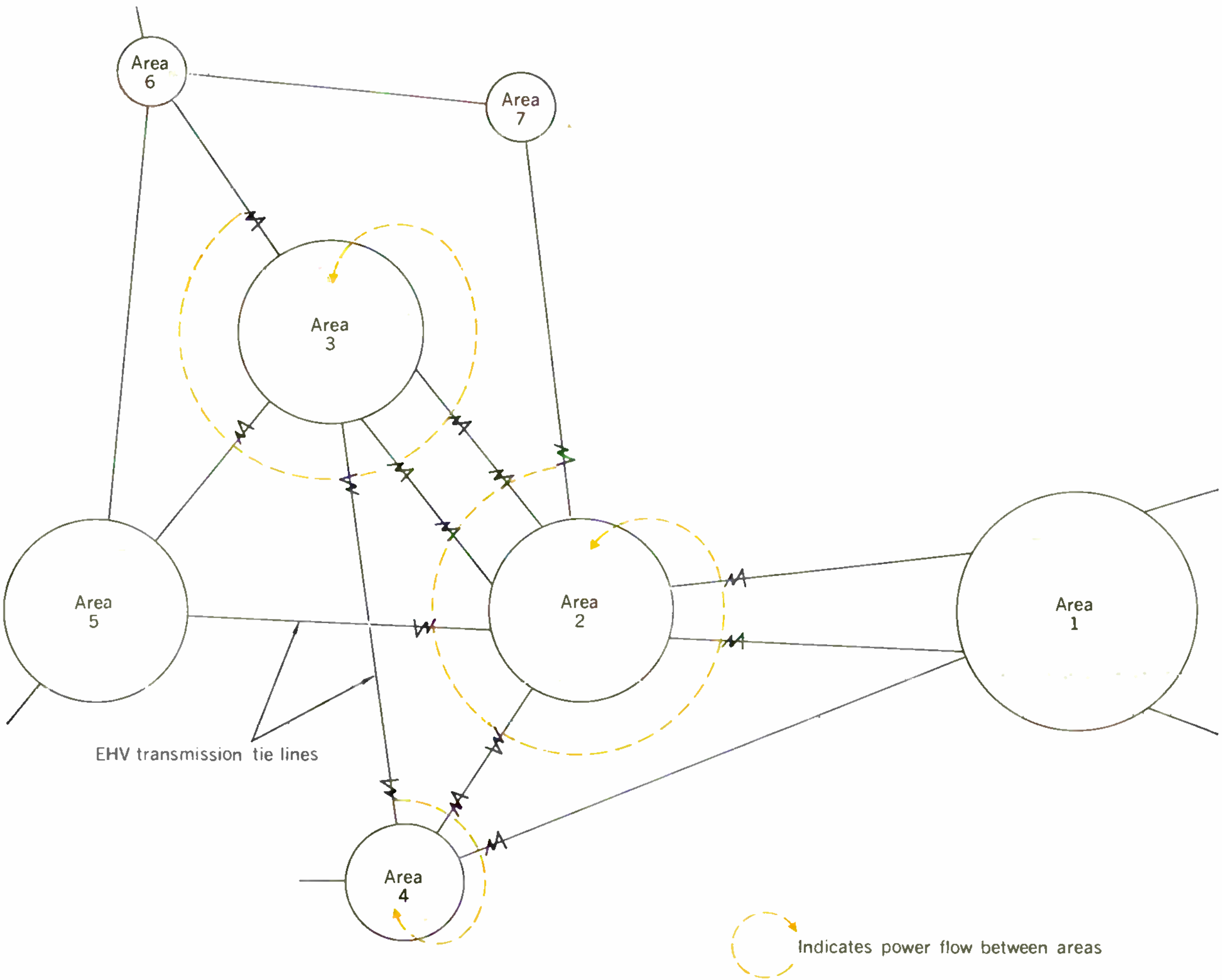

Area 6
Area 7
Area 3
Area 5
Area 2
Area 1
Area 4
EHV transmission tie lines
Indicates power flow between areas

Fig. 1. Schematic diagram showing the principle of area control, which is the accepted practice for all interconnected systems in the United States.

other areas, but with free-flowing ties within the pool area.

The benefits that accrue to each group in the power pool operation are the overriding reason for participation. These advantages include continuity of service, increased capability through the utilization of diversity, improved voltage and frequency regulation, financial advantages of lower overall production costs, deferment of capital expenditures that might be necessary in meeting the increased demand of independent system operation, and associated economies.

Perhaps the fundamental principle for efficient interconnection is in the statement: *each area must be responsible for the absorption of its own load changes within its perimeter.*

The principle of area control

For a power system operating on an isolated basis, it is necessary to keep the generation equal to its load. Thus the changes in the base frequency are the index of efficiency. The cooperative organization of grid systems was the natural reaction to this fundamental fact. If the independent area joins a large interconnection, and continues to keep its area generation equal to its area load, it will not introduce additional operating problems for the grid. If the load changes of that area, however, are corrected by generation changes in another area, there will be interarea power flows that are unsolicited and unscheduled—and unwanted. Therefore, when interconnected, it is essential that each area measure the interchange at all of its tie points with the interconnection, and then compare actual net power flow to determine whether the load changes of the area are being corrected by generation changes within that area. If an area has an excess of capacity, it may schedule a certain net delivery to its neighbors in the interconnection, but this value then becomes part of the base for determining how closely its generation changes are following its area load changes.

Figure 1 illustrates the area control principle, which is the accepted practice on all of the large interconnected systems in the United States. Hypothetical Areas 1 through 7 are indicated in the diagram, with the size of the shaded circle representative of the relative output capability of each.

Extensive telemetering is frequently necessary to obtain the net interchange of an area with the rest of the interconnection. The effective utilization of this information is obtained by its collation in a system operator's central office for each area where the area control requirement is acquired in accordance with the general equation

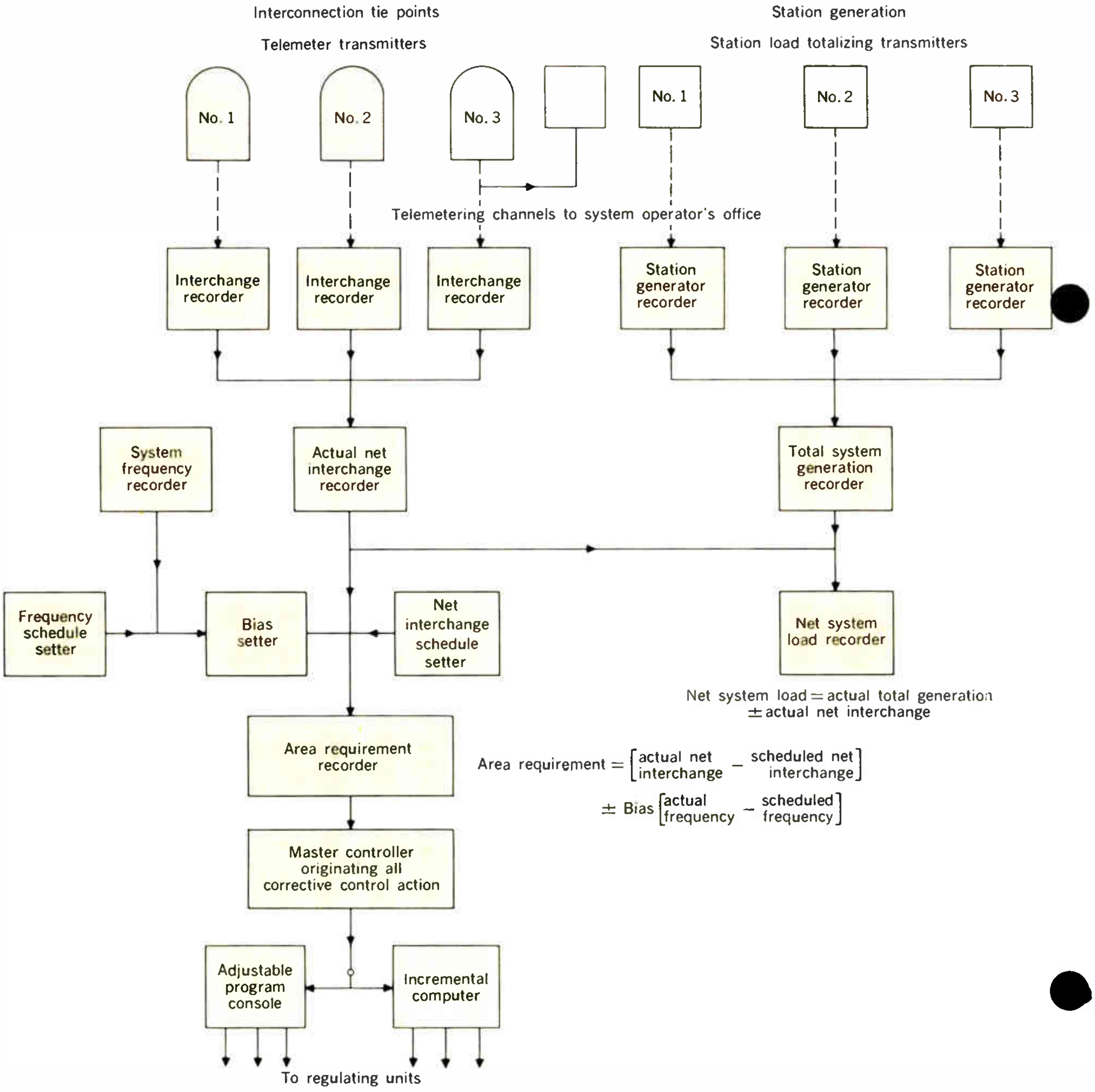
Area Requirement (Area Control Error) = (Actual Net Interchange − Schedule Net Interchange) ± Bias (Actual Frequency − Scheduled Base Frequency)

Note that from the equation the determination of area requirement includes tie-line bias control, which recognizes the inherent regulating characteristics, or bias, expressed as megawatts per 0.1 c/s of frequency change. It is experimentally determined for each area, and varies from one area to another, dependent on the type of lead, inertia constants, prime-mover governor action, etc. All values for systems in the United States have been within the range of 0.5 to 2 per cent of the capacity on the line in the area. For a given system, this bias characteristic provides an essential part of the reference, which determines whether any load change was within a local area or in some other area.

As indicated in Fig. 2, this area requirement (or area control error) is continuously computed on-line, and it is graphically plotted before the system operator. The total generation increase or decrease that should be made on the system to meet its fundamental load commitment is shown directly in megawatts. Figure 3 shows the graphic plot for a typical instantaneous net system load recorder.

The basic determination by each area of its responsibility to the entire network is the price for the benefits derived from an interconnected operation. This is a continuing investment, since communication and telemetering requirements constantly call for hardware updating that

Fig. 2. Block diagram of typical centralized control system. The area requirement is continuously computed on-line, and it is graphically plotted for the system operator.


Interconnection tie points
Telemeter transmitters
Station generation
Station load totalizing transmitters
No. 1
No. 2
No. 3
No. 1
No. 2
No. 3
Telemetering channels to system operator's office
Interchange recorder
Interchange recorder
Interchange recorder
Station generator recorder
Station generator recorder
Station generator recorder
System frequency recorder
Actual net interchange recorder
Total system generation recorder
Frequency schedule setter
Bias setter
Net interchange schedule setter
Net system load recorder
Net system load = actual total generation ± actual net interchange
Area requirement recorder
Area requirement = [actual net interchange − scheduled net interchange] ± Bias [actual frequency − scheduled frequency]
Master controller originating all corrective control action
Adjustable program console
Incremental computer
To regulating units

will give the greatest speed of response, accuracy, and reliability. When the improved performance of new equipment in this field has been demonstrated by operating experience, it often indicates the obsolescence of existing equipment.

Analysis of area regulation

The fundamental principle for area efficiency is found in the statement: *the proximity by which each area maintains its area requirement at zero is a measure of its operating competence as a participating member of the interconnection.*

When the necessary increase or decrease in megawatts to return the area requirement to zero has been determined, the next problem is to decide where this change in generation should be made within the area. The essential regulating requirements are

1. The provision of sufficient regulating range.
2. The provision of a sufficient regulating rate.
3. The provision of an incremental rate which will match the system average over the proposed regulating range.

An area that has predominantly steam-electric generation may find that regulating requirements 1 and 2 (range and rate) can be readily met by its newer and larger gas-fired steam units, but that the area's incremental rate, under 3, is such that it is not economical to do so. This determination accepts the theory that the highest economy is obtained when all units on the bus are loaded to operate at the same incremental rate or cost in mills per kilowatthour.

For example, in the early morning hours when the load is coming on a system, it may be economical to regulate on a large unit that has an incremental cost of 2 mills/kWh, because other large units, with a similar incremental cost, are being loaded. It would not be economical, however, to provide continuous regulation on this 2-mill unit just to utilize its range and response rate capabilities after all the other units of 2-mill capability have been loaded, and it is necessary to proceed to 4- or 5-mill generation to carry the area load.

It is generally recognized that regulating requirements 1 and 2 for power systems in the United States are responsibilities of the area to the interconnection, regardless of the means taken by the area to meet them. If a single unit or plant fails to provide the required regulating range and rate at all times, then multiple unit or plant regulation must be employed simultaneously by the area. The simultaneous regulation of units involves regulation requirement 3, which relates to the relative loading of units, and therefore introduces incremental loading determinations and the internal economical operation of plants within the area.

Individual system needs. To determine individual system needs, it is necessary to analyze further the three essential regulation requirements just discussed. In doing so, we must consider the *actual continuous system load curve*, *unit incremental curves*, and *economy loading curves*, and their correlation for a given system.

Figure 3 shows the actual net system load curve for a typical area as continuously recorded from telemetered readings. Note that the curve is *not* determined from plotting hourly readings, but rather that it represents instantaneous on-line readings.

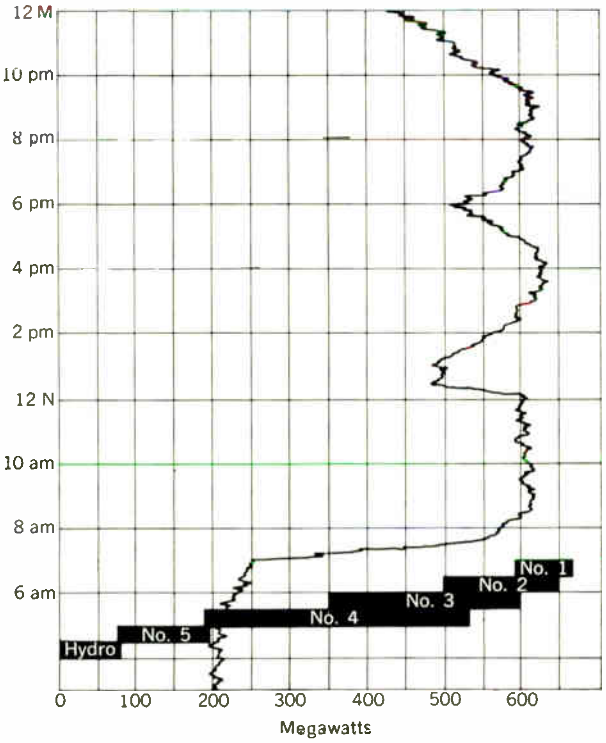


Fig. 3. Actual graphic plot of a typical instantaneous net system load curve and incremental loading program.

Fig. 4. Graph showing typical incremental cost curves which, for simplicity of illustration, are assumed to represent single-unit power stations.

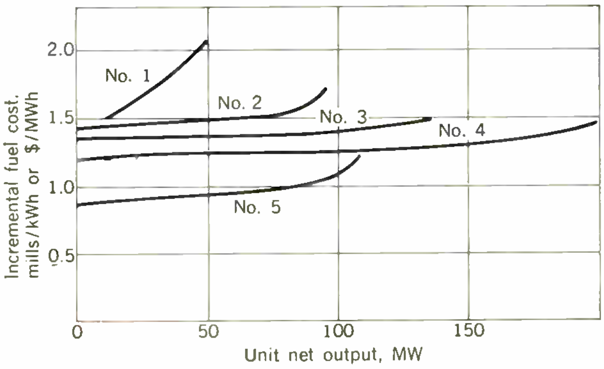


Figure 4 indicates typical incremental generation cost curves for units which, for simplicity of illustration, are assumed to represent single-unit stations. Most utility companies have obtained these data for each of their units. Actual operating tests are conducted to ascertain the Btu/hr fuel input to the boilers, as compared with the kilowatthour output of the generators, to obtain an incremental heat rate curve. When these values are multiplied by the fuel cost, the incremental cost of the next unit of generation for each value of unit output is obtained. If the Fig. 4 data are then related to the system operation by plotting against total system generation, the Fig. 5 economic dispatch curves are obtained, and these show which units or stations should be carrying any given system load for the lowest generation cost.

Correlation of curves. A correlation of the curves of Figs. 3–5 reveals a number of interesting factors.

First, by referring to Fig. 3, we note that the minimum load of our assumed system is above that of both the run-of-the-river hydro and Station no. 5. Thus it would appear that both would be fully loaded during the entire 24-hour period, and that no automatic regulation should be applied to these installations.

Because it has the next to the lowest incremental cost (Fig. 5), Unit no. 4 starts the pickup as load comes on the system in the morning. The actual rate of system load increase (Fig. 3), as indicated by continuously telemetered total generation readings, may be 10 MW/min. If Unit no. 4 has limitations on rate-of-load change, such as a maximum permissible sustained rate of 2 MW/min, it could not alter its generation as fast as the change in system load, and could not, therefore, meet rate regulation requirement 2. Power would flow in over the tie lines from adjacent areas to supply the load, and the area served by Unit no. 4 would have failed to meet the fundamental requirement for interconnected operation.

It is obvious, under these conditions, that regulation requirement 2 and the incremental loading program cannot be met simultaneously. And since the area requirements have precedence, an economic loading schedule cannot be rigidly followed. Therefore, to obtain the required rate, simultaneous response of other units in addition to no. 4 will be required.

For the assumed program, this means that, instead of

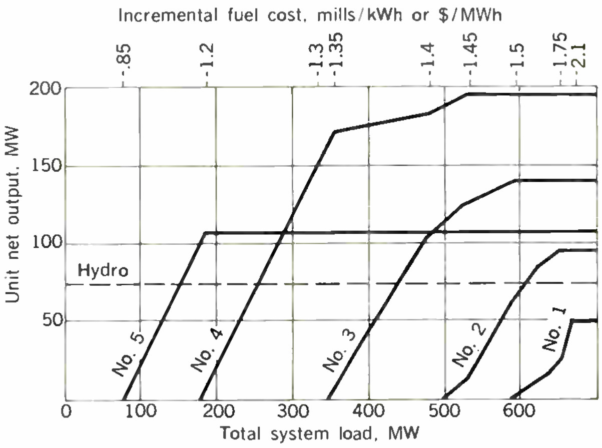
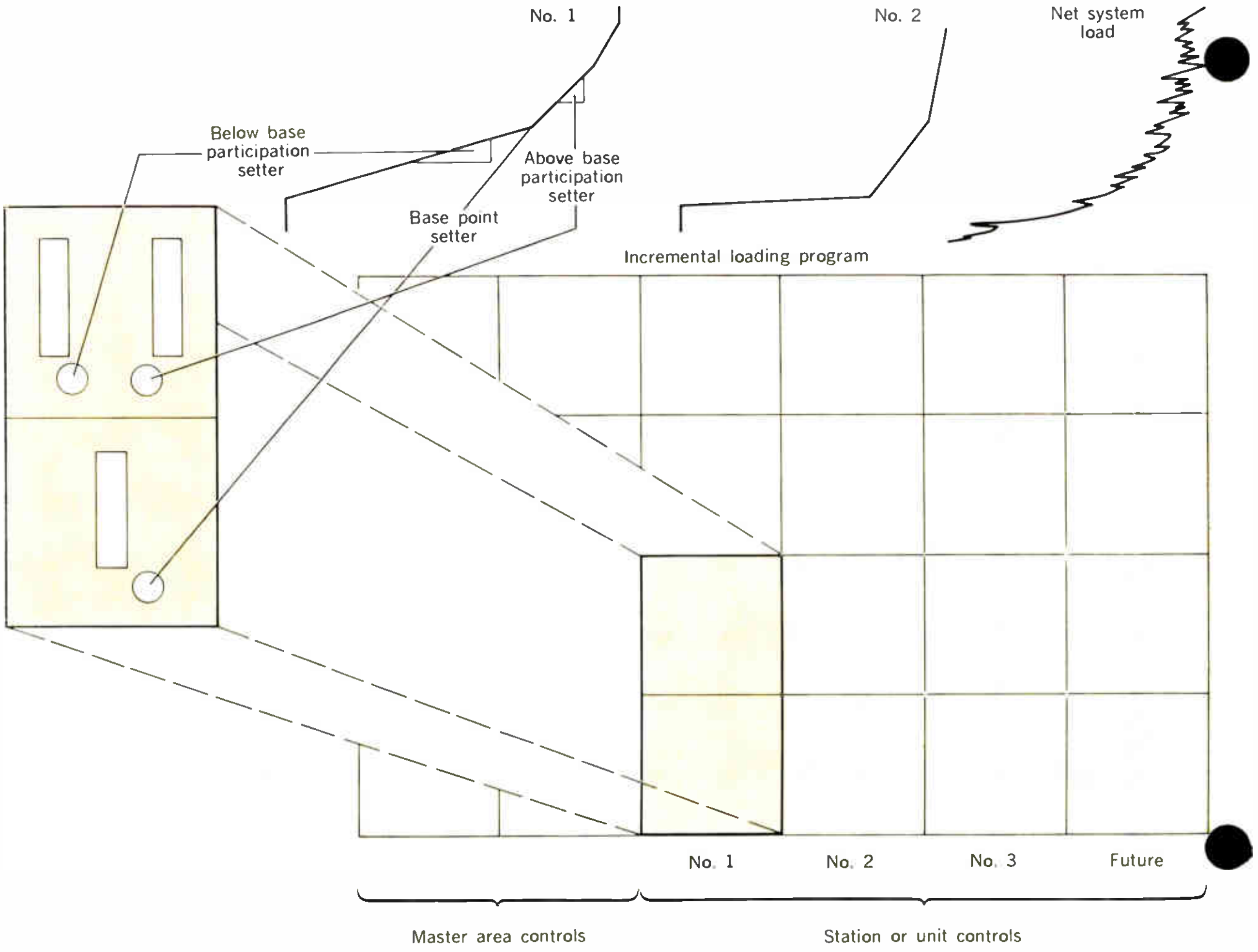


Fig. 5. Economic dispatch curves showing which units or stations should be carrying any given system load for the lowest generation cost.

Fig. 6. Block diagram indicating an adjustable economic program console. The system operator can match any given schedule by console setter dials to determine the breakpoint of the curve, and the slope of the curve above or below this point.



fully loading Unit no. 5 or the hydro, some regulating range must be applied to one or both of them; also, they should be equipped with *automatic control*—contrary to the "first glance" indications from Fig. 4. Further, it might be necessary to bring on Unit no. 3 earlier than anticipated—based on the incremental dispatch curves alone—to obtain the required rate most economically.

In actual operation some plants have inherent limitations on permissible rates of load change, while some limitations are abitrarily assigned. But few plants or units are initially designed to handle load regulation. After the base load period is terminated, however, by the addition of new capacity that will have an even lower incremental rate, the earlier plants or units will be used inevitably to handle variable loads. The lack of coordination between permissible rates of generation change and the rate of generation change required by consumer demand complicates the control problem. It indicates that control flexibility to meet changing load programs is an essential consideration.

The need for flexibility is illustrated further in Fig. 3 when the system generation is within the range of 500–525 MW. The Fig. 5 economic dispatch curves show that Units nos. 4, 3, and 2 should respond simultaneously, and during this period the attainable rate of generation change may satisfy both regulation requirement 2 and the incremental loading program.

For the time period when the system generation is between 525 and 600 MW, Units nos. 3 and 2 may furnish the necessary rate for the morning pickup, but may be unable to handle satisfactorily the more rapid noon change between these same values of system generation.

Transmission line losses. Up to this point, our application of regulation requirement 3 has related only to the incremental cost of regulating units on the station bus. To evaluate the true overall economic operation of a given area, however, one must also include the transmission line losses, which may produce a cost penalty that is caused by utilizing generation from a source remote from the load center. These transmission losses may be relatively small for a compact metropolitan system with heavy tie lines between nearby stations, but most interconnected areas in the United States, because of the geographical distances involved, have extensive transmission systems.

The transmission losses are influenced by the allocation of generation to the various system stations, and also by the interchange power flows between areas. After both factors have been considered, it is often found to be economically desirable to increase generation on a higher incremental cost unit near the load center, rather than to make this increase in generation on a remote unit that has lower incremental generation costs, but higher transmission losses.

If the incremental cost curves should produce an economy loading schedule that will always meet regulation requirements 1, 2, and 3, a rigid loading schedule may appear to be feasible. But experience on such a system, with consideration for the predictable conditions of interchange schedules, system load patterns, and the unpredictable conditions of the MW/min requirements of consumer demand, tie-line loading, and outage contingencies, must determine the extent to which these regulation criteria can be achieved.

On the basis of our assumed loading schedule, it may seem that the number of regulating hours per day for each unit could be determined. This oversimplification, however, may result in serious error when actual operating requirements are considered. After the system operator has used his best judgment in determining the available regulating capacity for his area load and interchange schedules, the automatic control should provide the necessary means for maintaining the desired conditions.

An adjustable program of automatic control is used in the United States for situations in which economic dispatch schedules, similar to Fig. 5, can be prepared in advance from available data. This control obtains its primary response from area requirements, and the corrective action is selectively routed to the regulating units in accordance with the economy loading schedule. The control does not fix the loading schedule, but the system operator can match any given schedule by console setter dials, as shown in Fig. 6, to determine the breakpoint of the curve, and the slope of the curve above or below this point. Although a complete schedule could be preestablished, most interconnected groups prefer to work from only two segments of the loading curve because of the simplicity and flexibility in meeting new conditions with a minimum of required coordination.

The use of computers for system control

In either the adjustable or fixed type of program control, the economies derived are no better than the precalculated loading schedules on which the control settings of the program are based. For this reason, many operating companies now use various types of computers to assist in the precalculations.

Many power companies are meeting their economic dispatch and area control requirements[3] with a combination analog computer and analog control system. If economic dispatch and simple calculation of economy interchange are the extent of the problem, this combination of equipment represents a proper and adequate solution.

The evolutionary technology. In recent years the concept of what a dispatcher should supervise has changed with the advances in technology, and the factors involved often go far beyond the economic dispatch and control problem. Functions such as economy interchange evaluation, energy accounting, interchange billing, unit selection, area security checks, data logging, hydro–steam coordination, transmission losses, system time error, and load-flow studies are now considered worthwhile for evaluation.

Table I lists in more detail some of the computer functions for power system automation.

As the problems of operation became more complex, it was necessary to develop more reliable hardware, and the digital computer has fulfilled this need to a considerable extent. There are four salient factors that recommend this equipment:

1. *Computer speed and throughput.* Computer hardware is available for an on-line, time-sharing operation, with input–output configurations that require no operator intervention. The computer's real-time throughput capability exceeds that required for the solution of real-time dispatching problems.

2. *Reliability.* The reliability of today's computer hardware has brought the control computer well within the range of practical utility application.

I. Computer functions for power system automation

Optimization	Implementation
Load forecast	Prediction of loads 24 hours in advance, based on historical data, trends, weather, etc., for use in other programs.
Generator incremental costs	The establishment of current incremental power costs for generators, either from fixed data or from results of computer heat rate calculation by power plant computer, for economic dispatch.
Economic dispatch	The computation of the most economic loadings for all units that are running, based upon generator incremental cost curves and transmission losses.
Unit selection	The selection of the most economic units to start, stop, and run for the next 24-hour period. The best solution is obtained by dynamic programming.
Spinning reserve	The computation of the amount of immediately available reserve needed to maintain acceptably low risk of power shortage while minimizing costs.
Interchange billing	The determination of the optimum amounts of power interchange and the computation of bills for the interconnected operation.
System security	A check for overload transmission circuits.
Load–frequency control	The loading of units to maintain frequency and interchange flows at desired levels. New settings for each unit are calculated every 5 seconds, or less, to follow load dynamics.

3. *Software.* Sophisticated programming, made possible by adequate data and operating experience compilations, provides a high degree of input accuracy and reliability.

4. *Increasing complexity.* The rapid growth of electric utilities in size and interconnection complexity dictates the acquisition of the computer to assist the dispatcher in planning and operating his system.

The computer may be used as an information processing system only, for direct control, or for control in conjunction with an analog system. In the first usage, a human operator performs the actual control on the basis of reports generated by the computer. In the second method, the digital computer is used for computation and also for handling the load–frequency control. And in the third technique, a digital computer is used to solve the complex computations, and it directs—through economic allocation—the operation of an analog-type load–frequency control.

The analog system. Figure 7 shows one type of conventional analog control system that utilizes base-point setters to determine the economic dispatch, and by means of an analog control loop and regulation participation setters, it performs the tie-line frequency regulation function. This system can be maintained manually near economic dispatch by an operator who periodically adjusts the base-point setters as the system load changes. Numerous details of the basic system configuration can vary since

1. Regulation groups can be composed of generator units, groups of generators, or entire plants.

2. The generator or plant controller can be located either in the central dispatching office or at the generating station.

3. Actual power output as relayed back to the dispatching office controller can represent a single unit or a group of units.

The analog control system has had an excellent record of reliability, and it can perform the basic load–frequency control function even if a digital computer is not directing it for strict economy.

All-digital system. As the term indicates, the all-digital-control system eliminates the analog regulating loop in the dispatching office. Figure 8 shows a system in which the generator control loops are also incorporated in the computer. Thus there is an apparent contradiction in terms, since an "all-digital-control system" can contain analog elements. Actually, there are two classifications for direct digital control.

1. *Digital control with analog subloop control.* In general, this type of system is used when extensive power station control equipment has already been installed. Area control error and the desired output of the generators are computed digitally, while the generators are balanced to desired power by analog means.

2. *Complete digital control.* This involves all loops, and evaluates system and unit response characteristics. This type of installation requires more computing capability. With proper programming, however, the ultimate in system–unit response, flexibility, and reliability can be obtained. This system can be applied where an older existing analog control system can be retained for backup.

Digital–analog, or hybrid, system. Actually, the Fig. 7 analog control system can be converted to a hybrid system merely by adding a digital computer that will solve the economic dispatch problem. Thus the computer, instead of the operator, will adjust the base-point setters. The updating of the base setter is often done on a fixed-time basis—say at intervals of 10 minutes—but, in general, the time interval varies with the application and the capability of the computer.

Figure 9 shows the hybrid dispatching system, with only the essential control loops designated. As in the case of the analog system, many variations in control configuration, grouping of units, and mode of control can be accommodated. Figure 9 shows that the digital computer will require all the usual inputs to perform a normal economic dispatch function. The computer input labeled "system status" represents necessary control information inputs that may include quantities such as maximum–minimum limits, tie-line telemeter channel failure, unit or group microwave telemeter channel failure, generator status (auto–manual–off), analog lockout, etc.

The elements of telemetering

As indicated in Part I of this article, telemetering is a system of measurement whereby the indication or the record of the measured quantity is produced at a location remote from the point at which the measurement is made. A simple telemetering system comprises three parts: a primary detector located at or near the measurement point, an intermediate means for transmitting the data to a distant location, and an end device that is capable of recording the values of the measured quantity.

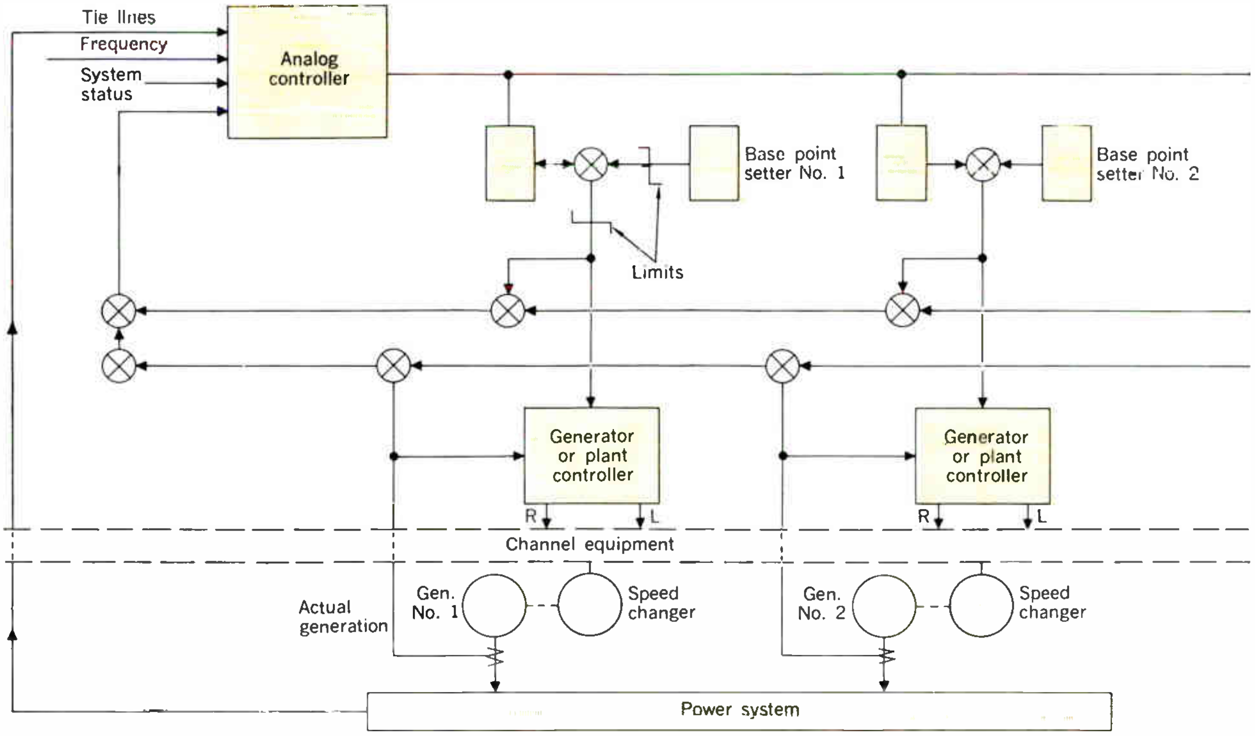
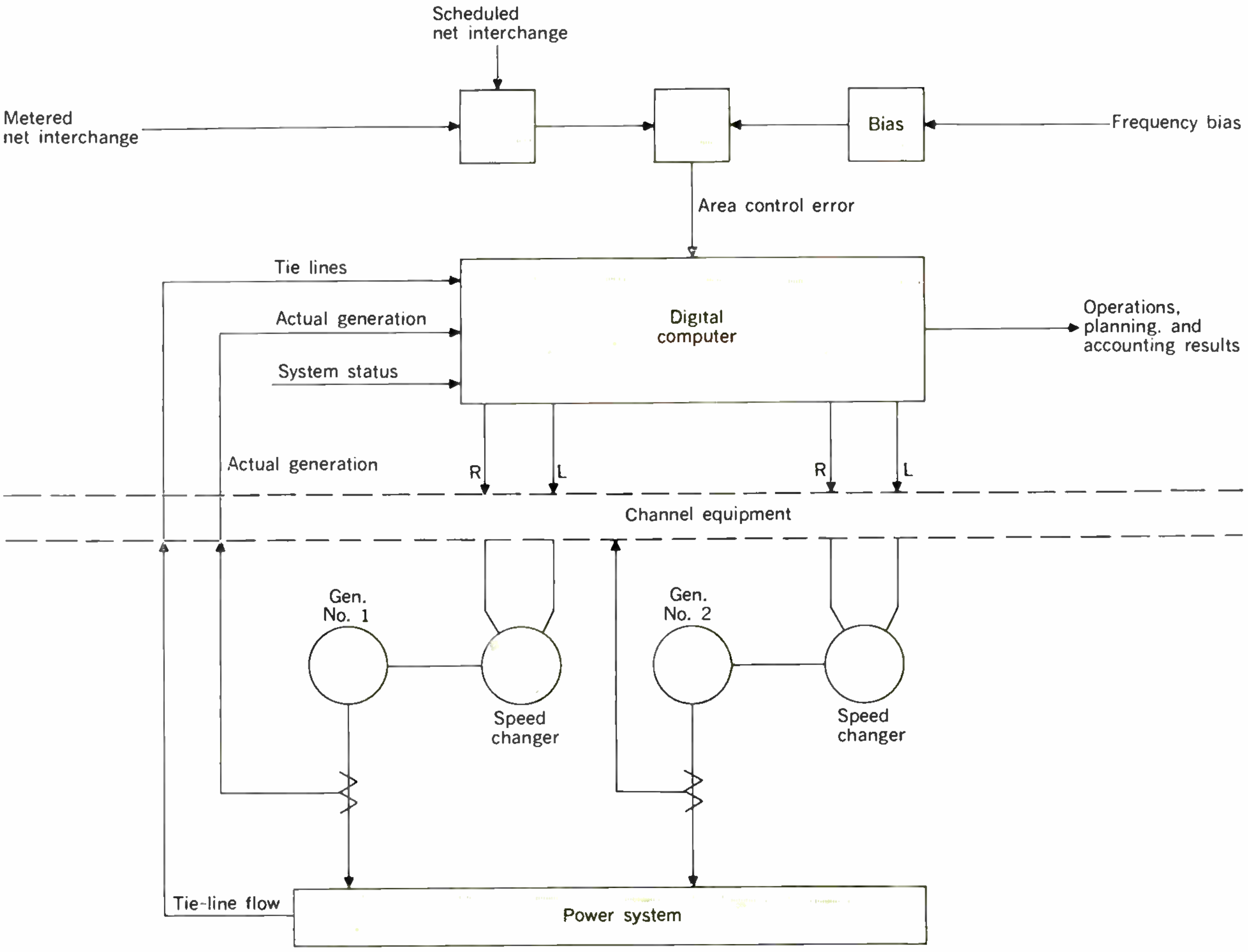


Fig. 7. One type of conventional analog control system utilizing base-point setters to determine the economic dispatch.

Fig. 8. An all-digital-control system in which the generator control loops are also incorporated in the computer.



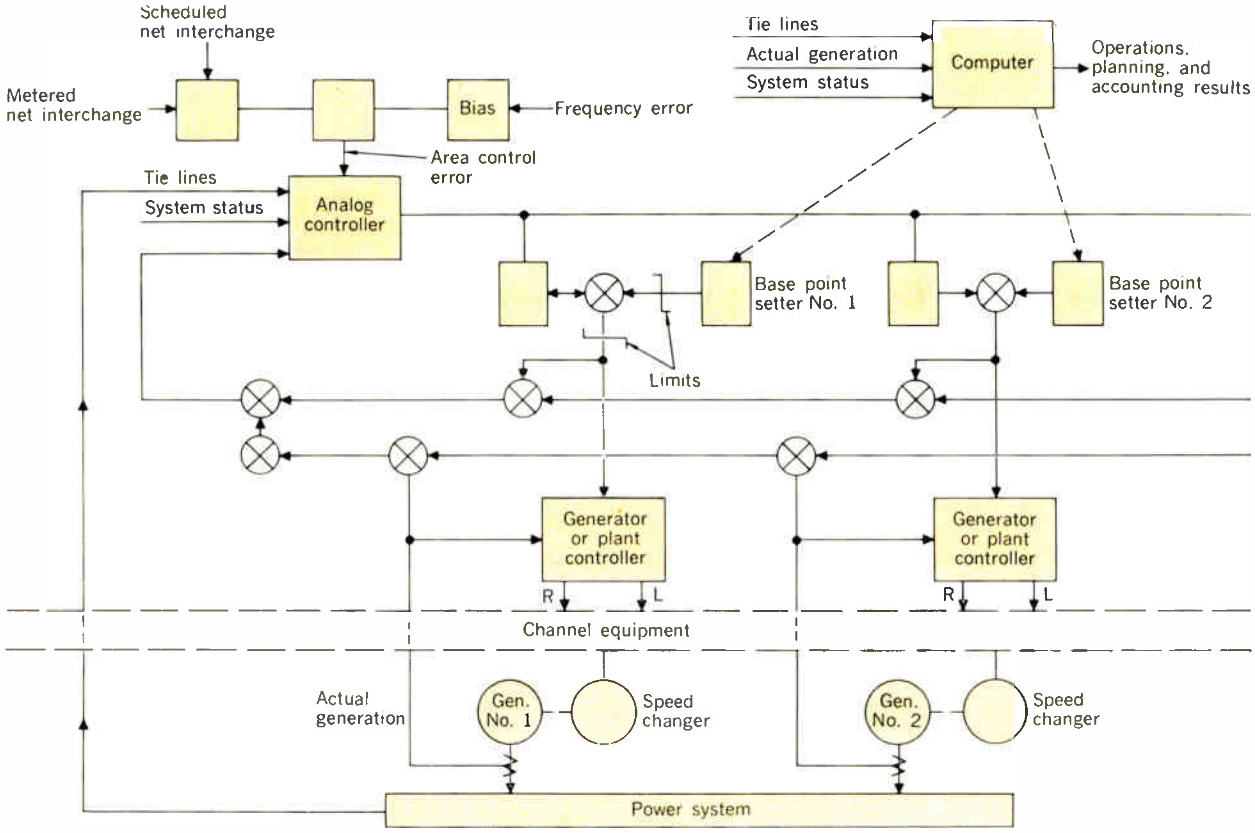


Fig. 9. Hybrid dispatching system (digital-analog), with only the essential control loops designated.

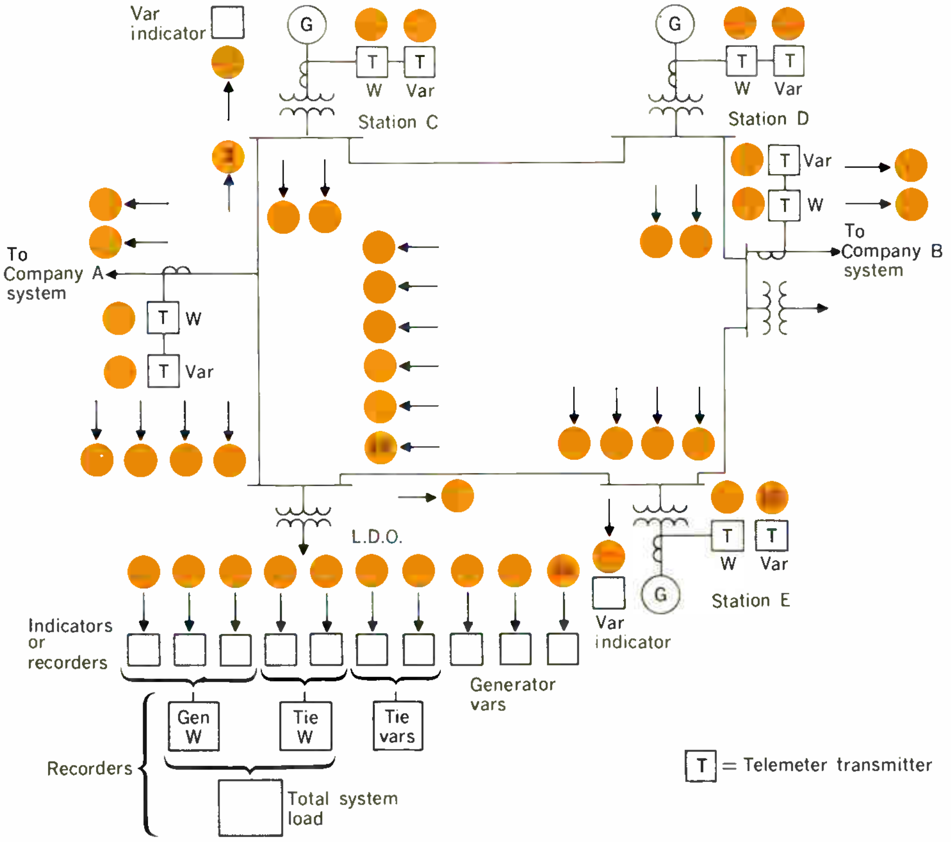


Fig. 10. Block diagram of a continuous, system-wide telemetering system. Note that all ten readings are transmitted to the load dispatcher's office where readings (1) and (2) are added together to show watts net interchange; readings (3) and (4) are summed to show vars net interchange; readings (5), (6), and (7) are added together to show total generated watts. This sum is then added to the watts net interchange reading to obtain the total system load. Reading (3) is also transmitted to Stations C and E, as the machines in these stations are used to regulate the vars interchange with Company A.

1. Watts interchange with Company A
2. Watts interchange with Company B
3. Vars interchange with Company A
4. Vars interchange with Company B
5. Watts output of Station C
6. Watts output of Station D
7. Watts output of Station E
8. Vars output of Station C
9. Vars output of Station D
10. Vars output of Station E

Telemeters are used for the remote measurement of current, voltage, power, pressure, and many other electric and nonelectric quantities. The operation of electric power generating and transmission systems is aided by the use of telemeters to provide the load dispatcher with these detailed data, and to give him on-line information on the amount and direction of power flow in interconnected tie lines.

Basic requirements. A telemetering system has certain basic functional requirements. It must be rapid in operation, simple, reliable, and readily maintained. Its action should be continuous, and the distance between the primary detector and the end device should not influence the performance of the system. These requirements usually make the frequency telemeter the first choice. The telemetering channel flexibility for communication between power stations of one system, or for intersystem area control, is quite extensive. It can include wire line, telegraph or voice channel, microwave, and carrier current.

System-wide telemetering. Figure 10 is a block diagram of a continuous, system-wide telemetering system. This system can instantly give the load dispatcher

1. A precise picture of the active power flow (watts) and the reactive power (vars) in all tie lines and at key points in the transmission system.
2. Continuous, on-line information on generating power and reactive power output.
3. Bus voltages at key points in the transmission system.
4. Reservoir water levels (in hydro and pumped-storage systems).
5. Loading of vital tie lines.
6. Total power interchange with network utilities.
7. Net power exchange (with all network utilities as a group).
8. Total generated power.
9. Total system load (exclusive of transmission losses).
10. Spinning reserve capacity.

In Fig. 10, readings 5–7 are needed for manual–automatic tie-line load control; readings 8–10 assist in scheduling the availability and distribution of generating capacity as required to meet load conditions throughout the daily load cycle. All of the readings are most important during times of power system disturbances.

Microwave communication systems

In recent years, a new generation of microwave communication equipment has been developed to serve the complex networks of large interconnected utilities systems. Microwave design improvements were undoubtedly accelerated by the urgent need for communication, telemetering, and control channels for which the existing carrier frequency spectrum and available leased-line facilities did not provide an adequate solution.

With the advent of computers, the centralization of more telemetered information is a prerequisite to the full utilization of automation potential. Therefore, microwave installations are increasing rapidly, and centralized information and control of many related power quantities will soon be possible.

Figure 11(A) shows a simple microwave communications system—a transmitting and receiving terminal connected by a repeater station. The equipment is of the IF heterodyne type, which does not demodulate to traffic frequency at repeaters where the station gain is principally at an intermediate frequency. Figure 11(B) illustrates a transmitting terminal that consists of an IF oscillator, frequency-modulated by the sending voice multiplex. The traffic, or baseband, spectrum, occupied by a 240-voice-channel signal, extends uniformly from about 60 kc/s to 1.3 Mc/s. The FM signal is shifted to the microwave channel by up-conversion in a mixer. The skeleton receiving terminal shown in Fig. 11(C) is merely a superheterodyne FM receiver, and, by interconnecting it with the transmitting terminal, a two-way heterodyne repeater is provided (Fig. 12).

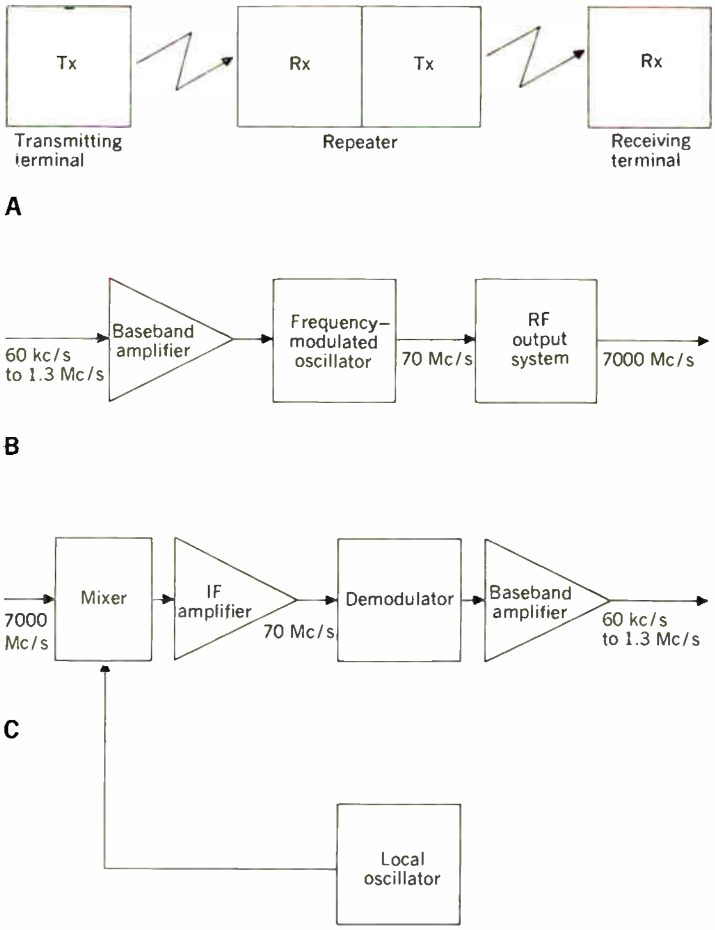


Fig. 11. A—Simple microwave communications system consisting of a transmitting and receiving terminal connected by a repeater station. B—Transmitting terminal consisting of an IF oscillator, frequency-modulated by the sending voice multiplex. C—Superheterodyne FM receiver.

Later in this article, we will discuss the application of this type of microwave system in the vast interconnected network of the American Power Company.

The Detroit Edison experience

A hybrid, or digital-directed analog control, system is in service at the Detroit Edison Company, a privately owned utility that serves metropolitan Detroit and a surrounding area. The system has been in service since April 21, 1963. The basic functions of the installation are

1. Generation and tie-line control.

2. Automatic economic dispatch, including transmission loss calculations, security and regulating calculations.

3. Data logging that includes periodic alarm and a record of communication by the operator with the computer system.

4. Operation studies to provide interchange evaluation and billing, plus unit scheduling and schedule preparations.

Economic evaluation. In the economic evaluation of the benefits of the computer installation, the owners believed that improved economic dispatch, reduced equipment running time, greater accuracy of interconnected transactions, and improved load assignment on two large generating stations could be achieved.

Three years of continuous operation have also indicated savings in the safer operation of the Ontario interconnections, improved system response and control, and better system information for the operator.

In a comparison of the credits and debits of actual operation, projected over a 10-year period, the computer and associated control equipment cost is $142 000 per year. And there is a uniform credit of $190 000 per year; therefore, the new equipment can pay for itself from credits, and provide the owners with an additional savings of $48 000 per year throughout the 10-year period.

Performance of basic functions. The primary basic function of the system is that of generation and tie-line control—the heart of which is the *area control error*. As previously explained, the area control error is the instantaneous measurement of the number of megawatts which a power system must increase or decrease to fulfill its responsibility of maintaining a balance between its consumer load and generation, plus the system's responsibility to its interconnection commitments. As shown in the Fig. 13(A) block diagram, the area control error signal, derived from the proper tie line, frequency schedules, and actual generation, enters the master controller and causes the generating units either to raise or lower their output to maintain a balance between system generation and consumer load.

Figure 13(B) shows the insertion of a flexible loading program console that is manually operated and inserted into the computer control loop. This method provides a noninteracting program between the controlling dispatch units so that the rate at which any source responds to the control action will not influence the desired megawatt value computed for the other dispatch units. The only common tie between the various dispatch units is that all of the generation changes will always be made to minimize area control error.

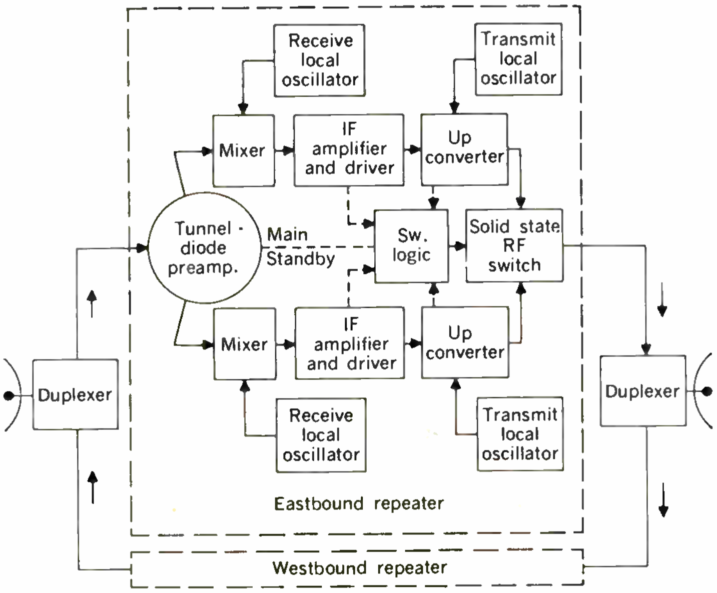
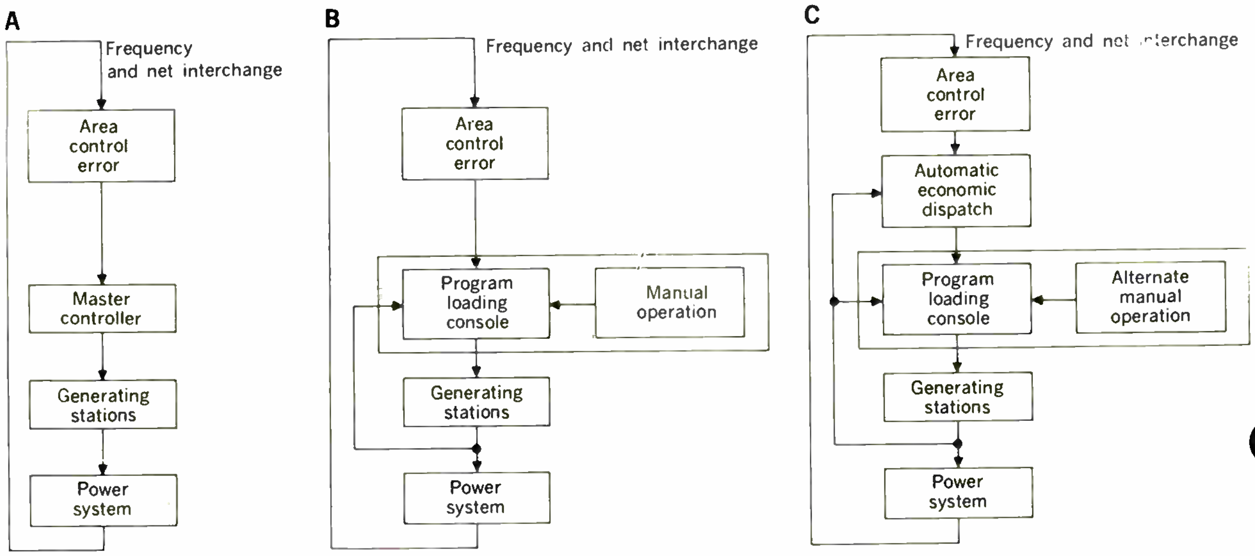


Fig. 12. When the superheterodyne FM receiver shown in Fig. 11 is interconnected with a transmitting terminal, a two-way heterodyne repeater is provided.

Fig. 13. A—Block diagram of the basic analog control action. B—Block diagram showing the insertion of a flexible loading program that is manually operated and inserted into the computer control loop. C—Block diagram of the basic digital-computer-directed analog control required to perform the functions of generation, tie-line control, and economic dispatch.



Automatic computer control. The basic analog control installed at Detroit Edison is the same as that just described, except that the base-point setters can be either manually set, or set by the digital computer, Fig. 13(C), as the new economic dispatch schedules are derived. This is the basic digital-computer-directed analog control required to perform the functions of generation, tie-line control, and economic dispatch.

In Fig. 14, we see that specific zones are represented on the area control error recorder. If a good control job is being done, and control error is retained within the normal band, all of the units are loaded exactly to their economic loading schedules. But if control error drifts beyond the strict economy loading bands, the following degrees of assist or emergency action are energized:

1. *Normal assist*—in which certain preselected dispatch units are permitted to divert a prescribed number of megawatts from their economic schedule.

2. *Emergency assist*—whereby all dispatch units within the regulating range receive control intelligence, regardless of economic loading.

3. *Scram assist*—in which all the dispatch units within the regulating range receive one-second control impulses for a period of 30 seconds.

Security calculations. In the security calculations and restraints function, ties between areas within the Detroit Edison system are examined to determine: the loss effect of the most heavily loaded unit in the area, the loss of the most heavily loaded tie, etc.

If an economic dispatch must be modified to meet security restrictions, the dollar cost will be logged on a typewriter to provide an updated record of the various security restraints on economic loading. Regulating margin restrictions are used to permit deviations from economic loading for specific operating units, so that higher cost generators may be put on the line and brought up to minimum load.

Interchange evaluation and billing. Interchange evaluation is an important preliminary perusal of a proposed contract to buy or sell power with a neighboring system to ensure that such an interchange would be mutually beneficial on a split-savings basis. This evaluation is performed, as a study function, upon request by the operator.

Interchange billing is achieved automatically when the net interchange schedule setter is not at the zero point (Fig. 14). Thus, at the end of a specified time period, the operator has a complete billing record of actual interchange cost.

Under the unit commitment and withdrawal program, the computer also computes the required spinning reserve, start-up costs, and operating costs. It then advises the operator by print-out whether to add or remove a system unit.

In Fig. 15, we see that the basic Detroit Edison block diagram is again modified to include the other digital computer functions just explained. And the final additions to complete the diagram are the operator's communication console and the associated readout typewriters.

Human engineering. The ability of the operator to understand and to communicate with the computer control system is a very important subject, since the "human engineering" element has a definite effect upon the success of an installation.

Detroit Edison evolved the operational concept of a simple transition from the digital computer control to manual operation of the generation and tie-line control equipment. This objective was met by using stepping motors as the setters in the analog control equipment.

When the digital system is out of service for any reason, the last dispatch remains undisturbed on the base-point setters. Control will continue, however, under base-digital-computer outputs to drive the base-point and participation direction. The base-point setting is adjustable by hand if the load level changes sufficiently to justify the resetting of the economic allocation.

AEP System computer center complex

The American Electric Power System (AEP) is the largest investor-owned producer of electric energy in the United States. The AEP System operates a group of power plants that provide electric service in parts of Indiana, Virginia, Michigan, Ohio, West Virginia, Ken-

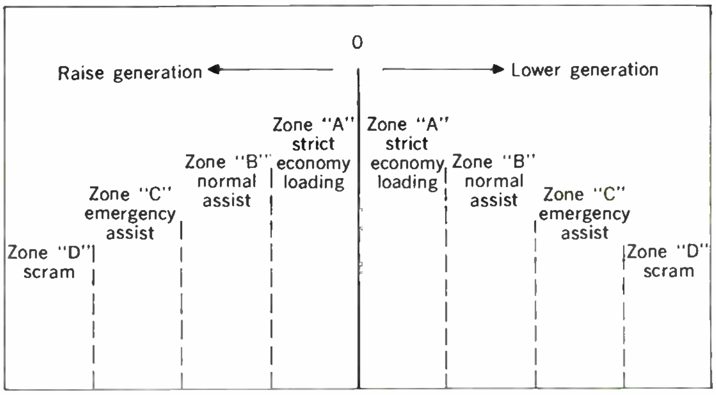
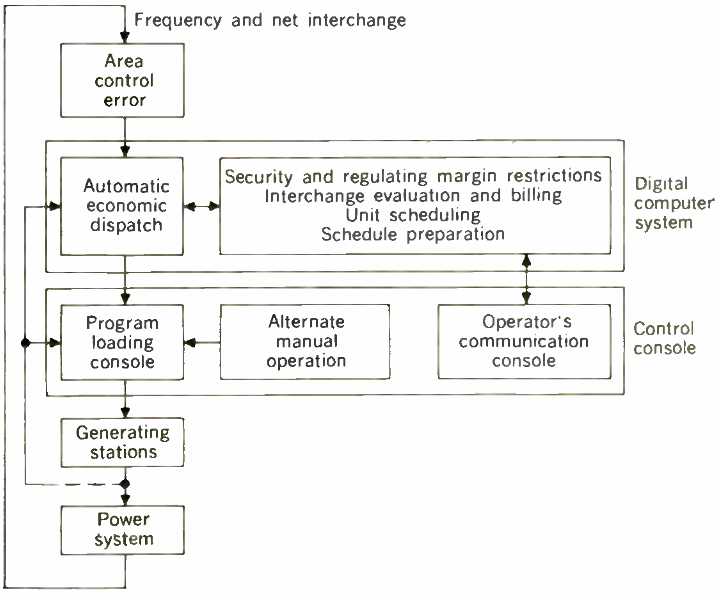

0
Raise generation
Lower generation
Zone "A" strict economy loading
Zone "A" strict economy loading
Zone "B" normal assist
Zone "B" normal assist
Zone "C" emergency assist
Zone "C" emergency assist
Zone "D" scram
Zone "D" scram

Fig. 14. Diagram showing the specific zones that are represented on the area control error recorder. Note that the device is at zero (mid-scale) when system regulating commitments are fulfilled.

Fig. 15. Final modification to the Detroit Edison basic block diagram includes all elements of the complete digital-directed analog control system.


Frequency and net interchange
Area control error
Automatic economic dispatch
Security and regulating margin restrictions
Interchange evaluation and billing
Unit scheduling
Schedule preparation
Digital computer system
Program loading console
Alternate manual operation
Operator's communication console
Control console
Generating stations
Power system

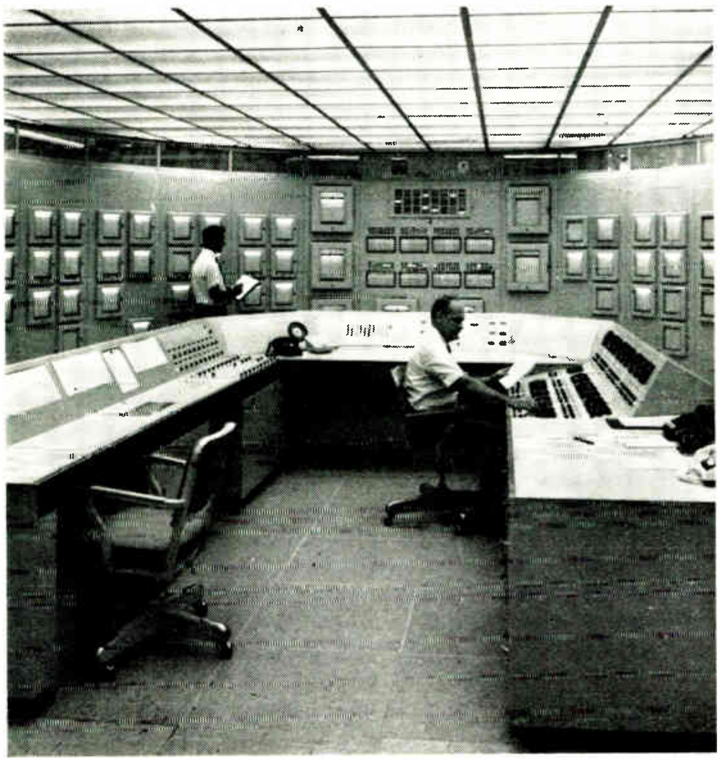


Fig. 16. View of American Electric Power's digital-directed analog control system at the Canton center. The equipment automatically controls the generation and distribution of power over a seven-state network.

tucky, and Tennessee. These plants have a total generating capability of more than 7800 MW. The system has a transmission and distribution network of 80 000 miles that includes 1600 miles of 345-kV EHV transmission.

The Canton "nerve" center. The AEP System has recently completed its integrated "nerve" center in Canton, Ohio. This installation consists of a *power control center* that automatically supervises the economic dispatching of more than 6000 MW of electricity from 15 major power plants, regulates power frequency, and controls the power interchange with neighboring utilities. In addition, a *data-processing center* automates and centralizes the billing functions for the system's 1.5 million consumers, and it correlates accounting, financial, engineering, and management information.

The power control center combines an analog computer system, and digital telemetering, with a digital computer system (see Fig. 16). By the continuous monitoring of the demand for power, and the availability of the generating units to supply it, the installation attempts to meet the total electric power requirements of all AEP System customers—and the power commitments to the interconnected network—in the most economical manner.

Data collected by instrumentation at hundreds of key locations on the AEP System are telemetered via microwave to the center. After extensive evaluation and computation by the digital computer, automatic instructions are routed back through the system by microwave to the appropriate generating units where power production at 38 generating units is either increased or decreased automatically as required. The circular flow of data in a closed-loop system is accomplished in a few seconds and it requires no manual intervention.

One of the primary functions of the power control center is to assure constant load-frequency control at a steady 60 cycles per second. The digital computer is activated every three to five minutes for the economic dispatching operation and load-frequency control.

The automatic selection of power generating and dispatch alternatives takes into account:

1. The relative generating efficiency of each unit.
2. The relative fuel cost at each plant.
3. The relative amount of energy lost in transmission over a given power line for a given distance.

From these weighted factors, the computer accurately selects the units to load and how much to load them to produce the highest combination of generation–transmission efficiency and economy for the entire system.

Regular transaction studies provide on-line calculations of various blocks of power, and their price, for sale to interconnected utilities. The calculations also help to evaluate the system's past and future operating performance. The data acquired from more than 800 input stations are automatically updated and stored every ten minutes to provide minute-by-minute analysis.

The computer hourly produces a complete log that includes the system load, generation, spinning reserve capacity, transmission line losses, hourly and average fuel costs, etc.

Functions of the data-processing center. The data-processing center is the nucleus for the centralized processing of customer billing, general accounting, and management information for the entire network.

The installation consists primarily of a large-capacity, tape-oriented, general-purpose digital computer that is served by an auxiliary data-processing system.

Meter readings taken throughout the AEP's 27 divisions are converted into machine language and fed into satellite computers in Fort Wayne, Ind. (for customers of Indiana & Michigan Electric); and Roanoke, Va. (for the Appalachian, Kentucky, and Kingsport power companies). The Canton center, of course, handles the accounts of Ohio Power and Wheeling Electric customers directly. Information from the satellite computers enters the data transmission units at Fort Wayne and Roanoke, and is microwaved to Canton at the rate of 15 000 accounts per minute.

Upon receipt of the data, the master computer prepares bills at the proper rate from its memory storage. The system simultaneously records on tape the account status of each customer—plus information for the future processing of bills—at the rate of 3600 accounts per minute. As each billing calculation is performed, statistics are accumulated for rate and revenue analyses, marketing studies, and Federal and state regulatory reports.

Finally, tapes containing billing data and other information are fed back through the microwave network to the computers in Fort Wayne and Roanoke where customers' bills and accounting reports are printed.

The microwave system. The basic microwave installation consists of terminal and heterodyne repeater stations. It operates in a band close to 7000 Mc/s, and it is capable of handling 300 voice channels for distances in excess of 1000 miles (see Figs. 10–12). The equipment uses the heterodyne principle at repeater stations, thereby making it unnecessary to demodulate the signal at each station. The data transmission rate is 15 000 characters per second.

The 1700-mile network provides the communications link for the entire computer center operation at Canton. In the ten years since its original installation, the network has been expanded and modernized. The manufacturers, in developing the latest phase of the equipment, have used transistors, tunnel diodes, and varactors throughout to eliminate all mechanical relays, vacuum tubes, and klystrons. Thus the new microwave equipment has more compact physical dimensions, greater reliability, and improved economy of operation than predecessor models.

The hub of the microwave network is a 390-foot-high tower, topped by a 10-foot-high antenna for the transmission and reception of signals between Canton and the power plants and the satellite computers. The base of the tower forms an 80- by 80- by 80-foot triangle, and its three legs rest on concrete foundations that are 16.5 feet deep.

The Leeds & Northrup Company was the prime contractor for the AEP's electronic nerve center at Canton, with IBM, General Electric, and RCA acting as subcontractors for various phases of the computer, microwave, and data-processing equipment installations.

An industry first—a complete digital system

The Kentucky Utilities Company in Lexington has the first complete digital system operation computer, with all major functions incorporated in the machine. Unlike most hybrid or digital-directed systems, the computer will provide load-frequency control and economic dispatch, and will output raise-lower control signals over three frequency-tone channels to each generating unit on the system. And the computer will accept analog inputs from 12 individually telemetered tie-line quantities and from eight generating units located at five stations. The computer is the most extensive all-digital control system application to date, with 8000 words of high-speed core memory and 32 000 words of drum memory that will also provide extensive off-line capability for load flow computations and FORTRAN programming.

Load–frequency control. The digital computer is programmed to compute area control error from analog input quantities, and output individually adjusted control impulses to each generating unit. This control cycle, including the scanning of telemetered inputs of unit and tie-line loads, system frequency, computing and outputting control impulses, is accomplished every four seconds. The load–frequency control program utilizes penalty factors that are calculated by the economic dispatch program so that each dispatch is economic within the limits of permissible response rates of respective units.

Area control error is computed by scanning individual tie lines every second, by converting these quantities to digital units, by application of scaling factor, and by comparison with manual input quantities of desired net interchange and scheduled frequency.

Each generating unit is also controlled by means of digital programs in the machine. The generating unit controller subroutine compares the assigned power for each unit with the actual telemetered power and produces control impulses to bring unit generation to the desired value.

Economic dispatch. The economic dispatch program utilizes the standard transmission loss formula equation. Two sets of B-coefficients, representing different system configurations and load, are stored and called into use as required. Each generator incremental cost curve is represented in the computer by a series of straight-line segments, and a subroutine is provided to combine cost curves into station and system total cost curves. The economic dispatch program develops for the load–frequency control program all necessary information for individual unit assignments and also checks the adequacy of spinning reserve for each dispatch.

Data logging. The scan-log-and-alarm program, adapted to the requirements of the company, provides an hourly logging of desired system quantities and also provides alarm messages for emergency conditions. In addition, a circuit-breaker logging program affords the supervision for the operation of 200 circuit breakers by giving the time of the breaker operation, the name of the breaker, and the action.

The off-line, or shared-time, program capabilities of the computer include unit commitment, and evaluation of sales and purchase of power.

Arizona Public Service System

This system consists of a digital computer with 16 000 words of high-speed core memory, 32 000 words of drum memory, and a computer-directed, transistorized analog control system. The equipment was placed in service at Phoenix by the Arizona Public Service Company in 1963. The initial system controls 13 generating units in seven power stations.

Reasons for digital–analog approach. The digital–analog approach and the eight-channel computer were chosen to allow maximum utilization of computer capacity for tasks other than the traditional economic dispatch. These functions include

1. Control and economic dispatch for the power system.
2. Scheduling and forecasting.
3. Network analysis studies.
4. Data-processing assignments.

Versatility of system. The eight-channel computer, with an elaborate and versatile priority system, will permit the time-sharing of many tasks of on-line system operation. Thus each area can operate independently of the others. For example, all turbine generating units can be operated manually from the local generating plant control, the analog control system is superimposed on the generating plant to regulate generation automatically, and finally, the digital computer is superimposed over the analog control system to obtain completely automatic economic dispatch of the power system.

The computer is coupled to the rest of the control system through contact closure outputs, inputs, and analog inputs to provide system operating information, control information, and communication between the dispatcher and the system.

Emergency power supply. To protect against a loss of the digital-control system either from power failure or surges, the system receives its ac power from a solid-state 7.5-kVA battery inverter system. While the battery supply is maintained by continuous charging, sufficient capacity is available to permit 30 minutes running without recharging, in the event of ac power loss. This will provide sufficient time to transfer to another emergency ac supply during long-term outages.

The equipment for the Kentucky Utilities Company and Arizona Public Service Company installations was furnished by the Westinghouse Electric Corporation.

Con Edison's energy control center

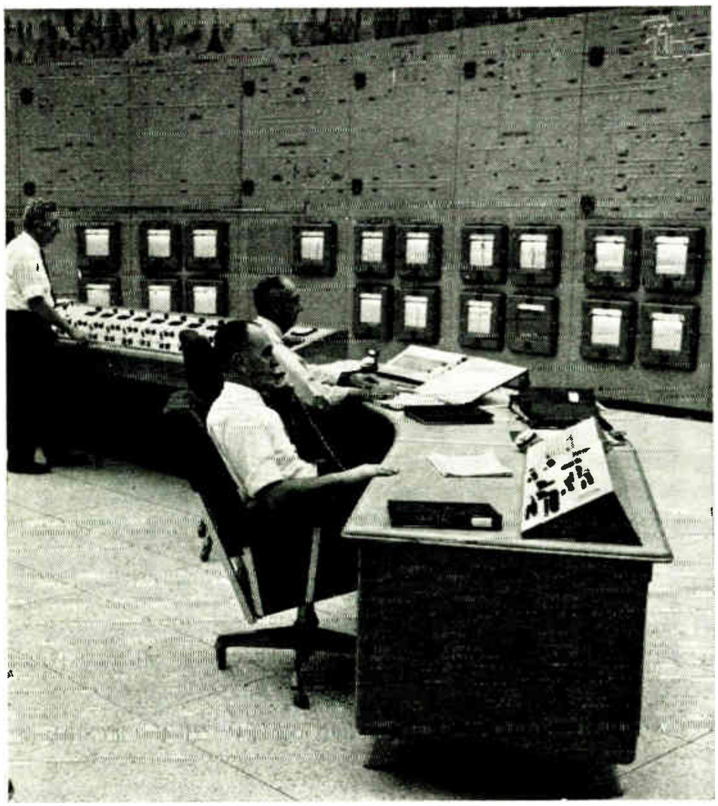
Early in 1959 it became apparent to the Consolidated Edison Company of New York[4] that additional facilities would be needed for the coordination of operations of the system's rapidly expanding electric generating and transmission system. There was also a need for additional space to accommodate instruments and indicating or alarm devices required by the establishment of new and unattended bulk power stations.

It was estimated that the functional requirements of the new control center would be best met by a new structure, measuring about 80 feet in width by 120 feet in length, free of interior columns, to accommodate the control room, the observation area, and the necessary spatial adjuncts.

The site selected for the construction of the control center is at 128 West End Avenue in Manhattan, near Lincoln Center.

The supervisory control panels were installed in the control room for all of the existing and future major distribution substations in the Borough of Manhattan, and also the control facilities for a major transmission substation near Poughkeepsie, N. Y. Thus it was necessary to locate the associated relay cabinets on another floor of the building, thereby necessitating a two-story structure.

Fig. 17. View of the recorders, visual displays, and control console at the Consolidated Edison control room as seen from the system operator's desk.



Control room facilities. The primary functions of the control room facilities are

1. To display information as to the exact instantaneous status of each element of the company's electric, gas, and steam systems.

2. To permit reliable remote control for those elements of the electric transmission and distribution systems that are specifically intended for control and indication at the center.

3. To afford quick and positive communication with each generating station, attended substation, gas holder station, and steam production station in the company's system. Similar communication facilities are provided to reach the dispatchers of other electric systems with which the Con Edison system has interchange agreements.

Information display. The information display installed at the Energy Control Center differs in some respects from those used by other electric utility systems. For example, there is no complete diagram of system transmission and distribution circuits on the display panels. The use of line diagrams is limited to the upper portions of the panels directly in front of the working positions of the system operator and his assistant (Fig. 17) who are responsible for

1. The regulation of generating units to meet the estimated system electric load at all times (estimated at a peak of 6000 MW for 1965), with established objective margins of spinning reserve.

2. Forecasts of anticipated electric system load for the ensuing 24-hour period.

3. The supervision and control of the generation level assigned to each generating station for economic dispatch.

4. The negotiation of hourly schedules for capacity and energy interchange with the companies that comprise, with Con Edison, the Southeast New York power pool.

5. The coordination of schedules for capacity and energy interchange between individual utilities of the Southeast New York pool and the Niagara Mohawk Power Corporation.

6. The supervision and control—as required—of the direction and magnitude of power flow in the loop-type 138-kV transmission system to avoid the overloading of component parts of the underground cable circuits that comprise this system.

The system diagram displayed for the operator's use shows individual 345- and 138-kV circuits from terminal to terminal, transmission substation buses, all circuit breakers, and all disconnecting switches—except those especially used for circuit breaker isolation.

In most cases, the transmission circuits between adjacent stations consist of two or more parallel feeder circuits. These are shown on the transmission system diagram, but they are grouped so that an indicating wattmeter, mounted on the diagram panel, clearly shows the direction and magnitude of the power flow in each *group* of transmission feeders at each generating station or substation feeder terminal. This feature is very helpful when more than one automatic circuit breaker trips at the same time, and it assists the system operator when he directs the change of taps in the phase-angle regulating transformers to correct undesirable flow of power in ring-type circuits.

Recording wattmeters are provided to totalize generation at each station. The instruments for the 60-c/s stations are the two-pen type that show the total megawatts generated at 60 c/s, and megavars, in or out.

Between the working positions of the system operator and his assistant is the console for tie-line and frequency control, which has panels for a total of ten generating stations to participate in automatic control. Seven of these are connected for use, but not all are used simultaneously. Tie-line load regulating duty is assigned to the stations that have the highest incremental production cost during those hours when the higher cost stations must be operated to provide the required system capacity. At such times the lowest cost stations are base-loaded and off control. And during the hours when the higher cost stations are shut down, or held at minimum load for area protection, the regulating duty is transferred entirely to the lowest cost units, which are then generating at less than normal capability.

The remaining display panels for the electric system are primarily used for the district operators who have the responsibility of operating the distribution system in Manhattan and the Bronx. Separate district operator control centers are maintained for the other boroughs of New York City and for Westchester County.

Manhattan distribution. The Con Edison distribution system in Manhattan is the most complex in the company's service territory. Also, the highest load densities exist here. At present there are 24 separate 60-c/s network areas, with supply feeders of 13 800 volts. The status of each feeder must be indicated on the display panels that are used for reference by the district operators.

The Manhattan substations that supply networks are usually unattended facilities. Supervisory control for each of these stations is provided at the Energy Control Center, where the district operator can open or close any circuit breaker, change taps on the main transformer banks to regulate bus voltage, and take indicating instrument readings of certain quantities. Provision is made, however, at these substations for automatic control of bus voltage. Any automatic circuit breaker operation or transformer tap position disagreement is automatically indicated on the display panel by lamp and audible alarm.

Con Edison's 25-c/s distribution system is supplied from three of its generating stations by 11 400-volt feeders. The loads consist of 25 synchronous converter substations in Manhattan that supply the 120–240-volt dc networks in the borough. Other 25-c/s feeders at 11 400, 22 000, and 33 000 volts supply traction loads for four main-line railroads that terminate in New York City. The status of these feeders, as well as the 60-c/s feeders, is displayed by feeder identification plates in card holders, grouped according to substation or high-voltage customer destination.

Leased wire facilities. The company considers that the leased wire circuits used for communication, telemetry, and supervisory control and operation of the tie-line load and frequency equipment are so important that unusual precautions are taken to ensure continuity of service. One of the major hazards to underground utility and communication cable installations under city streets is caused by water main breaks. For this reason, the leased wire circuits for the most important functions are provided in duplicate, and they are carefully supervised both by representatives of the New York Telephone Company and Con Edison to assure maintenance of route segregation to the most practicable extent.

Computer-directed control system. A $1.3 million digital-directed hybrid computer system is presently being installed at the Energy Control Center in anticipation of large power blocks from interconnections with other utilities, the proposed pumped-storage plant at Cornwall-on-Hudson, and possibly from Hamilton Falls, Canada. Also, the new Ravenswood Unit 3, rated at 1000-MW capability,[5] will go on the line this summer.

The new system, scheduled to be in operation in 1966, will allocate load to any of 23 combinations in 12 power plants, and will perform automatically and continuously for maximum economy of system operation. Units in any combination will be selected for their similar heat rates. Therefore, with the new equipment, unit commitment will replace the present station commitment.

Information from the digital computer will be transmitted to the analog unit to set generator schedules automatically throughout the system. The digital equipment will provide memory and logic, while the analog units will furnish the dynamic qualities of simultaneous control, continuous monitoring, and backup facilities.

In addition to more economic and efficient load dispatch, the hybrid control will provide automatic interchange scheduling between Con Edison and neighboring utilities in its interconnection.

The inclusion of var scheduling and control in the new system is under consideration. As this is a compact metropolitan system, the transmission losses at present are relatively small, and will only become considerable after Cornwall goes on the line.

Of things to come

It is predicted that within the next decade we will have computer systems reporting to and controlling each other. Computers located in power plants will report to a system "controller," which will then receive instructions from an overall policy-making machine through which management will exercise its ultimate control.

A maximum R & D effort on the part of computer system manufacturers, scientists, engineers, and programmers will be necessary for an even greater understanding of controlled processes and systems. Faster and larger computers are being designed and built, and advanced compilers for easier programming are on the way.

The author wishes to acknowledge the courtesy and cooperation of the Leeds & Northrup Company, General Electric Company, Radio Corporation of America, and the Westinghouse Electric Corporation in making available technical information, diagrams, and photographic illustrations for this two-part article.

REFERENCES

1. Hopkins, W. A., "Why Go Digital for System Operation?" presented at 1964 Utility Engineering Conference, East Pittsburgh, Pa., Apr. 26–May 8, 1964.

2. Morehouse, S. B., "Recent Developments and Trends in Automatic Control of Large Interconnected Power Systems," presented at the 1956 Session of the Conference Internationale des Grands Réseaux Electriques à Haute Tension, Paris, France.

3. Graham, J. H., and Hissey, T. W., "Latest Trends in Automatic Load-Dispatching Computers," paper submitted to Southeastern Electric Exchange Conf., Washington D. C., Apr. 16–17, 1964.

4. Otten, H. C., "Consolidated Edison's Energy Control Center," presented to Electrical Systems and Equipment Committee, Edison Electric Institute, Oklahoma City, Okla., Oct. 15, 1963.

5. Friedlander, G. D., "Giant Generators for Growing Power Demand," *IEEE Spectrum*, vol. 2, no. 2, Feb. 1965, pp. 70–86.

Trends in the development of modern radioelectronics

V. I. Siforov *Corresponding Member of the Academy of Sciences of the U.S.S.R.*

At the invitation of the IEEE, a delegation of seven members of the Popov Society of Russia attended the International Convention in New York. The delegation was headed by Dr. V. I. Siforov, the President, who is a corresponding member of the Academy of Sciences of the U.S.S.R. During the Convention, a special session was held which featured an address by Dr. Siforov entitled "Trends in the Development of Modern Radioelectronics." This article contains the full text of that address. I am sure that the readers of SPECTRUM will find Dr. Siforov's point of view very interesting. We are happy to be able to make the text of his talk available to members of the IEEE.

F. Karl Willenbrock
Editor

The role of science in the progress of mankind grows steadily. One of the leading trends in modern science is the emphasis on radioelectronics.

Seventy years have passed since radio was invented, i.e., the significant day of the 7th of May 1895 when the Russian scientist A. S. Popov published a design for a radio receiver and demonstrated it publicly by transmitting messages in the form of elementary radio signals from an artificial source of electromagnetic waves. During this historic but rather short period, radio engineering, and then electronics and radioelectronics. achieved surprising successes.

Electronics started with the invention in 1906 by the American scientist Lee de Forest of an electron vacuum device—a triode. Electronics and radio developed successfully and simultaneously by cooperating with many sciences and fields of engineering to become radioelectronics.

Radioelectronics encompasses such fields as radiospectroscopy, radioastronomy, radiometeorology, electronic computing devices, and television, as well as many others. It is difficult these days to name any field of science, any kind of engineering, any industry, or any culture in our modern society which radioelectronics has not penetrated in some form.

Like many other sciences and fields of engineering, radioelectronics has developed very rapidly. Over the last 20 years, radioelectronics has progressed further than during the previous 50 years although, even then, significant results were obtained. During the last 20 years semiconductor triode-transistors (1948) and electronic computing devices were invented. A most important event in the history of radioelectronics was the discovery of quantum electronics. Investigations carried out in this field by the Soviet scientists N. G. Basov and A. M. Prokhorov, as well as works of the American scientist Charles Townes, resulted in their winning the Nobel prize in 1964.

The extent of progress in the field of radioelectronics can be seen from production figures for electronic components, devices, machines, and systems.

The number of broadcast receivers in the Soviet Union, including wired remote loudspeakers, currently exceeds 60 million. During one year alone—1964—approximately 4.5 million radios and radio-phonograph combinations were produced. High-quality ultrashort-wave broadcasting with frequency modulation has also been developed successfully in the Soviet Union. Further building of ultrashort-wave broadcasting stations and increases in production of radios for the reception of the programs transmitted by these stations will lead to a sharp improvement in the servicing of the Soviet people by broadcasting. Even now, high-quality ultrashort-wave broadcasting reaches a total of more than 75 million people in the Soviet Union.

The production of television sets was increased in the Soviet Union to meet a growing need in telecasting. The total number of television sets now in use is about 12 million. In 1964 alone, 2.9 million television sets were produced. The Soviet Union television network covers a territory with a population of more than 90 million people. It is anticipated that by 1980 telecasting will reach all of the Soviet Union.

In the Soviet Union much attention has been paid during the last few years to the development and production in great quantities of transceivers and amplifiers for radio-relay communication lines. The number of these lines increased by 19 times during the last ten years. About 60 republican and regional centers as well as a great number of other cities now see television programs originating from Central Television in Moscow thanks to a network of radio-relay and coaxial-cable communication lines. These lines have steadily extended the exchange of television programs between the cities of the Soviet Union and other countries by including the Soviet television network in the international "Intervision" and "Eurovision" networks.

In addition to the manufacture of broadcasting and television equipment, much attention has been paid in the Soviet Union to the development and production, in required quantities, of various electronic and radioelectronic equipment for providing automatic industrial processes, for carrying out space research investigations, for study of the structure and properties of different substances, for various biological and medical investigations, and for investigations in many other significant fields of science, engineering, and industry.

Radioelectronics developed very rapidly in the Soviet Union just as it did in many other countries. For example, by the end of 1964, there were 17 million television sets in Japan while that country's total telephone network covered about 15 million channel-kilometers. Great sums of money are invested in the United States for the development of radioelectronics. This can be illustrated by the fact that the price of only one electronic computing machine—type CDC-6600 performing 3 million mathematical and logical operations per second—is very high. Radioelectronics has also developed successfully in Hungary, Czechoslovakia, and many other countries.

Radioelectronics has played an important role in space research. Some of its applications in space research are: complex mathematical computations of trajectories for the flights of spaceships and interplanetary stations through use of high-speed electronic computing devices; high-precision launching guidance to precalculated orbits; precise determination of spaceships' locations and flight speeds; transmission of telemetry information to the earth from aboard spaceships; transmission of television images from space; and radiotelegraphy and radio telephony communications.

Modern radioelectronics permits the transmission and reception of radio waves over long distances. Radio communication over a distance of more than 100 million kilometers was established during the flight of "Mars 1," the automatic interplanetary station. During the radar observations of the planet Jupiter carried out in the Academy of Sciences of the U.S.S.R. by Academician V. A. Kotelnikov and his co-workers in 1964, the radio waves sent from the earth reached the surfaces of the planet and returned to the earth covering the distance of 1.2 billion kilometers. Modern radiotelescopes receive the natural radio emission of space objects spread from the earth over a distance of more than six billion light-years.

Radioelectronics also played a significant role in the Soviet Union in the space flights of cosmonauts Yu. Gagarin, G. Titov, A. Nikolaev, P. Popovich, V. Bykovsky, V. Nikolaeva-Tereshkova; the first space team of cosmonauts V. Komarov, K. Feoktistov, and B. Egorov who performed, on the 12th of October, 1964, unprecedented scientific experiments on the Soviet spaceship "Voskhod"; and also the famous flight of cosmonauts A. Leonov and P. Balajev on the 18th of March, 1965.

Radioelectronics has also played a great role in the investigations of atomic nuclei, of elementary particles of substances of the virus world, and of bacteria. Without superpower radioelectronic devices, such as proton accelerators, it would have been impossible to achieve such remarkable results in the investigations of many properties of atomic nuclei and elementary particles. Similarly, it would have been impossible without electron microscopes to achieve successful results in the investigations of the properties of various viruses and bacteria. Radioelectronic devices have also played a great role in the improvement of public health and in medical progress. Very complicated economic and scientific-technical investigations and works are carried out at present with the help of electronic computing machines.

The question now arises: Why is it that radioelectronics, its methods, devices, instruments, and systems, has found such wide and successful application in so many different fields, and what are the prospects for its development and importance for mankind in the near and more distant future?

The reasons for such wide application of radioelectronics lie in the fact that radioelectronic devices, instruments, and systems possess so many valuable properties. A few of them are: high-speed operation, flexibility and universality, precision and sensitivity, miniaturization, and ease of designing complex devices and systems by using combinations of many primary components or elements. Many of these properties can be explained by the flexibility of electric energy in general, and, in particular, by its very high flexibility in the region of high and super-high frequencies of electromagnetic oscillations. High-speed operation comes about because radioelectronic devices, equipment, instruments, and systems are based on the use of electromagnetic processes, the rate of which in the long run is determined by the velocity of electromagnetic energy propagation—the most rapid velocity available in nature.

In order to answer the question about the future of radioelectronics, some general trends in its development should be considered.

The first general trend is the mastery of higher and higher frequencies, i.e., shorter and shorter electromagnetic waves. Long and superlong waves for radio communication and radio navigation, medium and short waves for broadcasting and communication, and ultrashort waves for television and radar purposes are widely used at present.

Intensive experiments are being carried out to master millimeter, submillimeter, infrared, optical, and shorter electromagnetic waves.

A second trend is the improvement in quality of radioelectronic components (tubes, semiconductor devices, and other elements), units, devices, equipment, and systems. In this case we bear in mind such qualitative indexes as high-speed operation, reliability, miniaturization, sensitivity and precision, noise immunity, range, etc.

A third trend is the broad penetration of radioelectronics into all aspects of life in our society.

Finally, there is the technological improvement of radioelectronic components, units, devices, equipment, and systems, as well as the ability to produce them in great quantities.

These four trends have been present through all the periods of development of radioelectronics in the past and will continue for many years in the future.

The history of radioelectronics shows that not only have there been evolutionary changes but there also have been fundamental qualitative jumps both in properties of radioelectronic devices and systems and their practical applications. The invention of the triode by Lee de Forest; the discovery of triode properties to generate electromagnetic oscillations by the German scientist A. Meissner in 1913; the invention of the transistor and electronic computing devices; the discovery of quantum electronics—these are all examples of the fundamental periods in the development of radioelectronics. Mastery of each frequency band is usually accompanied by a fundamental change in the properties and nature of radioelectronic equipment. The mastery of short radio waves made possible long-distance radio communication while the mastery of ultrashort waves gave rise to radar and high-quality television.

There is no doubt but that more qualitative jumps will be in the offing for the development of radioelectronics in the future. These changes, to a certain extent, can and should be foreseen.

Quantum electronics with coherent lasers giving electromagnetic emission in an extremely narrow frequency band of the optical range, high-precision frequency and time standards, low-noise masers, and other devices and methods open up great possibilities for the solution of many difficult modern problems.

Devices providing the precision of "electronic clocks," which corresponds to an error of one second during several thousand years, are constructed with the help of the methods and means of quantum electronics. Recently, quantum electron generators have been designed that use hydrogen for the working substance. These generators provide a high relative precision of about 10^{-13}. The frequency of such generators according to the measurement made by N. Ramsey is equal to 1420 405 751.800 ± 0.028 c/s.

More widespread use of semiconductor devices and the creation of miniature radioelectronic devices are important trends in the development of radioelectronics during the last decade.

In contrast to ordinary devices, which are assembled from different interconnected components, thin-film solid-state schemes are used in miniature and superminiature designs. Thin films are prepared from resistive, dielectric, magnetic, semiconductor, and other materials, and have thicknesses of tenths or hundredths of microns. For these devices, sputtering, chemical deposition, photolithographic, and other methods are used for applying films onto a substrate.

By using solid-state schemes (e.g., semiconductor plates, the separate sections of which act as electric resistors, capacitors, inductors, diodes, transistors, connectors, and other elements) it is possible to reduce not only the number of contact interconnections but also to ensure significant increases in element density and consequently to achieve considerable decreases in size and weight.

A number of investigations with the aim of studying the basic possibilities and limitations for the design of technical systems of processing and storage of information with the help of the atomic–molecular level of processes were carried out recently in the Soviet Union. For example, the Soviet scientist M. S. Neiman showed that an equation

$$\frac{N \cdot MBQ}{k} = C$$

takes place between the degree of reliability N, the degree of microminiaturization M, and the degree of high-speed operation B, where Q is the quantity of parallel channels for processing of independent series of information, k is a multiplier taking into account the favorable action of catalytic agents, and C is a constant. These and other scientific investigations are the basis for a new field of radioelectronics—radioelectronics of superminiature, reliable, and high-speed operating systems in which not the tubes and transistors but rather the atoms and molecules of some substances, and the electromagnetic wave interacting with them, are the primary elements of the equipment.

Fundamental laws for the development of modern science should be taken into consideration in order to predict correctly the further progress of radioelectronics. One of these laws is the interaction of various sciences and fields of engineering. Creation of electronic computing devices from this point of view is the result of a blend of the ancient science of mathematics with the young science of electronics. Astronomy and radioelectronics combined to give rise to radioastronomy. Radioelectronics works successfully not only with such natural sciences as physics, chemistry, and biology, but also with some social sciences such as economics, linguistics, and statistics. One of the remarkable fields of modern science of our century is cybernetics, which successfully coordinates radioelectronics with many other sciences in nature, society, and thinking.

One of the important problems suggested by Soviet scientists is the creation of a common system of optimal planning and control of a national economy on the basis of wide application of economic–mathematical methods and the use of electronic computing techniques. The solution of this problem, together with the wide introduction of radioelectronics into all fields of the national economy, permits the more expeditious use of all the resources of the country—labor, natural, and technical—to ensure more sufficient satisfaction of the rising material and cultural needs of the Soviet people. Coordinated efforts should be made by scientists working in the fields of economics, mathematics, radioelectronics and electronic computing techniques, cybernetics, and many other sciences, and fields of engineering in particular, for the solution of this problem.

Achievements of modern radioelectronics, atomic power engineering, automatics and telemechanics, cosmonautics, chemistry, physics, mathematics, cybernetics, and many other sciences and fields of engineering ensure the immense growth of man's power over nature.

By reasonable use of the achievements of modern science and engineering there are many possibilities for increasing considerably our material and cultural values and for raising significantly the living standard of all members of the society, for reducing the working day, and for satisfying more completely our material and cultural requirements. But unreasonable attitudes toward modern engineering can bring on unimaginable disasters for the enormous masses of people. That is why Soviet scientists, as well as all the Soviet people, who carried the burden of the last war, defend and will continue to defend the cause of peace and the policy of peaceful coexistence of states irrespective of their social system.

Radioelectronics brings together the peoples of all continents and states. Use of radio-communication broadcasting and television opened up vast opportunities for an exchange of information and achievements in different fields of science, engineering, and culture. Soon the problem of global television will be solved with the help of artificial earth satellites and the television relay stations located in them. In the not very distant future there will be an opportunity to create—with the help of artificial earth satellites and radioelectronics—a system for carrying out observations and study of meteorological data and their transmission together with the data obtained by the ground meteorological stations. This "global" system should cover our entire planet. Processing of all this meteorological information will be done by electronic computing machines. Inexhaustible opportunities lie in the further development of cybernetic radioelectronic machines, devices, and systems. Even now electronic machines free us from the laborious and tiresome work connected with carrying out a lot of calculations. Electronic cybernetic machines are capable of performing not only the computations but also complex logic operations. Many of these machines are capable not only of operating according to a precisely given program but also are capable of learning and adapting. Many investigations during the last few years were carried out in the field of designing machines for the recognition of patterns and other machines intended for the improvement of the efficiency of the learning process in particular.

Some authors in the West believe that in the far future there will be electronic self-adapting cybernetic machines containing more primary elements than there are brain neurons or nerve cells. It has been said that these machines will "think" better than a man, and that even a society of such machines can be organized. Some authors say that this society will become hostile to the society of people and then there will be a disaster which will end with the death of mankind.

Such a *pessimistic point of view* is unfounded. Cybernetic electronic machines are designed and constructed by the people, and the general trends in their development still depend on people. These general trends in turn are determined by the economic and cultural requirements of the society which, in the long run, define the directions of the development of science and engineering in general and cybernetic techniques in particular. Scientists and engineers, recognizing these laws of development, will carry out research and design in the field of cybernetics in those directions that will not be harmful to the welfare of mankind.

On the other hand, the application of more and more complex structural schemes of interconnections of great numbers of primary elements, improvements in high-speed operation, application of superminiature units, use of complex connections of feedback, both determinate and random, and improvement of self-programming and self-adapting properties will lead to qualitative changes in the properties and applications of the whole future cybernetic technique. It is possible that these changes will be of such decisive importance that they will give rise to some new forms of matter motion differing from the presently available ones.

The kinds of changes and how rapidly they take place will depend on scientists. It may be that when the conditions under these new forms of matter motion are created they will lead to the state where their properties will turn out to be fundamentally different from the properties of inanimate nature, living organisms, and the properties of a man and his consciousness.

The English scientist Maxwell in the last century generalized the laws of electromagnetism experimentally worked out by Faraday. Maxwell prepared and solved his famous equations and predicted the theoretical existence of electromagnetic waves, defined their basic properties, and determined in particular that these waves should be propagated in free space with the velocity numerically equal to the factor $c = 300\,000$ km/s.

By carrying out the investigations of complex structural systems with a great number of units and with constantly rising structural difficulties the scientists and engineers of the future working in the fields of radioelectronics and cybernetics may prove the possiblity of the existence of new forms of motion.

Future electronic cybernetic machines will help man to solve problems and questions that are fantastically difficult for us now.

It is possible with the help of radioelectronics to solve the problem of maintaining contact with the intelligent beings of other worlds.

This will be a remarkable and exciting event in the life of mankind and will simultaneously be a starting point for the origination of many new sciences and also of practical applications, the significance of which is rather difficult to appreciate now.

The high standards of Soviet science and engineering and their rapid progress, and the tendency of Soviet scientists to cooperate with the scientists of all countries in achieving the noble aims of scientific progress and the improvement of the well-being of humanity, are actions that will hasten the solution of the most difficult and urgent problems of the present day.

It is obvious that mankind has now entered a period of scientific–technical revolution and is on the threshold of new great discoveries. There is no doubt that radioelectronics will occupy its proper place in these achievements of a human mind and bring much good to the people of our planet.

Submarine telephone cables

Although the transmission of telephone messages by transatlantic cable has been a dream since the 1870s; economic and technological obstacles prevented it from becoming a reality until nine years ago. How these obstacles were overcome is the subject of this article

E. T. Mottram *Bell Telephone Laboratories, Inc.*

The concept of telephone communication over trans-ocean submarine cables dates back to the latter half of the 19th century. As early as 1879 Alexander Graham Bell, using his recently invented telephone instruments, tried to talk over one of Cyrus Field's telegraph cables, but without success. The need for bandwidths far greater than the few cycles of the early cables was not identified until some years later, and it was the middle of the next century before the necessary disciplines had matured to a point where such projects became technologically feasible and economically attractive.

Many well-known names appear in histories of the efforts to "speed up" telegraph cables that would eventually yield the telephone cable. Heaviside, in the period 1885–1887, showed theoretically that the way to reduce the distorting effect of cable capacitance was to add inductance. By the turn of the century Krarup, a Danish engineer, had proposed the addition of continuous loading by winding soft iron wire helically along the length of the cable; Pupin proposed to achieve the same end by lump loading, adding inductance coils at regular intervals. Both schemes were used successfully.

In the early 1920s Dr. O. E. Buckley and his associates at Bell Telephone Laboratories took another large step forward by use of the newly developed high-permeability permalloy for continuous loading. With these cables telephone circuits were extended from the U.S. mainland to Catalina Island and to Cuba, and by the end of the 1920s a telephone cable across the Atlantic was receiving serious consideration.

The development of high-frequency radio for long-distance communication made such a cable economically less attractive, and the project was deferred. But by the end of World War II the demand for reliable communication had increased to a point where it could not be satisfied by the number of usable radio frequencies available. Further, ionospheric storms were causing interference, disrupting communication in some areas for long periods of time. Clearly any major improvement would have to come from a different medium; thus, interest in cables was stimulated once again.

Many developments that had evolved since the 1920s contributed to a more favorable outlook, and influenced the decision to lay a telephone cable across the Atlantic:

1. Polyethylene had been discovered and its properties as an insulator had been explored. High insulation resistance and a low dielectric constant, together with imperviousness to water, made it well suited to submarine cable applications.
2. Carrier techniques developed for land lines made it possible to transmit a large number of messages simultaneously over the same conductors.
3. Stable amplification with low distortion needed for the transmission of broadband carrier signals had been realized with the invention of the feedback amplifier.
4. Long-life electronic components, both active and passive, permitted amplifiers to be placed at sea bottom with a reasonable expectation of long periods of failure-free operation.

With the new tools in hand, development was undertaken which yielded the first transatlantic cable telephone service in 1956.

The problem

The transmission problem posed by a submarine cable system is identical with that of any land-based carrier system using amplitude modulation. But an ocean-bottom environment also imposes a need for a capability of withstanding hydraulic pressures as high as about 10 000 pounds per square inch for long periods of time and makes prohibitively costly any extensive adjustments or maintenance. The result is that equipment functionally similar bears little resemblance to land-based equipment.

Figure 1 shows in schematic form what happens to a signal in a carrier system. A high-level complex signal is placed on (in this case) a coaxial cable. As it progresses along the cable, the signal is attenuated until it approaches the noise level of the system. It is then amplified and transmitted along the next section. Limits of noise and overload are indicated, and the margins allowed must be adequate to permit performance objectives to be met over the life of the system.

Because the loss of coaxial cable is roughly proportional to the square root of frequency, the gain characteristic of the repeater amplifier is not flat. It must be shaped to match cable loss over the entire transmitted band of frequencies. Should repeater gain be greater than the loss of a length of cable, the resulting misalignment is shown by the dotted curve. After passing through a number of repeaters the signal would overload the amplifier, cause intolerable distortion of the signal, and shorten the life of the active devices. Should the reverse be true, the signal might drop into the noise and cause errors in transmission and loss of intelligibility.

To give a sense of the magnitude of the problem, a typical repeater amplifier provides amplification of about 50 dB at its top frequency. This process is repeated 180 times to provide 128 two-way message channels between the New Jersey coast and Cornwall, England, a distance of 3500 nautical miles. Losses of 9000 dB are compensated by amplification. In terms of nonlogarithmic numbers, the signal is amplified by a factor of 10^{900}. This amplification must be accomplished with a minimum addition of distortion and noise, and misalignment must be held within comparatively narrow limits (10 to 15 dB) over the life of the system.

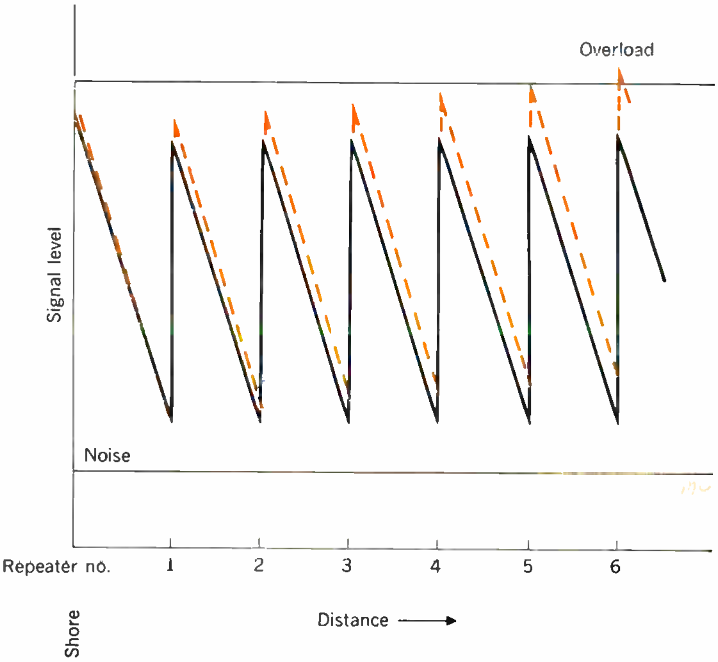


Fig. 1. Carrier signal-level diagram.

Objectives

Basically, the objective of a cable system is to provide a highly reliable transoceanic facility capable of transmitting a complex signal at a quality level at least comparable to the land plants of the continents connected.

In normal use the signal sent over a cable may have traveled several thousands of miles before reaching the cable system and may go thousands of miles more on the far continent before reaching the terminal, in the process suffering in the land facilities most of the degradation usually considered "tolerable." Not much can be done about distortions accumulated in the land plants, but it becomes desirable to hold any additions to a minimum.

The noise objective used for American-designed broadband systems is 35 dBa0* for a 4000-mile telephone circuit. This is about 3 dB better than is usually found in the land plant, and because addition tends to be on an rms basis, the contribution of the cable is small.

This objective must be based on a load appreciably higher than is found in conventional AM carrier systems. A tendency for people to speak louder over long circuits has a measurable effect on system loadings. In addition, because the facility is costly, it is economical to make more efficient use of bandwidth by stacking message channels more closely together. This situation, together with the higher activity factor that results from the application of TASI (Time Assignment Speech Interpolation), further increases loading.

An increase of about 6 dB compared with CCITT (International Telegraph and Telephone Consultative Committee) standards for an equivalent eight-group carrier system is shown in Table I.

Systems

A great number of systems can be devised to meet a given set of objectives. The curves of Fig. 2 describe systems of varying bandwidth using cables with several values of loss. The repeaters shown are the number needed for a 3500-mile system.

In a system of given span and cable size, the cable cost will be constant regardless of the number of channels. Because cable attenuation increases as $\sqrt{f}$, the number of repeaters also increases approximately as a square-root function. The resulting rapid decrease in cost per channel mile is illustrated in Fig. 3.

This function approaches a limit as the number of repeaters increases to a point at which their noise contribution is significant. Before this point is reached in long systems, however, it usually becomes economical to use a lower-loss cable of larger diameter. The increased cost of the cable will be more than offset by the reduction in the number of repeaters.

Several factors influence the selection of system capacity. Clearly the fact that the incremental cost of additional

* The dBa0 is a measure of the relative interfering effect of noise frequencies at a zero-level point. It is a weighted measure referred to a 1000-c/s tone 90 dB below 1 milliwatt.

channels is small suggests systems of large capacity. There are several offsetting factors, however, that must be considered if the size is to be optimum.

It must be assumed that any system has a finite life, which will vary depending on the objectives and the skill and experience of the developers. Likewise, any route will have a rate-of-growth potential. If this rate is low, a high-capacity system may not be filled within its lifetime, and unused capacity is expensive no matter how little it may have cost. Even when systems can be expected to be filled eventually, the rate may be so low that it may be more economical to add channels in smaller units and defer part of the expenditure until additional capacity is needed.

A second factor is reliability. The system with greater capacity will use more repeaters, and the probability of failure will increase in direct proportion to the number of repeaters. Where reliability is of prime importance the smaller system will have an advantage in achieving redundancy at an earlier date, thus permitting division of the load, or at least making good on another facility.

Most recent cables have been of the two-way type. The transmission band of a single cable is split in two by so-called cut-apart filters. The upper band is used for transmission in one direction, and the lower band for the opposite direction. Although this technique increases the number of repeaters and the filters increase the cost of a repeater, the increase is more than offset by avoiding the use of a second cable and may result in lower channel-mile costs. Higher-capacity systems reported under development will also use two-way cables, although capacities may be anticipated in the future that will require one-way cables and exploit the cut-apart region wasted on two-way cables, as is done on most of the major trunk routes on land.

I. Cable loading

	Submarine Cable	CCITT
Average talker volume, VU	—10.8	—12.0
Standard deviation, dB	5.8	5.0
Activity	0.75	0.25
RMS power per channel, dBm	—9.6	—15*
RMS power per group, dBm	+2.4	—4.2
RMS power per band, dBm	+11.5	+4.8

* Includes an allowance for power of signaling tones.

The earliest transatlantic systems used one-way cables for quite different reasons. In the interest of greatest reliability, the number of electronic components in underwater repeaters was held to a minimum. The original one-way repeaters required 60 components as compared with 200 or more typical of current two-way units. An armored coaxial cable was used. The combined protection and strength member was a carryover from telegraph cables. One of its characteristics was that, under the tensions of laying, the spiral armor tended to untwist, which caused the cable to rotate. Any abrupt discontinuity which interfered with this rotation could cause damage to the cable. Further, the cable was to be laid by existing telegraph ships, with relatively few modifications. The solution was the long, snakelike, flexible repeater, which appeared as a small increase in diameter of the cable and could be handled in the same way as cable while it was in the ship's storage tanks and when it was passing over the ship's cable engine drum and sheaves. The small diameter minimized the interference with the twist of the cable but restricted the number of components that could be housed. The form factor necessitated long leads, which imposed limits on the top frequency when greater channel capacity was desired.

Cable design

Early cable was based on telegraph designs, with the addition of a concentric return conductor to form a coaxial structure. The multitape structure shown in Fig. 4(A) was carefully designed as to length and direction of lay to minimize change and permanent damage as the cable was bent around drums and sheaves and twisted

Fig. 2. Number of repeaters vs. top frequency for several sizes of coaxial cable, 3500-mile system.

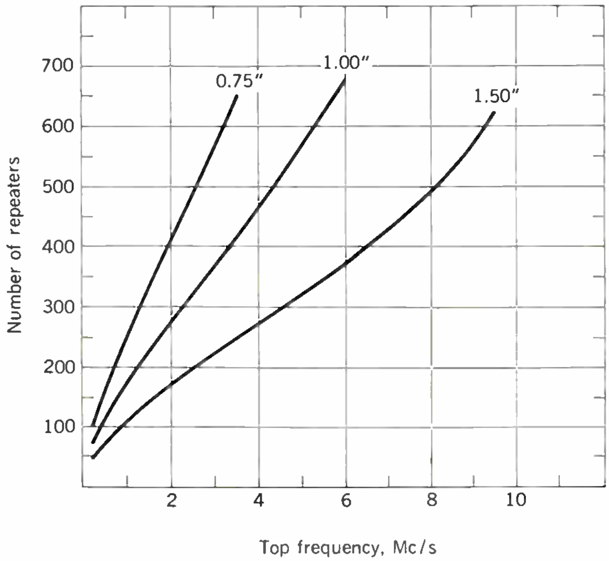
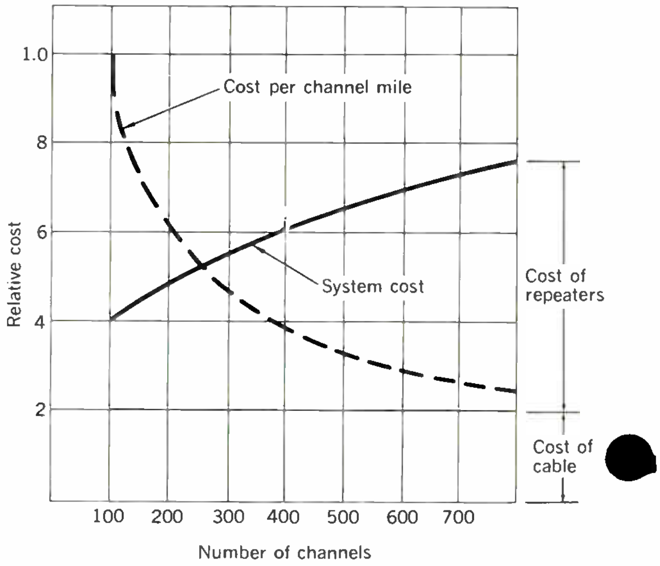


Fig. 3. System cost as a function of number of channels.



under stress. The teredo tape and the jute served as "protection"—a carryover from telegraph experience.

Once the success of the early cables had justified further development work, several shortcomings received attention. Early cables of transatlantic length have decreased in attenuation by 15 to 16 dB since they were laid. Under ocean-bottom pressures, the complex central structure tended to bond, decreasing the effective length of the conductor. It was also discovered that the tapes required to protect gutta-percha and paragutta insulation of early cables from teredos and other marine borers were not required by polyethylene; however, the jute used as bedding and to protect armor wires served to attract these borers. Changes obviously were in order.

As broader-band systems have been developed it has proved economical to decrease cable attenuation by increasing diameter. As the cable is scaled up in size, it is possible to place the strength member within the central conductor. Because of the use of high-tensile steel and the smaller margins needed in a strand protected from corrosion it is possible to produce within the same volume a lower-loss, lighter-weight cable. While a strength-to-weight ratio equivalent to the armored predecessor is maintained, the small diameter of the strength member causes a large reduction in twisting moment.

With similar objectives in mind the British General Post Office and Bell Telephone Laboratories have evolved different structures. The Post Office uses a reverse lay to achieve a "dead" strand, which is enclosed in a copper center conductor closed with a box seam. The polyethylene insulation is surrounded by a group of spiral aluminum tapes. An outer wrapping of overlapping turns of aluminum foil insulated from the return conductor by polyethylene tape is intended to reduce crosstalk interference between turns in the ship's tank during laying.

The Bell System has chosen the more compact strand permitted by a single direction of lay and the longer lay made possible by a welded enclosing conductor. The return conductor is a single copper tape with longitudinal overlapped seam. The relatively small twisting moment that results from the small diameter and long lay of the strength member is resisted by the series of concentric cylinders formed by the copper and plastics, which constrain twist to less than ¼ turn per 100 feet at breaking stress.

Both structures shown in Figs. 4(B) and (C) have proved to meet the needs and have made it possible to use larger diameter, rigid structures to house the repeaters, although stowing and launching remain a problem.

Measurements made over two years or more on both designs show no changes in attenuation greater than can be accounted for by shore-end temperature changes.

Repeaters and equalizers

The ability to lay large structures in deep water opens up many new possibilities. The large volume permits form factors suited to the amplification of higher frequencies.

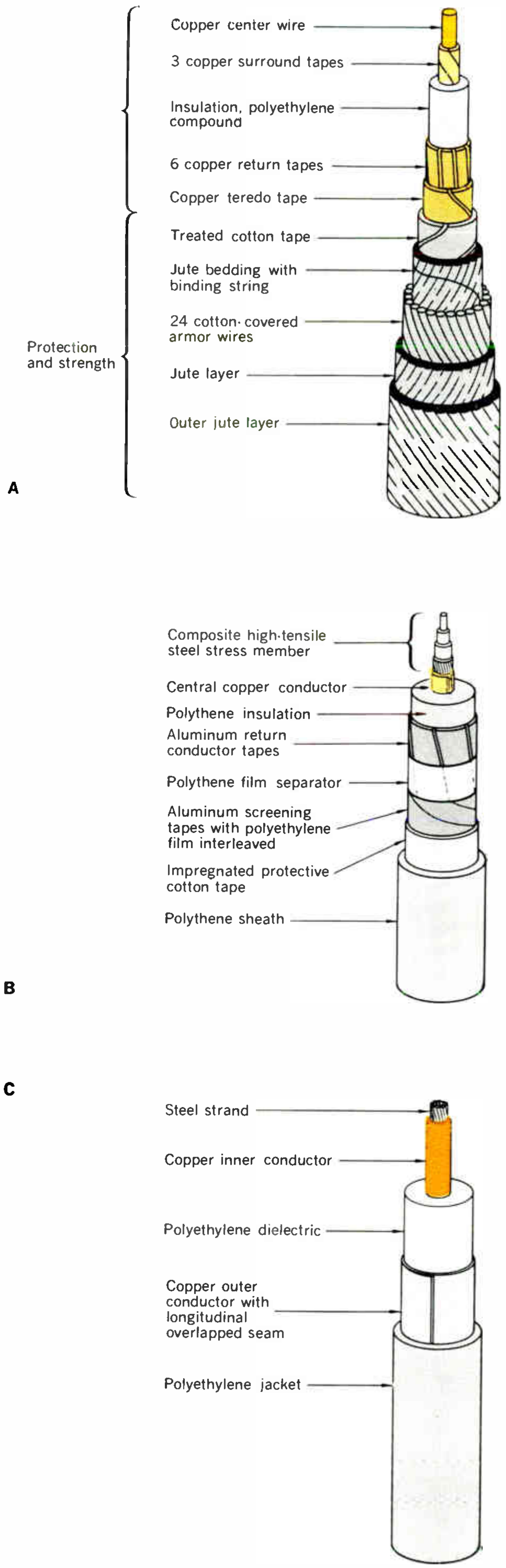

A
Copper center wire
3 copper surround tapes
Insulation, polyethylene compound
6 copper return tapes
Copper teredo tape
Treated cotton tape
Jute bedding with binding string
24 cotton-covered armor wires
Jute layer
Outer jute layer
Protection and strength
B
Composite high-tensile steel stress member
Central copper conductor
Polythene insulation
Aluminum return conductor tapes
Polythene film separator
Aluminum screening tapes with polyethylene film interleaved
Impregnated protective cotton tape
Polythene sheath
C
Steel strand
Copper inner conductor
Polyethylene dielectric
Copper outer conductor with longitudinal overlapped seam
Polyethylene jacket

Fig. 4. Deep-sea submarine telephone cable. A—0.620-inch armored cable. B—British lightweight cable. C—American lightweight cable.

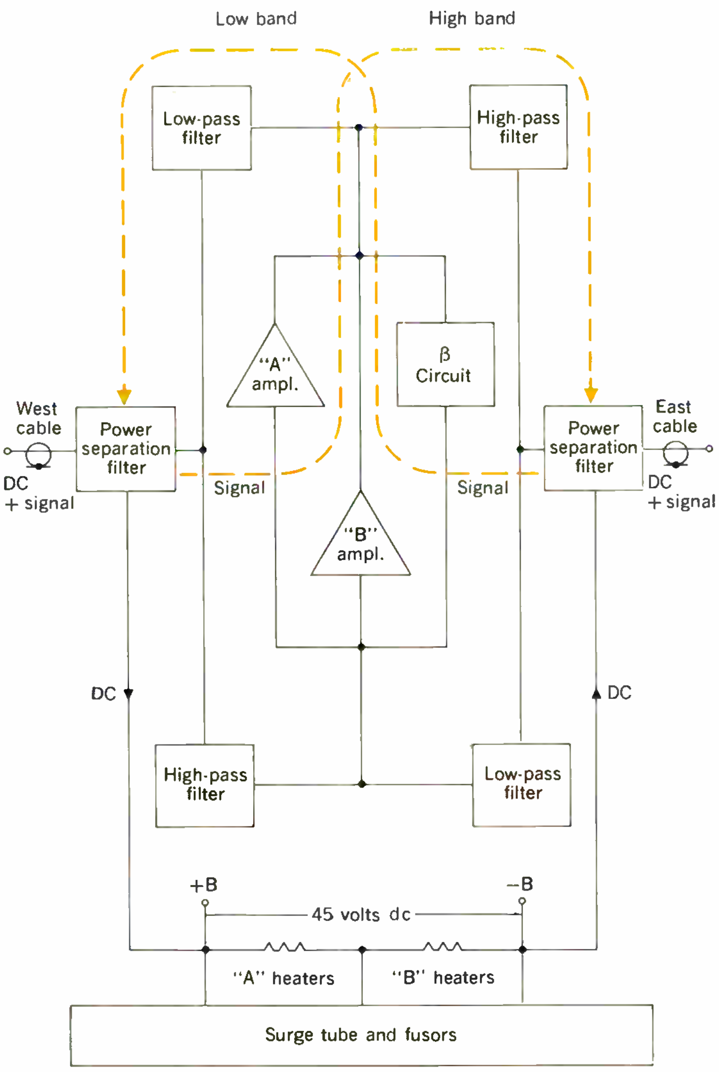


Fig. 5. Two-way repeater block diagram for SD system.

Space can be made available for larger components and greater spacings needed with the high voltages required by longer systems and closer repeater spacings. It is also possible to find room for the cut-apart filters of two-way systems and for redundancy to improve the reliability of systems using tubes having a higher figure of merit and closely spaced elements. All of these are to be found in a variety of systems currently in use.

Most of the repeaters installed during the last few years are two-way units using a single amplifier to amplify signals in both directions of transmission, as shown in Fig. 5. Electron tubes are the active devices and two amplifiers are used in parallel with a common feedback path.

Objectives for the feedback amplifier are conventional. About 30 degrees of phase margin at gain crossover and 10 dB of gain margin at phase crossover with smooth transitions in between are desirable. In a system designed for long life and many repeaters it is essential that advantage be taken of optimum phase relationships in the in-band feedback path to minimize the effects of aging of active devices on amplifier performance.

Gain of the repeater may be shaped by the feedback path of the amplifier, the power separation and cut-apart filters, or some combination. Examples of each can be found in commercial use. Out-of-band gain roll-offs serve an important function in preventing system noise loading.

The directional filters, which permit use of the same amplifier for both directions of transmission, have unusually rigorous requirements. Because they and the power separation filters provide several spurious feedback paths around the amplifier that would affect transmission performance, out-of-band losses must be maintained at a high value throughout while in-band losses, especially at the top frequencies, must be kept to a low value.

Although a coaxial cable inherently has a smooth characteristic, small differences between cable loss and repeater gain will exist. To avoid the use of a large number of components in each repeater to achieve a nearly per.'ect match, these deviations are usually allowed to accumulate over a number of repeaters and are then "mopped up" in equalizers placed at intervals along the cable. Reduction in the number of components results in lower cost and a higher degree of reliability.

In addition, the equalizer usually houses adjustable or switched networks. These are simple, smooth shapes which are used during cable laying to make corrections for the effects of temperature and pressure deviations encountered on a route.

Power

Repeaters are powered over the cable. A constant-current dc supply provides heater power; the drop across heaters provides the plate voltage. The signal being transmitted is superimposed on the dc voltage. At each repeater the signal and power are separated and reassembled by the power separation filters.

To hold voltages on repeaters to a minimum, long systems are normally powered from both ends. Because tube life may well be shortened by the thermal stresses that occur on heating and cooling, various forms of redundancy have been used in supplies to avoid power failures.

To hold current constant, voltage must be variable to aid or buck earth potentials. During periods in which aurora is present, voltages as high as 2500 volts have been recorded across the Atlantic, and in shorter links voltages as high as 4 volts per mile have been noted.

Because of the energy stored in the cable, precautions must be taken in case of cable damage. A break near a repeater may well result in surges as high as 60 amperes or more. Various protective devices have been used, including spark gaps, fast-acting gas tubes, diodes, and combinations of these. The essential considerations are speed of breakdown and the ability to carry high currents repeatedly.

Shore terminals

The shore terminal is the point at which signals are fed to and received from the cable. A wide variety of arrangements are available to meet the conditions found in both system and connecting links. A terminal that is typical in functions to be performed is that of a Bell System SD cable system.

The basic arrangement of the terminal is shown in Fig.

6. A broadband signal, which is a composite of 128 3-kc/s or equivalent channels, is derived by conventional frequency division multiplex.

The frequency allocations (Fig. 7) are shown for 128 channels in each direction of transmission. Pilots (96 kc/s), inserted in each group modulator and appearing on the high-frequency line at the frequencies indicated, are used for monitoring, equalization adjustments, and automatic switching to alternate terminal equipment. The nominal power level of each pilot is −20 dBm at zero-transmission-level points.

Two-order wire channels are provided in the gaps between supergroups. One of these channels is split so that part of its spectrum can be used for voice communication and part for teleprinter exchange (TWX).

In the transmitting path, this signal is shaped or pre-

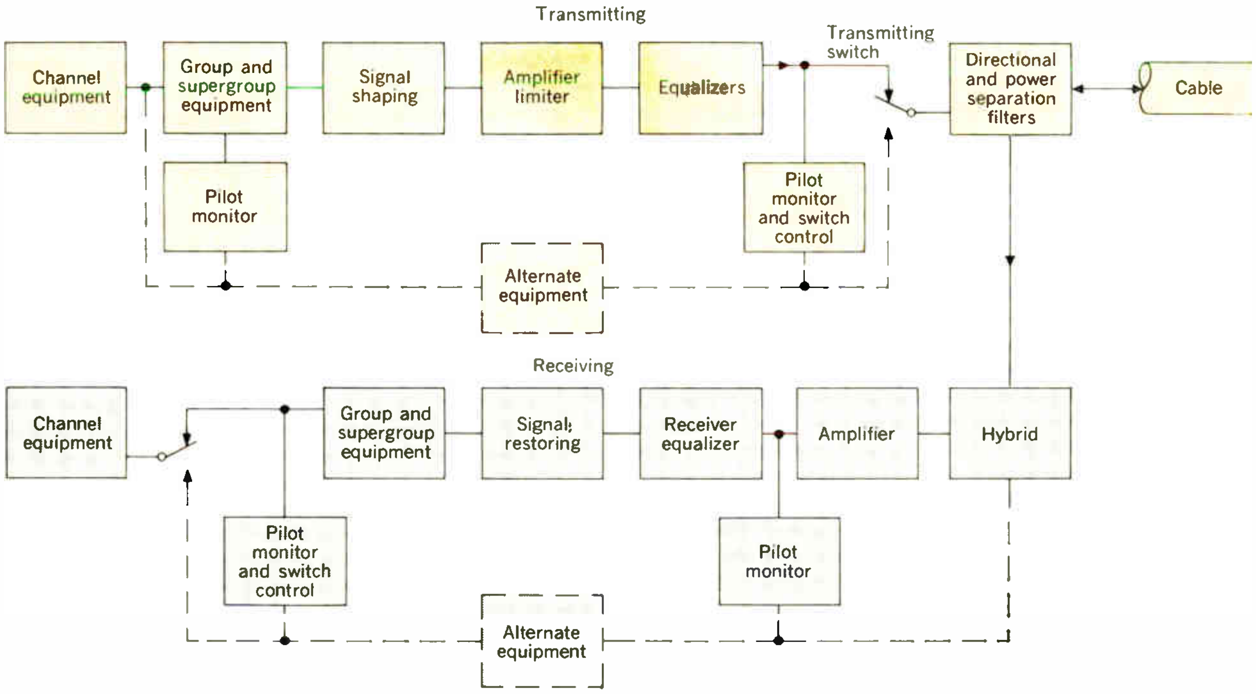
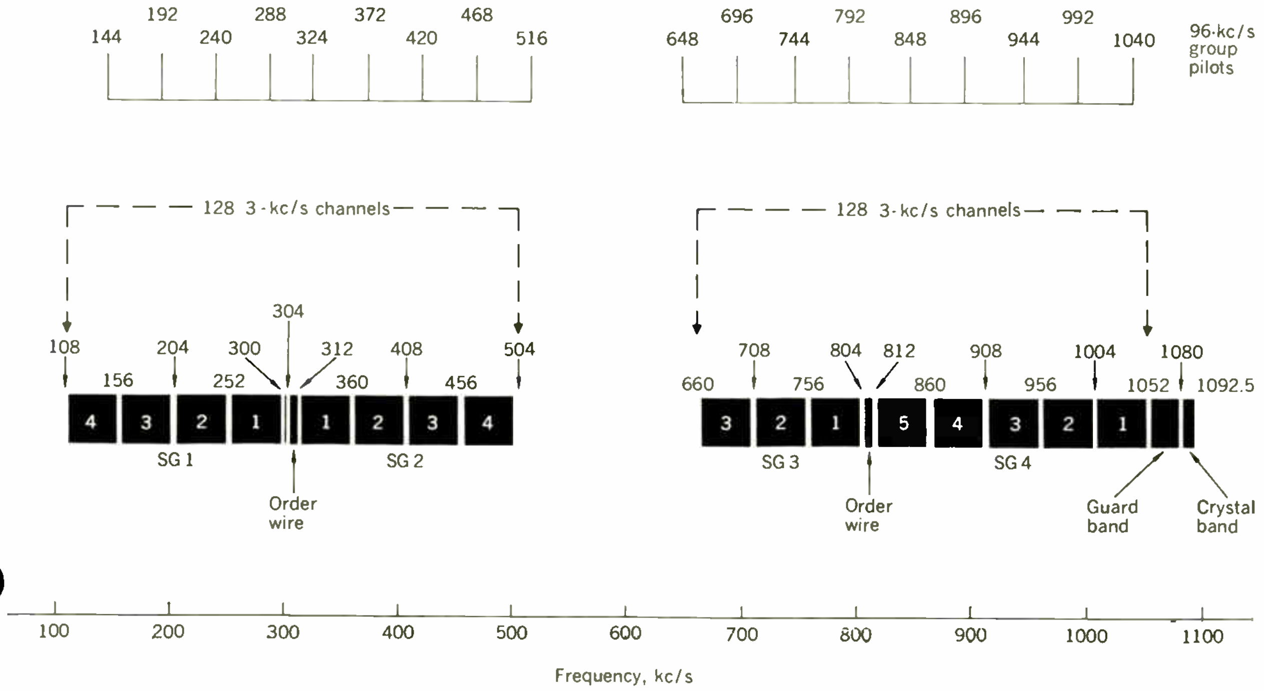
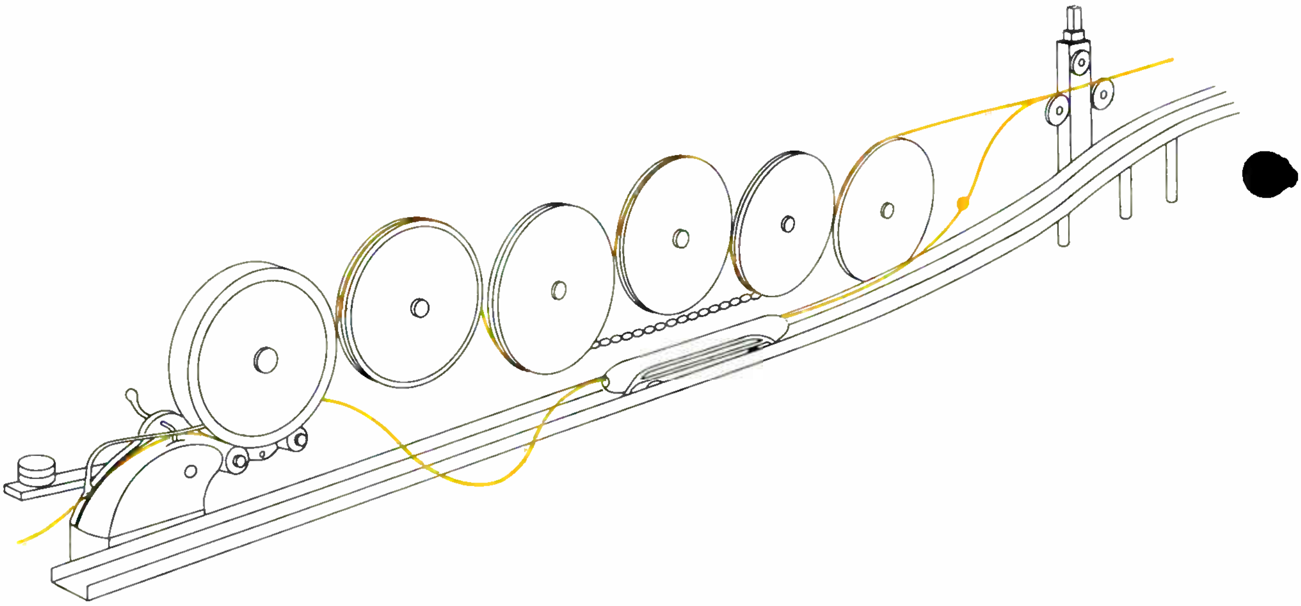


Fig. 6. Submarine cable terminal block diagram.

Fig. 7. Frequency allocation for submarine cable system.





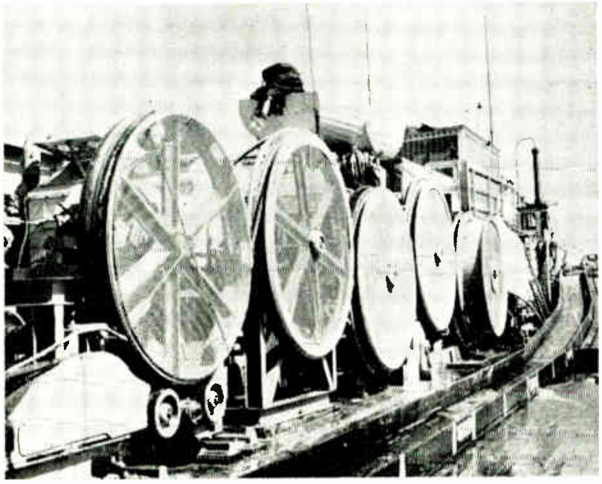


Fig. 8 (above and left). British GPO five-sheave drum-type cable engine.

distorted to equalize the noise in all channels and then limited to prevent overloading of the undersea repeaters by excessive inputs that might be caused by trouble conditions. Further equalization following the limiting amplifier is used to adjust for various misalignments in the undersea system. These equalizers, both fixed and seasonally adjustable, allow signal levels to be optimized from the point of view of signal-to-noise performance. At the receiving terminal, equalization and amplification are provided to make the frequency response of the system flat from end to end. A signal-restoring network is provided to complement the shaping introduced in the transmitting direction.

Duplicate equipment is provided for both transmitting and receiving terminals. Switching is under the control of a pilot monitor, which effects a transfer in case of a change in level of one of the pilots. The equipment also serves a useful function in making out-of-service maintenance possible.

Cable laying

During the past several decades, increasing interest in the science of oceanography has resulted in funds becoming available to organize the known facts regarding ocean-bottom conditions and to increase that knowledge as new and improved measurement techniques have been developed. Application of this knowledge to the selection of cable routes has been stimulated by the high loss incurred when a cable of large capacity fails.

The shortest cable route has obvious advantages in first cost; but a route that avoids a single failure may well save the cost of many miles of cable. Areas to be avoided include those in which there are rough ocean bottoms, steep grades, deep chasms, or strong currents; those in which bottom conditions are unstable because of earthquake or volcanic activity; and those in which the discharge of rivers may build up silt to a point of instability. Even more troublesome are man-made hazards, such as damage from otterboards of trawlers in areas of bottom fishing, and anchor damage in rivers, harbors, and roadsteads.

Bottom conditions on a proposed route are usually surveyed by means of a precision fathometer or depth indicator and, at the same time, checks are made of bottom temperature. Because cable has significantly large temperature and pressure coefficients, this information is needed for controlling misalignment during laying.

The data from a route survey require skilled interpretation. The beam width of a typical fathometer may be as great as 20 degrees; thus, at a depth of two miles the indication will be from a circle more than 4000 feet in diameter. The amount of detail available from present measuring techniques is far less than can be utilized effectively by modern laying techniques.

The entire operation of evaluating potential routes—balancing cost against hazard—is an exacting task requiring the exercise of sound engineering judgment.

Cable-laying equipment has also improved as new designs of cable and repeater have evolved. For many years telegraph cable, as it was laid, was passed around a large-diameter drum. The payout rate was controlled by braking the drum. Several turns of cable were wrapped around the drum to prevent longitudinal slippage. Heavy wedges, known as fleeting knives, were used to slip the cable laterally across the drum and keep it from running off at the side. In some of the early systems using rigid repeaters in relatively shallow water, the cable was laid first and the repeaters were added later in a repair type of operation.

As systems became longer, repeaters were laid with the cable. To get a rigid repeater past the drum, the cable was stoppered at a point outboard of the cable engine. Turns of cable could then be removed from the drum and placed back on the drum after the repeater had passed. The repeater, which had been moved along the deck on a wheeled dolly, was lowered overboard by a gantry. Stopping the ship to carry out this operation in rough seas and strong winds presents hazards to life and limb as well as to the cable.

To avoid these hazards in deep water, where cable tensions are high, the first transatlantic repeaters were flexible (or articulated) and could be passed around the drum without removing turns, and without slowing the ship to a speed at which control was poor.

For long systems using rigid repeaters better solutions were needed, and several schemes have found application. For example, the German ship *Neptun* uses a 12-foot drum type of engine and passes the relatively short 500-pound repeaters around the drum. This requires slowing the ship, and careful application of hold-back tension to insure that the repeater is held tight on the drum. Precise timing is essential.

The British General Post Office developed the scheme shown in Fig. 8. The drum-type cable engine has been modified by replacing the drum with five V-groove sheaves, three of which are geared together and braked. At each repeater a bypass rope is spliced into the cable, leaving the repeater and adjacent cable in a slack bight. The bypass rope is passed over the sheaves while the repeater and its bight are walked alongside the cable sheaves. Hydraulically operated sheaves are used to take up slack in the bypass. This arrangement also requires a carefully timed operation at reduced ship speed. It has been used successfully for laying the CANTAT, COMPAC, and SEACOM cables.

The AT&T Company ship *Long Lines* is equipped with a so-called linear cable engine. It consists of two sets of caterpillar tracks pressed together by a series of hydraulic cylinders to grip the cable. As a repeater passes through the engine, the caterpillar tracks are spread apart by the repeater body.

Repeaters are stowed in racks on deck. There is a slot in the top of the cable storage tank and along one wall, and retainers are used to hold the leads to and from the repeaters in the slot. Prior to the time for launching, the repeater is moved into the launching position. Cable is payed out of the tank. When a repeater is reached its lead pulls out of the retainer, the repeater is accelerated, and the trailing lead of the repeater feeds back into the tank to resume payout from that point. Cable and repeaters have been laid at speeds up to 8 knots without slowing the ship as repeaters are launched.

Cable ships are provided with precise transmission-measuring instruments. During laying, measurements are made from shore to the end of a group of repeaters terminated in an equalizer and known as an ocean block. As the cable reaches the bottom its loss, which was high while on shipboard because of the relatively high temperature, begins to stabilize at nearly its ultimate value. Shortly before the equalizer goes overboard, transmission performance is extrapolated to the block end and the equalizer is adjusted for optimum system performance. After adjustment the measuring equipment is connected to the output of the subsequent block and the process is repeated until all of the cable has been laid.

Reliability

A comprehensive discussion of the many facets of reliability is beyond the scope of this article. The subject cannot be dismissed without mention, however, because a high degree of reliability is a prime requirement of cable systems. Exceeded in importance only by the needs of manned spacecraft, in many ways it presents an even more exacting problem because of the requirements for longer life.

The performance that has been recorded on many cable systems has been achieved only by a broadly based effort. The design of parts has been based on similar parts that have demonstrated excellent performance in commercial service. After suitability for service, ease of manufacture and inspection have been controlling considerations in design. Materials used have been carefully selected for stability and long life.

The systems and circuit designs have been concerned with using parts well with adequate margins in circuit configurations that provide the greatest assurance of stability and long life. Manufacture has been called upon for careful handling of materials to provide perfect parts carefully checked in comprehensive detail. Installation must have been carefully completed on routes selected for minimum hazards. And, finally, operation must be carried out with continuing care and good judgment.

The performance record of these systems is a tribute to the skill and care of the members of the various teams. Systems in use vary in age up to about 15 years. Faults other than damage from trawlers have been few in number. The AT&T Company reported an amazing performance of repeaters in a wide variety of systems and locations. As of January 1, 1965, 3 billion component hours had been accumulated without a single component failure.

It is not surprising that, with such performance, systems have increased rapidly in number and length. Service on the first transatlantic cable route was initiated in September 1956. There are now in service more than 30 000 route miles of cable to provide more than 4 million telephone message circuit miles. Repeaters and cable are being manufactured for additional installations in 1965 and still more are planned for the future.

BIBLIOGRAPHY

Bullington, K., and Fraser, J. M., "Engineering Aspects of TASI," *Bell System Tech. J.*, vol. 38, Mar. 1959, pp. 353–364; also in *Elec. Eng.*, vol. 78, June 1959, pp. 639–643.

Special Issue on the SD Submarine Cable, *Bell System Tech. J.*, vol. 43, July 1964.

Special Issue on the Transatlantic Cable. *Ibid.*, vol. 36, Jan. 1957.

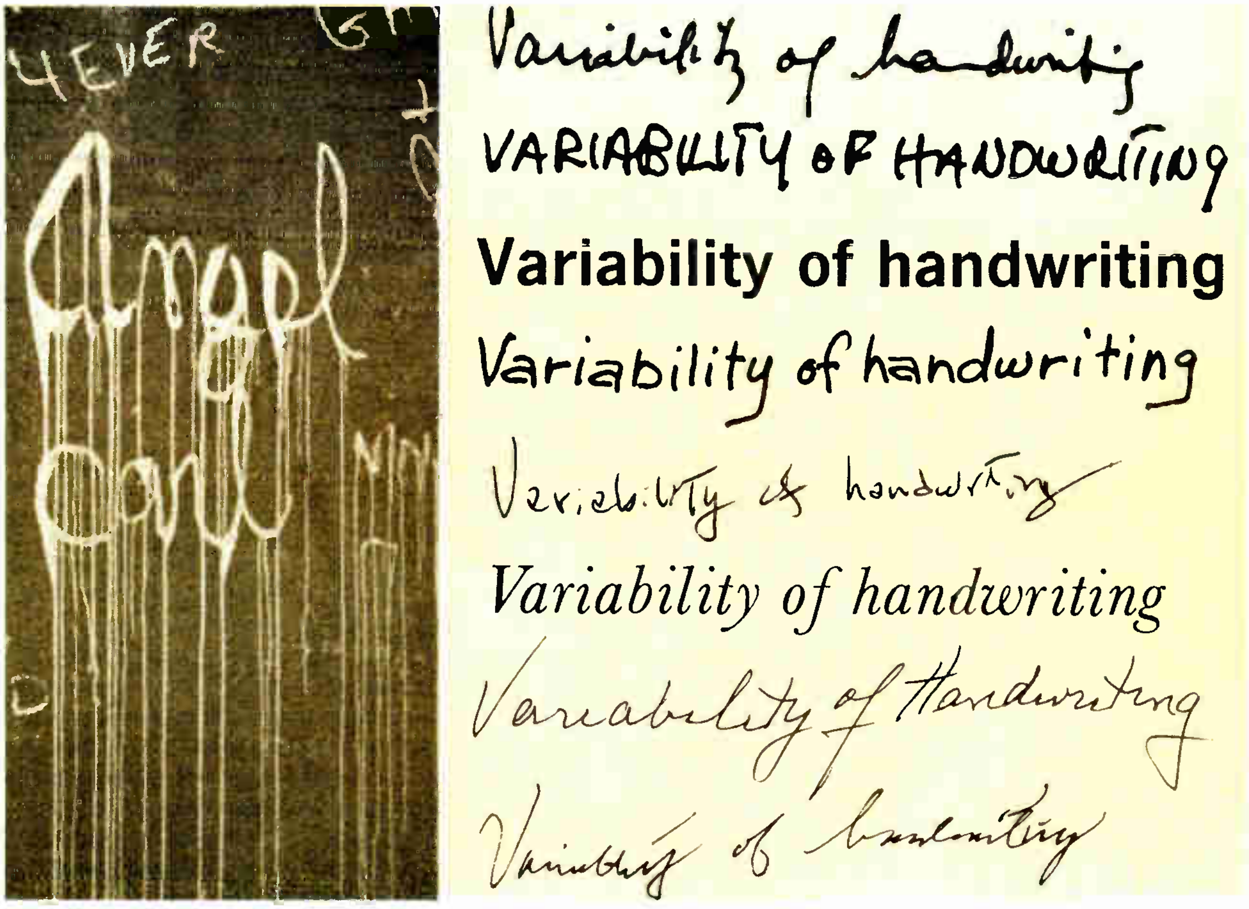
Special Section on the Anglo-Canadian Transatlantic Telephone Cable (CANTAT), *Proc. IEE* (*London*), vol. 110, July 1963, pp. 1115–1164.

Machine recognition of human language

Part III—Cursive script recognition

As with automatic speech recognition, attempts to mechanize the recognition of natural handwriting must face problems inherent in highly variable continuous patterns. Several different approaches to these problems are outlined

Nilo Lindgren *Staff Writer*


4EVER
Angel
Carol
Variability of handwriting
VARIABILITY OF HANDWRITING
Variability of handwriting
Variability of handwriting
Variability of handwriting
Variability of handwriting
Variability of Handwriting
Variability of handwriting

One might imagine, if one were not told otherwise, that the stimulus to research on automatic handwriting recognition stemmed largely from the research and achievements in automatic character recognition. Certainly, over the past decade the momentum of character-recognition developments has been impressive, and progressively more difficult problems have been undertaken. Where not so long ago machines were reading only machine-printed alphanumeric characters of fixed type faces, printed with special inks, and so forth, there are now in existence machines that will read, more or less satisfactorily, multiple, intermixed fonts at high speeds. Although the first optical character reader was installed commercially just ten years ago, in a Labor Department report it was estimated that 100 such machines were already in operation by 1963 and that as many as 300 more would be in operation by 1968.[1] As to the magnitude of the continuing research effort,[2] a rough index can be obtained from an awesomely large state-of-the-art report assembled by the National Bureau of Standards several years ago. It includes, in addition to a comprehensive review of all aspects of this field, a 549-item bibliography.[3] In recent years, there has been increasing research attention to automatic recognition of hand-printed and handwritten letters and numerals employing contour tracing and other techniques.[1-9]

With all this burgeoning activity, then, it would seem a logical step to undertake research on the next order of difficult problems, those inherent in "natural" connected handwriting. However, according to the pioneers in this field, the impetus to get going on handwriting research came from the other direction. Leon Harmon of Bell Telephone Laboratories relates that he first became interested in the problems of cursive script recognition because such research might throw some light on the more difficult problems associated with automatic speech recognition.[10] Murray Eden at M.I.T. states that there is very little background work in character recognition that would lead to cursive script studies. Rather, he says, it was the studies going on in speech (especially those involving generative grammar) that led to the idea of pursuing similar studies on handwriting generation, because there were very similar problems involved in both script and speech.[11]

However, in sharp contrast to speech research, those who have seriously undertaken automatic handwriting recognition are very few in number. And even among these few, none yet have attempted to deal with what legitimately can be called natural handwriting, that is, handwriting not in some fashion restricted.

Handwriting and speech

Undoubtedly one of the basic linguistic processes is the transformation of speech to writing. Both of these forms of verbal behavior display interesting similarities, as well as differences. Although their physical manifestations are different, and although different sets of muscles and body regions are involved in the execution or articulation of the appropriate gestures, both forms exhibit a great variability in their patterns, and challenge the researcher with difficult problems of determining the invariants in such patterns.

There is in handwriting, as there is in speech, the problem of individual variability. Both in speech and in writing, individuals manifest idiosyncratic differences that must be taken into account by a machine that intends to "recognize." And because handwiiting presents a more or less continuous signal, it too raises the specter of the segmentation problem.

Although handwriting research is relatively new (most work seems to have started around the beginning of this decade), its proponents appear already to have arrived at the same conclusion that speech researchers took many more years to formulate and accept—that a machine meant to recognize natural handwriting must be armed with a knowledge of linguistic constraints (context).

Even with very legible handwriting, machines are bound to make mistakes, which they will need contextual information to resolve. And if a machine were to attempt to read scrawled or scribbled sentences, it would be lost without such information. To become even more hypothetical, what would happen were a machine to be confronted with a "noisy" field as in Fig. 1 (left), in which normal orientation is missing, and in which there is no certainty that a "message" is inscribed in the space at all? Here, at last, it would seem, is a safe corner where those humanists who wish to preserve their innate superiority over the machine will prevail.

Although the physical signals of speech and handwriting are different, the problems they present for machine recognition beyond the primary levels (acoustic features in one case and visual features in the other) appear to be the same. That is, beyond the primary level, both must relate to the same grammatical structure of the language in question. At some level in the brain, the abstract representation meant to govern the production or perception of speech and the abstract representation meant to govern the production or perception of handwritten words must be in a one-to-one relation. Presumably, then, many of the problems and insights relating to the higher levels of speech processing will be directly applicable to the higher levels of visual language processing.

Physiologically, however, there also appear to be some deep and interesting differences between human aural and visual language processing. For one thing, the auditory perceptual system appears to handle incoming information in a serial fashion whereas the visual perceptual system handles its information input in parallel. Thus, the visual system apparently takes in a picture, or a string of words, at a glance, so to speak, whereas the auditory system must take in the beginning, middle, and end of a sentence in that order, that is, stretched across a definite period of time.

Whether such differences, at deeper levels, are more apparent than real is yet to be determined. Recent experiments, mentioned in Part II, suggest that a human stores up chunks of auditory information, and then in some parallel or hierarchical fashion makes a decision on its meaning. Similarly, in visual processes, one engages in serial-type action. In viewing paintings, for instance, one seems to take in a whole painting at one glance. However, experience reveals that one's eyes scan a painting in definite patterns (according to the school of com-

Fig. 1. One of man's minor victories over the machine: machines can't read graffiti although humans can. Machine printing, hand printing, and handwriting present different levels of difficulty.

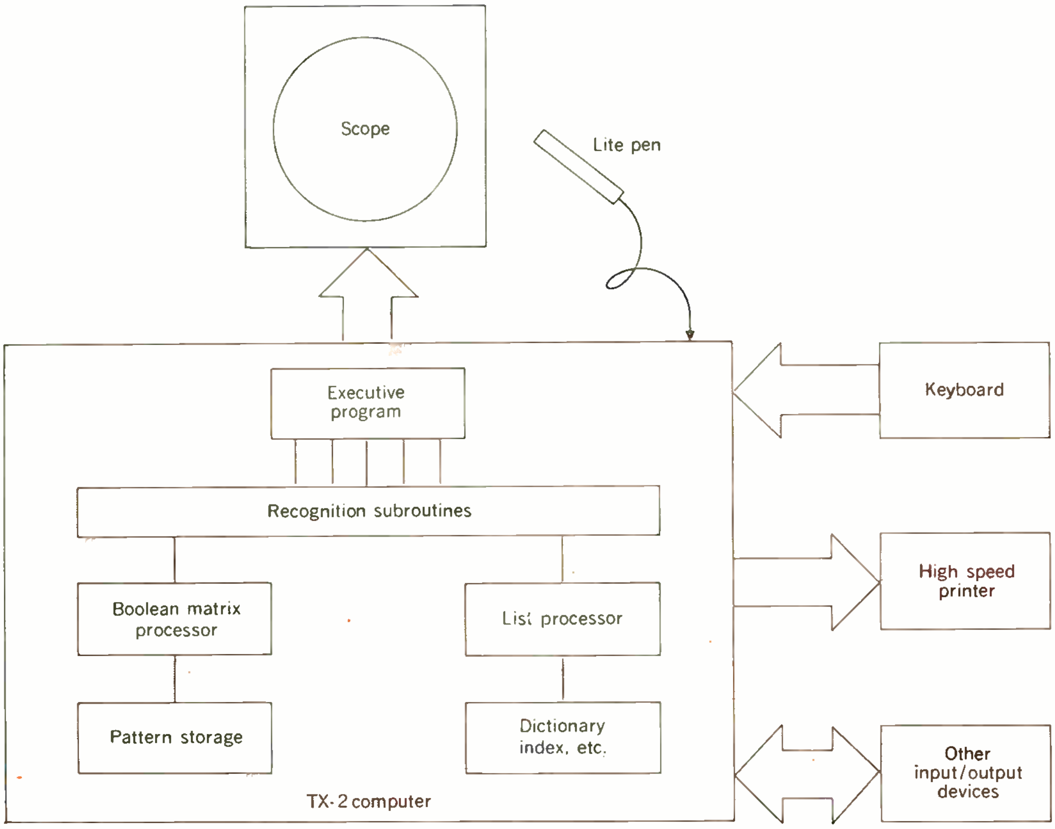


Fig. 2. Experimental system configuration devised by L. D. Earnest uses Boolean matrix processing techniques. Sample computer printouts of the recognition procedure appear at right.

position), and that as a result of this serial scanning, one sees more and feels more of the meaning of the painting. It would be very tempting to pursue speculations and relevant research along these lines, were there but space to do it. In any event, such basic differences and similarities as prevail among the human perceptual processes are nowadays of profound scientific and engineering interest because they form part of the foundation for all the specialized areas in the general field of pattern recognition.

Decision strategies

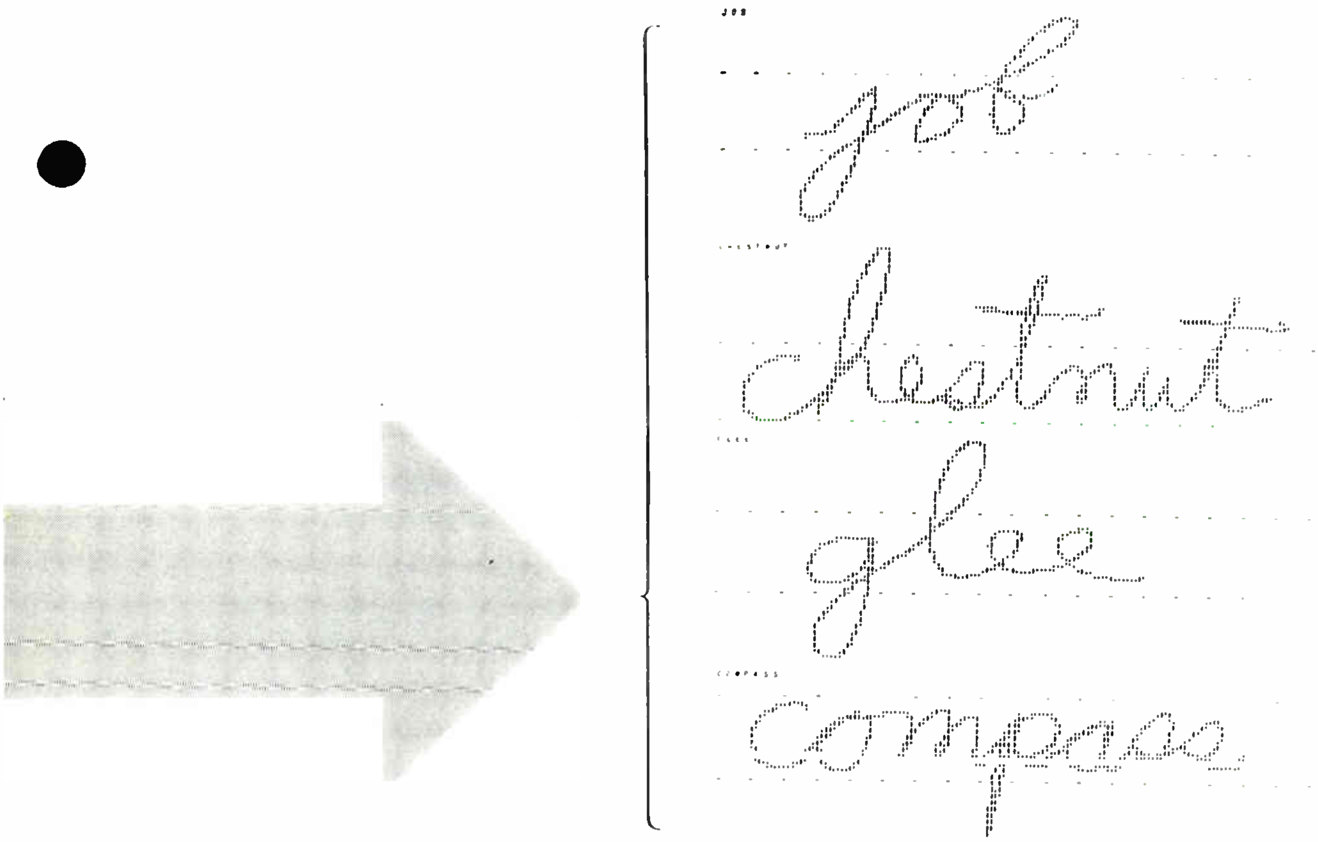
Three major efforts at devising methods of automatic handwriting recognition are outlined in the following pages. Each represents a different approach, and each has achieved considerable success. Although the description of these efforts is couched here in the most qualitative terms, the reader should realize, of course, that all have been pursued in a much more formal fashion.

As with the descriptions of speech research, decision strategies are little discussed. Our main interest is merely to sketch out somewhat comparatively the feature-selection and feature-abstraction methods. In any case, the question of appropriate strategy is an entire field unto itself, and a very large and extremely formal field at that. The interested reader should see, for instance, the work of theorists such as G. S. Sebestyen[12] or Professor King-sun Fu and his associates at Purdue University. Fu and Chen,[9] in addition to propounding their sequential-decision approach to pattern recognition and learning (especially in relation to problems of hand-printed and handwritten characters), also provide some perspective on the work of people like Abramson, Braverman, Daly, Glaser, Cooper, and Brick.

Recognition with few constraints

The first handwriting recognition program we shall consider was not the first to be undertaken. In fact, it was begun relatively recently by L. D. Earnest[13] of The Mitre Corporation and is still being actively pursued. It seems a good introduction to the feature-extraction problem, however, for it deals with whole words, avoids the segmentation problem, and is concerned with relatively few constraints.

Earnest identifies three categories of constraints on the handwriting process: graphic, dynamic, and linguistic. The graphic constraints he calls the learned conventional rules—i.e., writing horizontally from left to right, using small letters, high or upward-extending letters, and downward-extending letters, formed in conventionally specified ways, and so forth. Dynamic constraints are those imposed by the capabilities of the human hand, the musculatory control system, which may be compared to the articulatory system in speech. The linguistic constraints include some of those previously discussed, that is, words and the rules for making well-formed sentences in a language. But with handwriting, there are no phonological rules; rather, there are conventions of proper spelling. However, there are dynamic (muscular) phenomena of handwriting (how strokes are formed) that have a correspondence with the phonological rules of speech.



In Earnest's system, no explicit use has as yet been made of dynamic constraints; the only linguistic constraint exercised is that the words be from a known, finite vocabulary. Graphically, the writer is constrained to write more or less horizontally, and he uses a "lite" pen to write single words on the face of a cathode-ray tube (thus deferring certain problems related to actual writing on paper). The present system, which is set up on the M.I.T. Lincoln Laboratory TX-2 computer, recognizes about 10 000 common English words. The system deals only with whole words, and so circumvents, in part, the segmentation problem.

Earnest describes the three principal kinds of operations carried out by his recognition system as follows:[13]

1. Input patterns are encoded as two-dimensional arrays of binary elements; i.e., Boolean matrices with *0* and *1* standing for *white* and *black*. Transformations and tests on these patterns are performed by a Boolean matrix processor.
2. Certain arithmetic operations and tests are performed, mostly in connection with modeling of the physical handwriting process.
3. Language constraints are imposed through symbol manipulation, which uses list processing techniques.

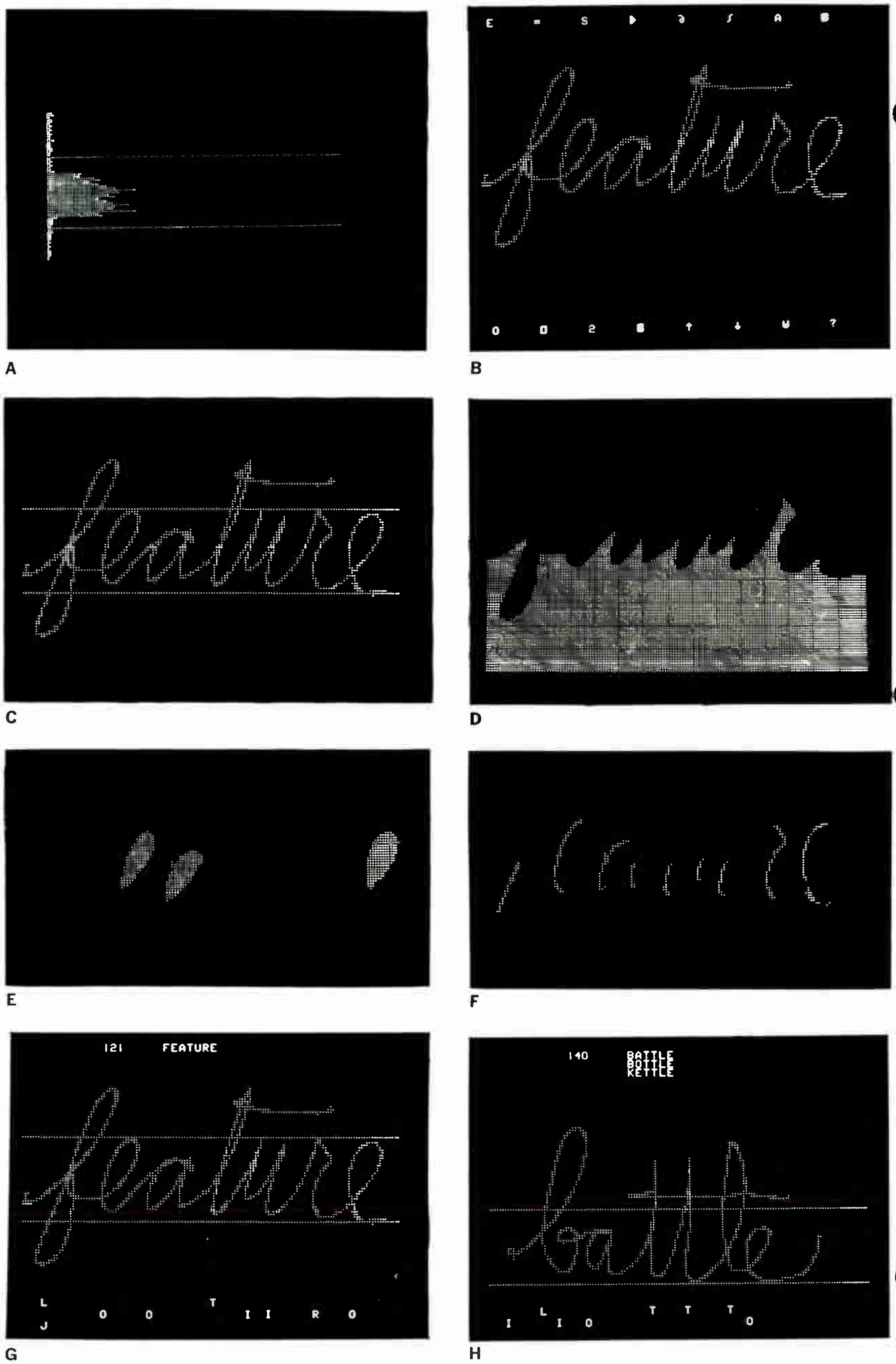
Figure 2 shows the system configuration. In this system, the recognition subroutines employ the matrix processor to extract features from the pattern and compare these features with those stored for the words in the dictionary.

With this system, the experimenter writes a word on the scope face, while the computer follows and records successive positions of the lite pen. The computer estimates the "envelope" of the small letters (those without upward and downward extensions), extracts key features, and forms a category code of these features. Dictionary words of about the right length are then chosen for comparison, and the horizontal coordinates of key features are tested for reasonableness against each word on the list. This process, which in the TX-2 requires about 15 seconds, may result in an output of no words, of one word, or of several.

The detailed procedure is traced out in Fig. 3, which shows each step from the estimation of the envelope through to the computer-printed output. Only six different features are used in this system to achieve recognition of 10 000 words.

In this recognition system,[14] the writer is not constrained to write the small and large letters within certain fixed lines. Therefore, one of the first tasks of the system is to estimate the bounds or envelopes of the letters, as in Fig. 3(A), before proceeding to extract the word features, as in Figs. 3(B)–3(G), which in this case results in perfect recognition. Figure 3(H) shows a different example in which the display only indicates that the word is one of three possible ones, either *battle*, *bottle*, or *kettle*. However, this example has the advantage of revealing something about the method of encoding.

Figure 4 illustrates the method more explicitly. Certain features of words are used to form a category code, and all the words of the dictionary are organized according to the code. This three-numbered code is seen in the upper left of Fig. 3(G). The first number tells whether or not there are crossbars, the second tells the number of high strokes, and the third tells the number of low strokes. Thus, the word *feature* has one crossbar, two high strokes,


E S A
feature
0 0 2
feature
121 FEATURE
feature
L J O O T I I R O
140 BATTLE BOTTLE KETTLE
battle
I L I O T T T O

A

B

C

D

E

F

G

H

and one low stroke; the word *battle* has one crossbar, four high strokes, and no low strokes, as do the words *bottle* and *kettle*. In Fig. 4, it can be seen that the word *word* is encoded 010—i.e., no crossbars, one high stroke, and no low strokes. Words with the same code, and more or less the same length (in this example, *ale*, *ash*, *work*, and so forth), are stored together as shown. In addition, there is a table listing the different ways each letter may be decomposed. For instance, the letter "a" could be represented by the features "O," "OI," or "II." Each dictionary word is transliterated into its possible alternatives of features for comparison with the observed features, and those that match are presented on the output. Some computer printouts appear in Fig. 2.

Although Earnest does not give any comprehensive figures on the success of his recognition system (an earlier report cites about a 60 per cent success), the method is certainly interesting in view of the very few constraints he employs to obtain even such a margin of success. From this point on, he intends to find matrix operations to extract more feature information, and thereafter, evidently, to weigh the improvements that might be possible using dynamic and syntactic constraints.

Finally, he says, "whatever is done, the computer, like any good secretary, will occasionally have to respond with something like, 'What in the world is this?' "

In contrast to Earnest's work, we should next look at two complementary approaches that take into account many more of the observable features of handwriting, and that rely basically on binary decisions about the existence or nonexistence of such features.

Fig. 3. A—Method of estimating envelope of small letters.The computer scans the density histogram of the word (which characteristically bulges as shown here) from bottom to top, finds the outer envelope bounds, and then sets the small-letter envelope bounds above and below certain thresholds. B—The word "feature" to be recognized displayed on the oscilloscope just as it is stored in the computer memory, on a matrix of 100 × 176 elements. Control symbols on the periphery may be selected by the lite pen. C—The computer first estimates the small-letter envelope, adds linear extensions to the two ends of the word, dividing the space around the word into upper and lower regions. D—With a single matrix operation, the lower region is explicitly derived. The significant downward strokes, such as for f, g, j, p, q, and so on, are determined. The upward extending letters and their crossbars are similarly found. E—The upward and downward strokes are truncated, and closed curves within the envelope are detected. F—Downstrokes are extracted and are used to segment the word and to key the extraction of features for later processing. Downstrokes are to the left of closed curves or in the lower region (compare with D). G—The recognition output is in this case unique. Features extracted are printed below: L and J designate high and low strokes; O designates closed curves such as for "e" and "a"; the spikes of the letter "u" receive the designation II, and so forth. H—Nonunique recognition. Further feature extraction would be required to separate these three responses uniquely. The letter "l" was detected as T. The present program takes such interpretations into account.

Two analytical methods

Two complementary analytical approaches to the machine reading of cursive script (that is, the automatic transcription of handwriting to typewritten text) were undertaken by L. S. Frishkopf and L. D. Harmon of Bell Telephone Laboratories. Their first efforts were reported in 1960.[15] During the succeeding years, Leon Harmon has continued to dig deeper into the problems of cursive script recognition. This research constitutes probably the first and the most thoroughgoing analytical approach to handwriting recognition that has been undertaken.

In their original approach, Frishkopf and Harmon took separate approaches; Harmon tackled the problem of letter-by-letter identification and Frishkopf the problem of entire words without considering the letter structure. Their methods complement each other in that sequences of classified letters could be tested as words, and words could be run through letter-analytical tests.

Physically, the methods rely on the study of the motion of the writing instrument as a function of time. Continuous horizontal and vertical displacement information is derived through the use of a captive stylus of a Telewriter. Thus, in this method, as in Earnest's, there is no attempt to treat the recognition of script letters or of words already written.

Individual-letter method. In the individual-letter method of analysis, which was simulated on a computer, letter identification was based on the extraction of certain local features of individual letters incorporated in whole words. Thus, the difficult problem of segmentation of individual discrete letters in continuous strings was met head on. (The problem of trying to find invariances in letter connectives is analogous to the problem of segmenting phonemes in running speech.)

Several constraints were imposed on the writer of specimens: He was asked to write carefully (no scrawl or scribble). He was supposed to write on a fixed baseline on the Telewriter; all small letters (those without

Fig. 4. Schematic of how the set of extracted features (in this case RIOROL for word) are compared with the dictionary entries.

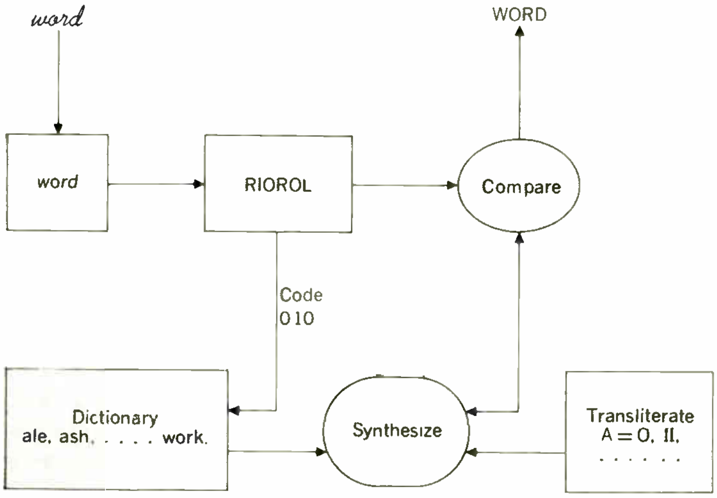
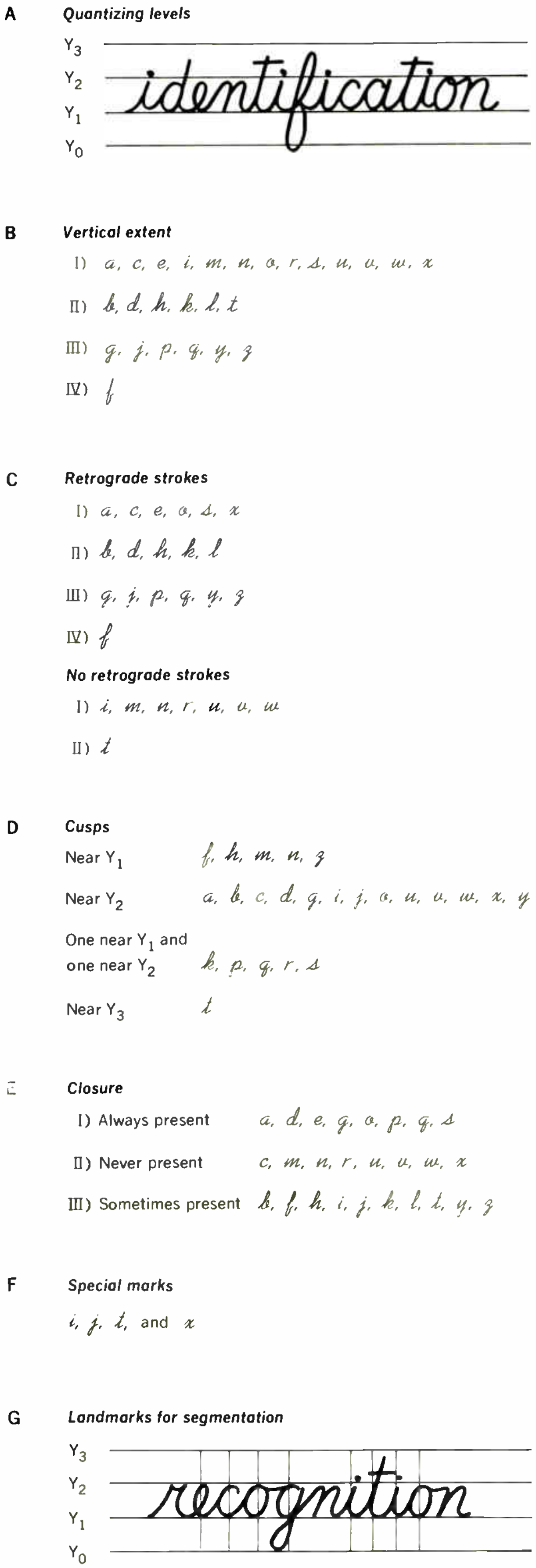


Fig. 5. A—Method of quantizing handwritten words. B–F—Features used in recognition system. G—One approach to the segmentation problem.



upward or downward extensions) were to be written in a fixed zone above the baseline. Furthermore, no capital letters were allowed. Figure 5(A) shows how words were quantized in levels.

Within this system of constraints, Harmon examined the features of the alphabetic characters and sorted them into groups according to the features he chose for his decision system. These sets of features were vertical extrema, retrograde strokes, cusps, closures, and special marks. The meaning of the vertical extent category and its subgrouping is clear in Fig. 5(B). Retrograde strokes are simply those moving from right to left, in opposition to the stream of ongoing writing from left to right. In Fig. 5(C), these are organized according to the vertical-extent categories. Cusps are pointed ends or sharp peaks, that is, abrupt changes in slope. The vertical locations of such cusps, grouped as in Fig. 5(D), provide added usable information. Closure [Fig. 5(E)] is defined as the return of a continuous line to within a very short distance of a prior location (that is, a loop or near-loop). Special marks, usable for identification, appear in the four letters of Fig. 5(F). (Although, in this first recognition program, Harmon regarded the dots for i and j as providing relatively weak identification criteria, he discovered later that such marks are quite good determinants of idiosyncratic writing behavior.)

Having formulated the features for his decision algorithms, Harmon next considered the problem of how to segment whole words into their letter components. This is the problem of how to assign automatically the recognizable features to their proper owners.

In his efforts to solve this problem, Harmon tried a number of segmentation methods, none of them truly satisfactory; however, in combination, especially with probabilistic methods he was to try later, he was able to increase greatly their effectiveness. The method he used in his original computer simulation study was first to locate certain rough "landmarks" by establishing those places in the script that are vertically extended, retrograde stroked, and specially marked, as shown in Fig. 5(G). One can see, with a few experiments on words picked at random, how these three features do roughly locate quite a few letter positions. Next, Harmon obtained average letter-width measures through counts of center-axis crossings (about three axis crossings per letter) and overall word length. By centering these average-width markers on the rough landmarks, and by dividing up the entire letter space equally with them, he segmented the letter space. The accuracy of this method fell off whenever nonextended, nonretrograde, nonspecial letters (m,n,r,u,v,w) occurred in groups of three or more. However, trigram combinations of these letters do not occur frequently. In later research, digram and trigram probabilities were utilized to improve recognition results.

Using this segmentation method, and setting up a decision tree (a similar, but improved decision tree appears further on in this article) based on the five classes of features set out above, fair recognition results were achieved. Correct letter identification for the entire group of specimen sentences, written by five persons, was about 60 per cent; segmentation was 87 per cent successful. The system diagram appears in Fig. 6(A).

Whole-word identification. In the complementary system used for the recognition of whole words, developed by L. S. Frishkopf,[15] fewer restrictions were put on the

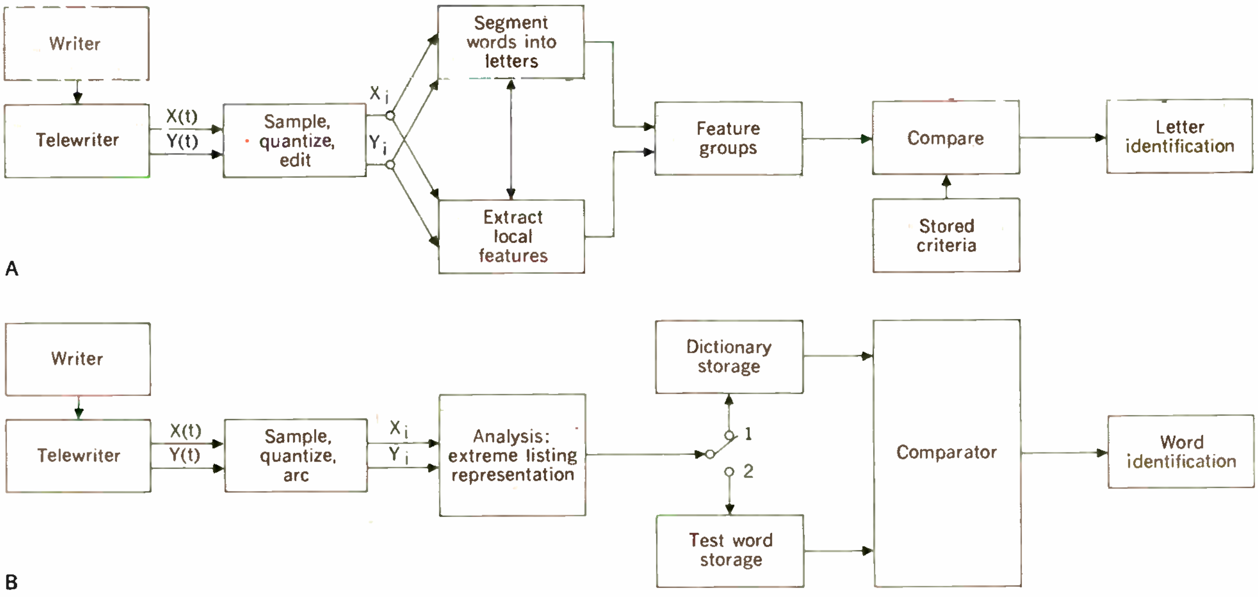
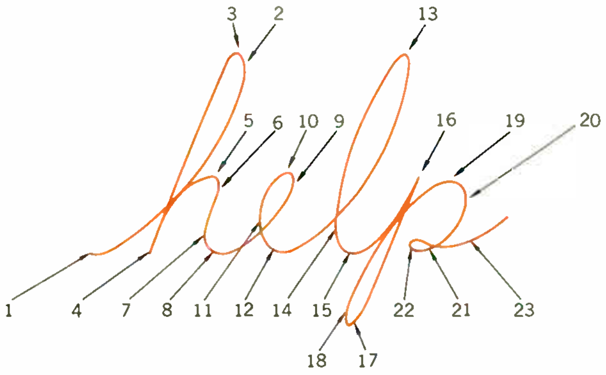


Fig. 6. A—System devised for individual-letter identification. B—System devised for whole-word identification.

Fig. 7. A string of extreme points occurring in a handwritten word are used in whole-word identification system.



writer. He was asked to keep to a baseline in his writing and to write legibly in Roman cursive script, but he was not obliged to keep within a fixed envelope. In this system, as in that of L. D. Earnest, whole words are compiled according to certain features in a dictionary with which input words are compared. However, the features Frishkopf selected for comparison are quite different.

The schematic of the system is shown in Fig. 6(B). Essentially, what this system does is transform samples of the real-time handwriting input into samples representing points equally spaced along the handwriting curve. The features that Frishkopf chooses to compile for his dictionary representations are the horizontal (X) and vertical (Y) extremes along the handwriting curve; that is, points of local maxima or minima. The order of the extremes is preserved in the listing. For instance, the word *help* in Fig. 7 is represented by a string of 23 consecutive extremes. Each extreme entry also contains information on: whether or not it is an X or Y extreme; whether it is right or left, upper or lower; what kind of slope exists between successive extremes; whether connective curves are convex or concave; and group classifications according to the zones of upper and lower extensions in which the extremes are found.

In the identification process, input words are separated from surrounding words, analyzed into an extreme listing, and compared with the entries in the dictionary. The strategy used in the comparator of this whole-word identification system (not exactly simple), will not be discussed here; its description is readily available.[15]

Tests of this whole-word identification method led to various results. When the same writer was the source of the dictionary and the test words, identification was about 30 per cent correct. In one case, when the writer was further constrained to write test words within specified zones, nearly 60 per cent of the words were identified.

In these pioneering studies of handwriting recognition, it should be noted, very little was attempted with linguistic context (aside from a fixed, limited dictionary, and so on). In this respect, these researches are directly comparable to the attempts at automatic speech recognition on the acoustic level alone. However, inasmuch as handwriting studies were stimulated by speech studies, and followed speech studies by many years, it was perhaps clear from the beginning that it would be necessary to make use of linguistic structure in any satisfactory recognition scheme. Indeed, it is interesting to see that such was the direction these researches took.

Subsequent analytical research. Since the 1960 report on his research, Harmon has spent several years improving handwriting recognition techniques based on individual letter identifications. During this period, he has accomplished three main things: (1) He improved the feature-extraction algorithms and the method of segmentation, and corrected some design errors. These changes raised the overall identification accuracy from 58.9 per cent to 87.4 per cent and the segmentation accuracy from 87 per cent to 97.8 per cent.[16] (2) He introduced error detection and error correction through the use of linguistic context; i.e., with digram- and trigram-frequency statistics. These changes raised overall identification accuracy to 93.7 per cent.[17] (3) Finally, he worked on methods of recognizing individual writers through their idiosyncratic writing features. Out of 100 writers, his system was able to identify 96 correctly, and of the other four, two were in second place. Because it is unrealistic to compare a machine's performance against a 100 per cent accuracy ideal, it is important, he points out, to find out exactly how well humans do in similar recognition tasks. A group of readers were asked to compare unknown writing samples with the 100 reference samples, and to identify the writers of the unknown samples. Overall identification accuracy for the group was 96.25 per cent against the automaton's score of 96 per cent![18] Following are some of the highlights of these improvements.

Segmentation errors were reduced by making letter-width estimates over an entire sentence rather than over a single word; the rules for segmentation were revised; and a new set of combinatorial rules for identifying letters in unsegmented spans was developed.

Certain inadequate feature-extraction algorithms were rectified by more optimum choice of threshold parameters, and the decision tree (a binary sequential truth table) was revised. The new decision tree, in Fig. 8, shows clearly how the individual letters are sorted. (Harmon has also implemented a much simpler decision-tree system in a small working model of a handwriting reader; it reads the whole words *zero* to *nine* with a recognition accuracy of 97 per cent.[19])

The essence of the contextual error-detection and -correction technique is described as follows: "Error *detection* is based on digram-frequency statistics; *location* of the single letter in error (given an 'improbable' letter pair) is accomplished by using letter-error statistics of the recognition system; *correction* of the error utilizes trigram statistics and letter-confusion error data obtained from prior performance of the system.

"In a study of digram statistics, we observed a simple but little-noted property of the English language; namely, that 42 per cent of all possible digram combinations have zero likelihood of occurrence, and 50 per cent of all possible digrams have $\leq$ 0.003 per cent likelihood of occurrence. If a single-letter error is produced in a word by random substitution, the resultant digrams thus generated have low probabilities of occurrence compared to naturally occurring letter pairs. It is on this statistical property that our principle of error detection is based."[17]

Unfortunately, it is not possible to go into details on the error-detection techniques (Harmon is preparing a paper on this work[10]); however, when errors are detected and located, they are corrected by the following procedure: "The errant letter is deleted from the word. A trigram sequence is then formed with the two adjacent letters (i.e., $L_1 - L_3$). Trigram frequencies are consulted to find which letter produces the sequence with the highest natural occurrence. This letter is said to be the correct one."[17] It is interesting to note that similar frequency-of-occurrence statistics are being used by J. W. Forgie at Lincoln Laboratory to facilitate correct recognition of speech sounds.

The program for identification of individual *handwriters* was based on the relative occurrence of distinctive features made by a writer. Relative occurrence is the ratio of the number of times a writer makes a feature to the number of times the feature could occur. These ratios Harmon discovered to be quite reliable identification determinants. For instance, a writer who tends to dot eight out of ten *i*'s in a sufficiently long writing sample holds quite consistently to that tendency, and he can be distinguished from a writer who tends to dot six out of ten or ten out of ten. In all, Harmon assembled a list of 74 features that could be used for identifications.

Analysis-by-synthesis of handwriting

We have already encountered Roman Jakobson's concept of the distinctive features in Part I and II. This concept has also been carried over, somewhat transformed, into the analysis-by-synthesis of handwriting. This work, undertaken by Murray Eden at M.I.T., was stimulated by the research going on there along such lines. In fact, Eden remarks that he became interested in pursuing the analysis of handwriting by such programmatic methods after hearing a talk by Morris Halle on the speech problem.[11] Speculatively, it appeared that insights into the nature of handwriting could also lead to insights into other forms of communicative behavior—in particular, "the acoustical analogon of script, speech."

With an eye cocked on the distinctive features of speech, Eden broke down writing into similar fundamental elements ("primitives"), a simple set of stroke gestures out of which all letters and words could be composed. However, although Eden points out that these basic line segments are similar to the distinctive features, he is careful to stress that there is no exact correlation, just as there is no exact analog between letters and phonemes. (For the arguments on this question, and on the philosophical question of breaking down quasi-continuous writing into discrete elements, one should see Eden's 1960 paper.[20])

In any event, because speech and writing are regarded as obviously closely related activities, the analysis of writing in Eden's work follows lines of development paralleling those of Stevens and Halle in speech. Once again we encounter *rules* of formation (in this case, rules of handwriting generation): these are based on the fact that the human hand and its associated control system seem to find certain types of cycloidlike strokes easiest to produce.

Eden's approach, and its evolution, is now well documented[21–26] in the literature, and is readily available, so we shall merely sketch out some of its major features here. In its earliest stages, the research dealt mainly with the process of generating handwriting. It was an attempt to describe, very formally, the structure of handwritten words, and an attempt to determine the necessary rules for assembling (i.e., generating) such word structures. More recently, the approach has evolved to the stage of being used for computer recognition programs.

In the beginning, Eden set up a formal description of cursive writing based on a set of four primitive symbols, namely, a set of four point pairs, and rules were defined for how lines were to connect these points.[20] This subset of four strokes, called segments [Fig. 9(A)], were transformed by rotation and reflection to generate a set of 11 distinctly different symbols. These symbols were moreover translated up and down into one of three partially overlapping horizontal fields to produce a set of 33 strokes. Of these strokes, 18 were sufficient to describe

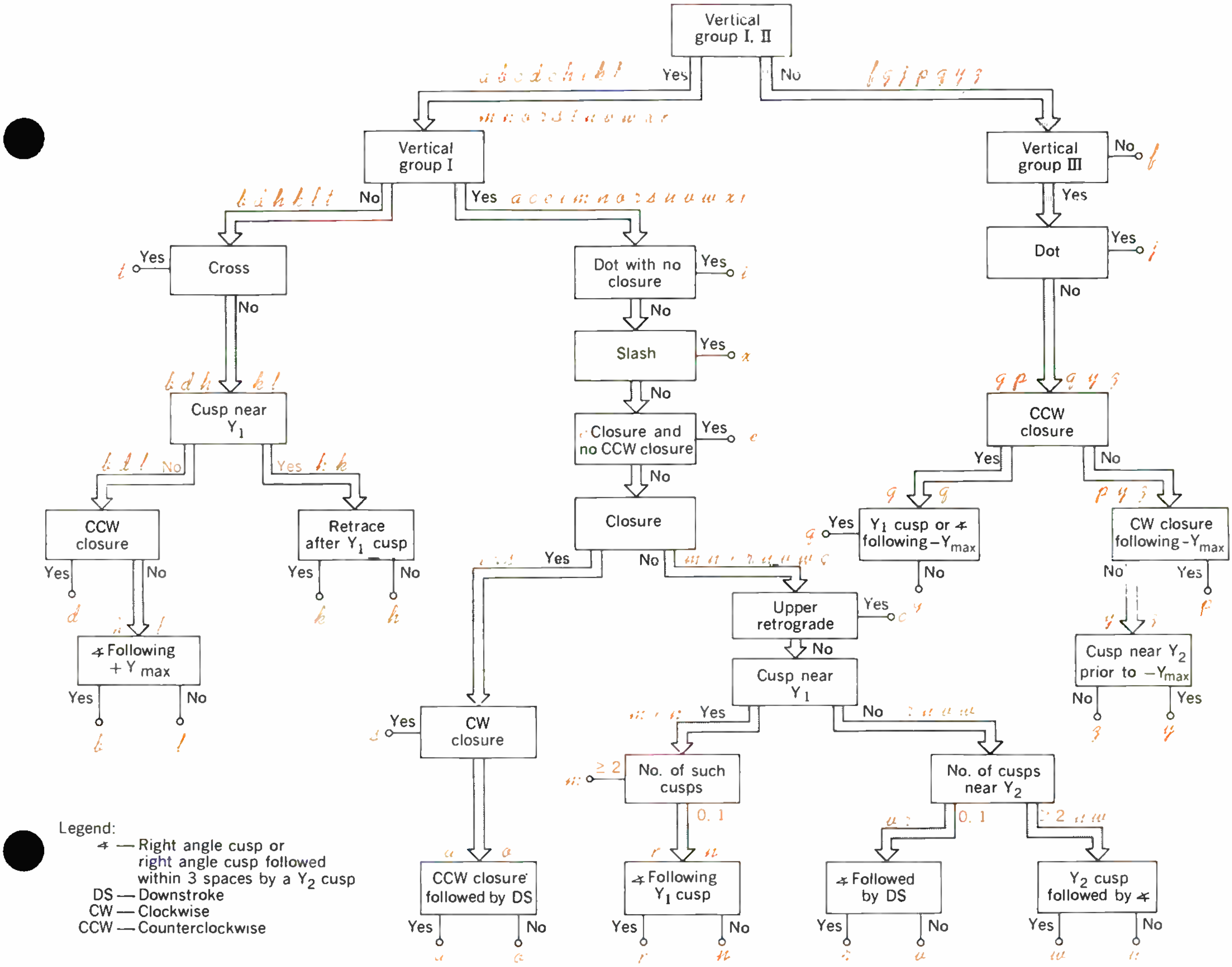


Fig. 8. New and more effective decision tree used in individual letter identification system.

all the English upper- and lower-case letters [Fig. 9(B)]; that is, it was possible to define each letter of the language as a unique, finite sequence of strokes. A few of the letters of this stroke representation of the alphabet are shown in Fig. 9(C).

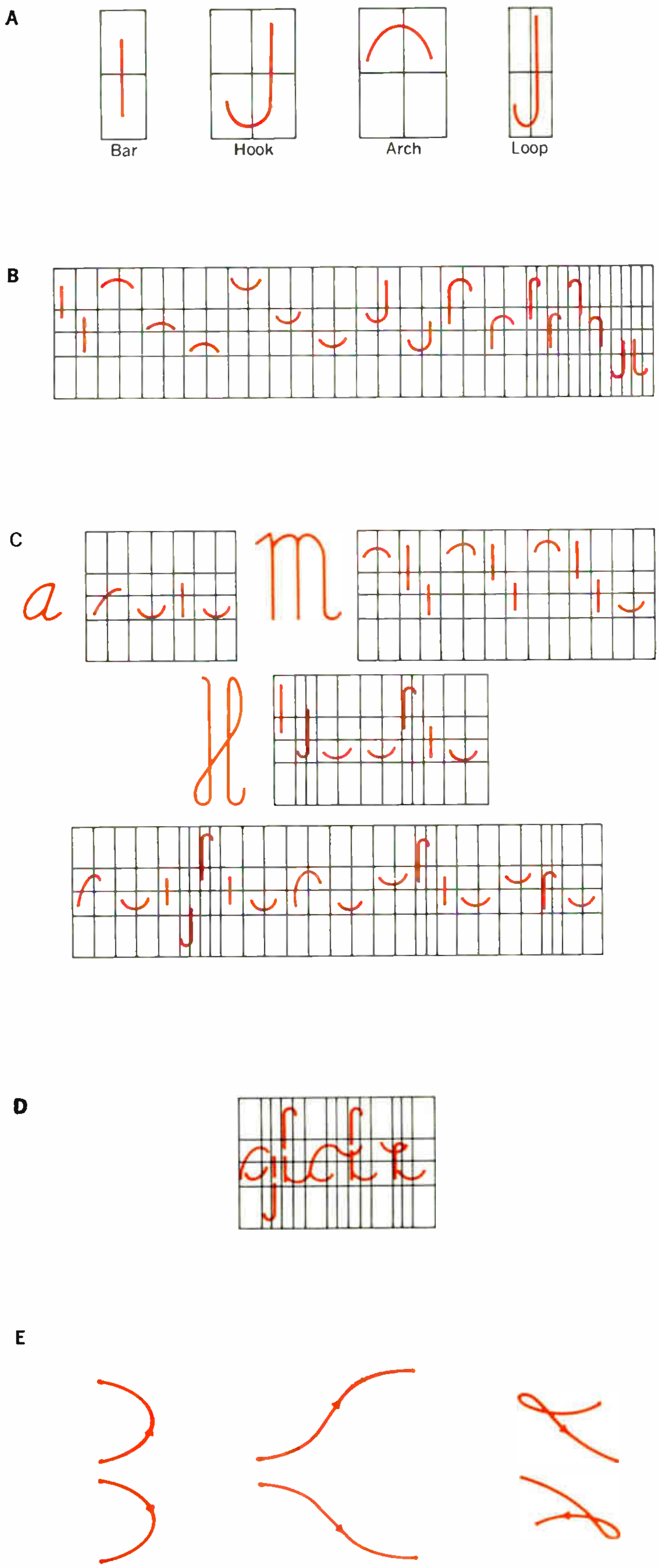
In this formal scheme, Eden points out that the line segments—the four primitives—are analogous to the distinctive features, the strokes are analogous to phonemes, and the letters composed of these strokes are analogous to morphemes.

An interesting historical comparison of these basic strokes, or invariants, has been pointed out by R. A. Fairthorne.[27] He notes their resemblance to the Cancerellesca hand of the 16th century, when public scribes, under the necessity of writing a great deal quickly, evolved letter shapes that remained recognizable even though they were deformed due to the speed of writing. These letter shapes corresponded to the invariants chosen by Eden and Halle.

Some of the rules for collating the basic strokes deal with strokes within single letters; others deal with strokes between letters. An example of how the word "globe" is represented in this formal system is shown in Fig. 9(D). The script defined by the formal system was, as Halle and Eden pointed out, a sort of idealized norm. No one, of course, actually writes this way. The first attempts at generating handwriting with these formal elements used a kind of Markov process, with a resultant handwriting that was rather awkward looking and childlike.

To generate writing that would match that of an individual writer, it appeared necessary to study the dynamics of actual writing, that is, to consider how the writing muscles governed the movements of the pen point. (This interest in the muscular dynamics has its parallel in the studies of the articulatory gestures involved in the production of speech.) Accordingly, Eden developed a somewhat qualitative sinusoidal description of actual writing dynamics, in which the strokes were described by segments of periodic functions. In this model of events, handwriting—i.e., the movement of the pen point—was seen as being generated by three groups of muscles acting more or less independently. One pair of muscles, simultaneously flexing or extending the thumb and index

Fig. 9. Progressive stages of analysis-by-synthesis method. A—Set of basic line segments. B—Set of 18 strokes. C—Stroke representations of letters. D—Stroke sequence for the word "globe" and its representation after being collated by rules. E—Function segments used for strokes and ligatures.



finger, governed vertical movements or movements along the direction of the predominant slope of the writing. A second pair governed the radial deviation of the hand at the wrist, and produced the fluctuating horizontal movement of the writing. The third pair, controlling adduction and abduction of the upper arm, produced the steady movement of the pen across the page. Each pair of muscles was viewed as generating a sinusoid in time. Thus, Eden viewed this formulation as a harmonic oscillator theory for the generation of handwriting.

In Eden's model, the velocity-function component of the writing was treated as a series of half-wave sinusoidal segments with an appropriate phase shift between the two components. By means of this phase shift, it became possible to generate suitable ligatures containing points of inflection. Examples of the kind of *function segments* generated by this means, and which could be used for both strokes and ligatures, appear in Fig. 9(E).

Further detailed studies along the lines of this formulation were carried out by Eden and Mermelstein. By using a special transducer, they studied the detailed properties of pen displacement and velocity as a function of time, from which they evolved two models[24] to describe such movements. The first of these models, a modification of Eden's 1962 model, had been extended to allow for differences between the generating parameter values corresponding to stroke segments having positive and negative vertical accelerations. In all, this formulation could generate any stroke with the specification of just seven parameters. Their second model, proposed at the same time, was suggested by the work of a number of Dutch researchers, J. J. Denier van der Gon and his associates at the University of Amsterdam, who had developed a handwriting simulator that could write words at a high speed.[28] This handwriting simulator, its authors felt, worked on the same principles as real cursive handwriting does. Although this handwriting simulator research has the same kind of relevance to an automatic handwriting recognition program as does Eden's and Mermelstein's, Dr. van der Gon reports that there has been no opportunity to carry their research in such a direction.[29]

In Eden's and Mermelstein's two mathematical models for the dynamics of handwriting generation, the generating parameters were obtained by analyzing actual handwriting through an operational segmentation process.

Eden and Mermelstein have most recently carried out experiments on computer recognition of connected handwritten words based on the researches described above, namely, by treating the handwritten pattern as a two-dimensional vector displacement function of time. In the recognition procedure, handwritten words were analyzed into strokes or function segments, and these strokes were "recognized" or classed according to the statistical likelihood of their belonging to preselected classes. Then, certain constraints were invoked at the stroke and letter level to limit the output sequence generated, that is, to maximize the likelihood of correct recognition.

The writing is done on a special transducer; thus, the problem of dealing with words already written is still avoided. This system of recognition, based on the search for pattern invariants by a consideration of the intrinsic movements that produce them, bears many of the earmarks of the formulations we have already encountered in speech research. Even though, it is said, there are wide differences between different writers, and even for the same writer under different conditions (substitute *speaker* for *writer*), the fact that such great idiosyncratic differences can be interpreted (understood) by many readers implies that there exist well-defined rules for handwriting generation.

Say Mermelstein and Eden: "We assume that the handwriting generating system of the hand and arm acts as a transducer, receiving by means of the nerve fibers a description of the graphical pattern to be executed. The system is noisy in that it introduces variations in the outputs corresponding to input signals thought to carry the same information. Any successful recognition system, human or artificial, must be invariant to just such variations."[25]

Furthermore, "Certain features of the pen movement considered as a function of time, are largely invariant for different writing samples of the same word written by the same subject, despite the possible existence of larger variations in the specific spatiotemporal muscle-innervation pattern. Some of the features considered in this context are the predominant direction of downstrokes, relative heights of letters, and the characteristic pattern formation of particular letters. Through a quantitative investigation of the invariance of these features we can characterize the generating system and separate the effects that are inherent in the system from those resulting from specific influences such as word and sentence context. To the extent that certain writing features maintain their invariance over different subjects and are found to carry information pertaining to the letter under execution, they may be used to recover the message from the handwritten pattern."[25]

Specifically, Mermelstein and Eden studied the variations in writing speed for several different persons, and they found that there were speed minima that were predictably located along the time axis. These speed minima, or the points of zero y velocity, defined the segmentation of the strokes or function segments. By this definition, writing samples were viewed as sets of upstrokes and downstrokes ordered in time.

Moreover, in the experiment cited, the stroke space was partitioned in such a way as to give all letters (nine letters were used for the input words, which consisted of such words as *fail*, *fall*, *feel*, *fell*, *fill*, *foul*, etc.) unique representations based on downstrokes only. Thus, for the most part, the information in the upstrokes was redundant once the downstrokes were specified. However, there were some cases in which the downstrokes were not known with certainty, so that the use of upstroke information improved recognition results.

Twelve parameters were used for stroke description; these included such measures as the x and y components of stroke displacement, initial and final x-velocity values, phase-shift parameters of x velocity relative to y velocity for initial and final segments, and so on.

What kind of results were achieved? The authors say: "Experiments carried out by computer simulation of the recognition system reveal that the system is capable of recognizing well-formed, legible handwritten words with a reliability that depends on the correspondence between the script of the test samples and that of the ensemble on which the machine's representation of handwriting is based. For an ensemble of 100 samples written by four subjects with a vocabulary chosen so that adjacent letters provided little contextual information, 91 per cent of the samples were correctly recognized if the machine had been previously exposed to the same 100 samples. Lower recognition rates are obtained in situations in which differences exist between the teaching and test ensembles."[25]

Whether or not this kind of analysis-by-synthesis approach to handwriting recognition can be extended to a general vocabulary is yet to be demonstrated. At the present time, Eden is working on the problem of how the approach might be applied to writing already written. One of the big problems, evidently, is how to gather time information from script that is already written.

Curve tracing might be one method to derive time information, but this method too presents some difficult and literally knotty problems. When curve tracing is used with printed characters, the researcher deals with topologically separate entities of relatively small complexity (see, for instance, the work of Freeman[6] and others); but with curve tracing in cursive script, the researcher is confronted with clusters of multiple crossings in small areas, so that the tracing mechanism is baffled about the direction in which to proceed.[30] Some way, evidently, would be needed to reduce this complexity, perhaps in combination with other types of strategy, probabilistic weightings, and so forth.

Another large problem, Eden indicates, is the determination of the patterns of interletter constraints. Inasmuch as context determines specific forms in script (for example, in the words *cat* and *bat*, the letter *a* has a differing initial formation owing to the influences of the *c* and the *b*), it is certainly conceivable that information on such constraints could be used in probabilistic matching schemes analogous to those mentioned earlier. In their latest research, Mermelstein and Eden worked with constraints on several levels—with constraints between strokes, with constraints between letters (but based on probabilities that they were letters belonging to words of the given set), and with constraints between words (that is, they had to be words from the given, restricted dictionary of 59 words; these included members of the input set, mentioned earlier, and other words with which they might be confused). However, Mermelstein indicates that he did not work with the kind of interletter constraints observed above.[30]

In any event, it is interesting to see how this program has evolved through a progression of ideas, and it should be of interest to see how it evolves in the future.

In general, the question may be asked: What will be the eventual achievements of programs such as these? Leon Harmon remarks that most automatic handwriting schemes prosecuted to date have not advanced much beyond the stage of being laboratory toys.[10] Of the future, L. D. Earnest concludes that although reliable mechanical recognition of cursive writing has not yet been achieved, such an achievement is not out of the question.[11]

Epilogue—What is recognition?

It may or may not seem odd that in a survey of machine recognition of human language we have really avoided an explicit discussion or definition of what constitutes recognition in the human. It is, in fact, a question that is rather considerably sidestepped in what we might call the "working" literature on pattern recognition. However, we should not terminate this kind of survey without at least realizing that the basic concepts governing some recognition work are open to question. In part, the very efforts to incorporate or simulate recognition in a machine pries open such essentially classical questions to renewed examination. Moreover, the answers one propounds to such questions—if one is working in the field of automatic pattern recognition—can have important, practical consequences.

A philosophic analysis of the concepts of recognition, especially in relation to current work on "artificial intelligence," has been undertaken by Kenneth M. Sayre in a book issued just this year.[31] It could, Sayre suggests, open up a new field of constructive inquiry. His major concern is "with human pattern-recognition, and with certain conceptual obscurities which are blocking current attempts to simulate it." These obscurities, he goes on to say, are in part methodological, but behind these lies a deeper obscurity, namely, "an unclarity about the nature of the human behavior we are trying to simulate. We simply do not understand what recognition is. And if we do not understand the behavior we are trying to simulate, we cannot reasonably hold high hopes of being successful in our attempts to simulate it."

"At the heart of these conceptual difficulties is a fundamental confusion between recognition and classification. That recognition has been conceived by researchers in the area to be a form of classification is evident to anyone familiar with the literature of mechanical pattern-recognition. Yet the distinction between recognition and classification is so unmistakable that once it has been made articulate it will appear strange that we should ever have fallen prey to the confusion."

Those words are in the introduction to Sayre's book and following them is a lengthy discussion of these conceptual problems besetting the researchers who would design recognition automata. These problems, Sayre points out, "in fact bear an instructive resemblance to problems with which, in a more general form, philosophers have been wrestling for centuries." Of special relevance is the fact that Sayre's discussion includes many of those programs described in this series.

Thus, instead of terminating this discussion with a neat conclusion, we must terminate with an open question. Nonetheless, it is right that this is so. For one function of a survey, it seems, is not to satisfy readers wholly, but to leave them in some measure dissatisfied so that they will not rest until they have gone further. Whether the means of achieving this are intentional or (what is more probable) unintentional, the result ought to be this: that the deeper, pregnant questions will be touched, and in one way or another, manifest their presence at the surface of the consciousness. In stirring up such questions and dissatisfactions in ourselves, we are at least in an empirical fashion disturbing and engaging our own functional systems of recognition, of classification, of concept making; and hopefully this will lead each of us to press his own investigations further.

The author thanks Dr. Lawrence S. Frishkopf, Leon Harmon, L. D. Earnest, Prof. Murray Eden, and Dr. Paul Mermelstein for their kind assistance and guidance.

REFERENCES

1. "Reading Machines for Data Processing—Their Prospective Employment Effects," Manpower Report no. 7, U.S. Dept. of Labor, June 1963.

2. Falk, H., "Optical Character-Recognition Systems," *Electro-Technology*, July 1964, pp. 42–52.

3. Stevens, M. E., "Automatic Character Recognition—A State-of-the-Art Report," NBS Tech. Note 112, PB 161613, U.S. Dept. of Commerce, Washington, D. C., May 1961.

4. David, Jr., E. E., and Selfridge, O. G., "Eyes and Ears for Computers," *Proc. IRE*, vol. 50, May 1962, pp. 1093–1101.

5. Gardner, L. C., "On the Recognition of Hand-Printed Characters," Tech. Note 400-6, New York Univ., Dept. of E. E., Sept. 1961.

6. Freeman, H., "On the Encoding of Arbitrary Geometric Configurations," *IRE Trans. on Electronic Computers*, vol. EC-10, no. 2, June 1961, pp. 260–268.

7. Kuhl, F., "Classification and Recognition of Hand-Printed Characters," *IEEE Internat'l. Conv. Record*, pt. 4, Mar. 1963, pp. 75–93.

8. Holt, A. W., "Character recognition using curve tracing," U.S. Patent No. 3 142 818, Aug. 1964.

9. Fu, K. S., and Chen, C. H., "A Sequential Decision Approach to Problems in Pattern Recognition and Learning," *Third Symposium on Adaptive Processes*, Chicago, NEC, published by IEEE, Sept. 1964.

10. Harmon, L., Private communication, Bell Telephone Laboratories, Murray Hill, N.J.

11. Eden, M., Private communication, Dept. of E. E., M.I.T.

12. Sebestyen, G. S., "Recognition of Membership in Classes," *IRE Trans. on Information Theory*, vol. IT-7, Jan. 1961, pp. 44–50.

13. Earnest, L. D., "Machine Recognition of Cursive Writing," in *Information Processing 1962, Proc. of IFIP Congress 62*, C. M. Popplewell, ed. Amsterdam: North-Holland Publishing Co., 1963, pp. 462–466.

14. Earnest, L. D., "Machine Recognition of Cursive Writing," unpublished paper, 1964.

15. Frishkopf, L. S., and Harmon, L. D., "Machine Reading of Cursive Script," in *Information Theory*, Symposium on Information Theory, London, 1960, C. Cherry, ed. Washington: Butterworths, 1961, pp. 300–316.

16. Sitar, E. J., "Computer Simulation of a System for Reading Cursive Script: Modifications and Experiment II," unpublished memorandum, Bell Telephone Laboratories, May 1, 1961.

17. Sitar, E. J., "Machine Recognition of Cursive Script: The Use of Context for Error Detection and Correction," unpublished memorandum, Bell Telephone Laboratories, Sept. 12, 1961.

18. Sitar, E. J., "A Handwriter Identification System," unpublished memorandum, Bell Telephone Laboratories, Feb. 1, 1963.

19. Harmon, L. D., "Handwriting Reader Recognizes Whole Words," *Electronics*, Aug. 24, 1962.

20. Eden, M., and Halle, M., "The Characterization of Cursive Writing," in *Information Theory*, Symposium on Information Theory, London, 1960, C. Cherry, ed. Washington: Butterworths, 1961, pp. 287–299.

21. Eden, M., "On the Formalization of Handwriting," in *Structure of Language and Its Mathematical Aspects, Proc. of Symp. in Applied Math.*, vol. XII, American Mathematical Society, 1961, pp. 83–88.

22. Eden, M., "Pattern Analysis," presented at 3rd IBM Medical Symp., Endicott, N.Y., Oct. 1961.

23. Eden, M., "Handwriting and Pattern Recognition," *IRE Trans. on Information Theory*, vol. IT-8, no. 2, Feb. 1962, pp. 160–166.

24. Eden, M., and Mermelstein, P., "Mathematical Models for the Dynamics of Handwriting Generation," *Proc. of 16th Annual Conf. on Eng. in Medicine and Biology*, Baltimore, Md., vol. 5, Nov. 1963, pp. 12–13.

25. Mermelstein, P., and Eden, M., "Experiments on Computer Recognition of Connected Handwritten Words," *Inform. Control*, vol. 7, no. 2, June 1964, pp. 255–270.

26. Mermelstein, P., and Eden, M., "A System for Automatic Recognition of Handwritten Words," *Proc. Fall Joint Computer Conf.*, 1964, pp. 333–342.

27. Fairthorne, R. A., Discussion, in *Information Theory*, C. Cherry, ed., p. 299, *op. cit.*

28. Van der Gon, J. J. D., *et al.*, "A Handwriting Simulator," *Physics in Medicine and Biology*, vol. 6, no. 3, Jan. 1962, pp. 407–414.

29. Van der Gon, J. J. D., Private communication, Dÿkzigt Hospital, Rotterdam, Netherlands.

30. Mermelstein, P., Private communication, Bell Telephone Laboratories, Murray Hill, N.J.

31. Sayre, K. M., *Recognition: A Study in the Philosophy of Artificial Intelligence*. Notre Dame, Ind.: University of Notre Dame Press, 1965.

Authors

E. Maurice Deloraine (F) received the B.S. degree in 1918 and the diploma of Ingenieur in 1920 from l'Ecole de Supérieure de Physique et Chimie in Paris. In 1949 he was granted the degree of Docteur-Ingenieur by Paris University. In 1917 he joined the French Army Signal Corps and later engaged in research work on the Eiffel Tower. He joined the International Western Electric Company in 1921, where he was responsible for part of the development in Great Britain of the first radio transatlantic telephone circuit. After acquisition of this company by ITT Corporation, he organized the International Standard Electric Corporation's Paris laboratory, which is now Laboratoire Central de Télécommunications. It was there, under his direction, that pulse code and pulse time modulation were invented. He came to the United States in 1941 to establish a laboratory for defense work for ITT's Federal Telephone and Radio Corporation. At present he serves as vice president of ISEC, president of LCT, and president of Le Matériel Téléphonique.

Alec H. Reeves received the diploma of the Imperial College of London University and also is an Associate of the City and Guilds Institute, an equivalent qualification to B.Sc. in engineering. In 1923 he joined the International Western Electric Company, which later became Standard Telephones and Cables Limited, British affiliate of ITT Corporation. He worked on the original transatlantic radiotelephone system; and later, while in the Paris laboratories of ITT, on the Madrid–South America high-frequency link, as well as on the microwave system across the English Channel. He also developed the first single-sideband high-frequency radio system. He pioneered in pulse methods, inventing pulse width and pulse time modulation systems, and quantizing pulse systems (pulse code modulation). He also introduced the use of flip-flop circuits for frequency division and pulse counting. The multipoint gas counting tube was another of his inventions. He is now in charge of exploratory research at Standard Telecommunication Laboratories.

Donald D. King (F) is director of the Electronics Research Laboratory at Aerospace Corporation, Los Angeles. He is responsible for research in many phases of electronics, including solid-state devices, millimeter waves, and electrooptics. He received the A.B. degree in engineering science in 1942 and the Ph.D. degree in physics in 1946, both from Harvard University. Subsequently he served as research fellow and as an assistant professor of applied physics at Harvard. In 1948 he moved to Johns Hopkins University, where he was active in various military research programs, and in 1955 became director of the Radiation Laboratory. In 1956 he organized the Research Division of Electronic Communications and served as its vice president and manager until 1964.

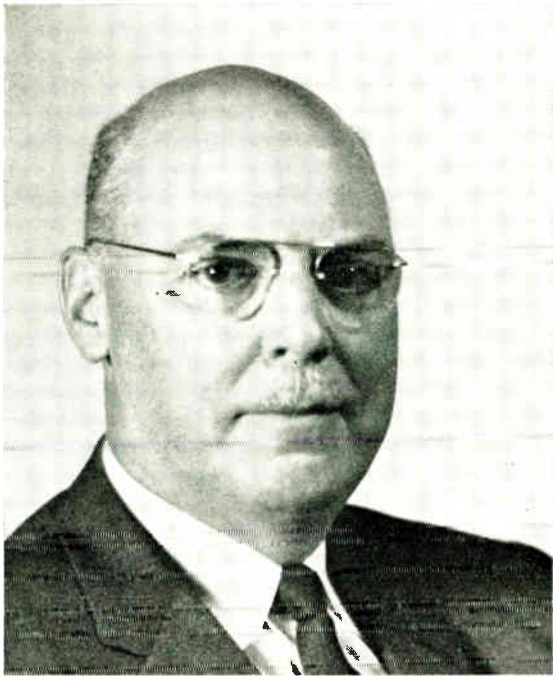
Dr. King has published a book and numerous papers on microwaves and antennas. He also served for four years as editor of TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, and in 1963 as chairman of the MTT Group. He is a member of Sigma Xi and the American Physical Society.

J. E. Gibson (SM) received the B.S. degree from Rhode Island State College in 1950 and the M.Eng. and Ph.D. degrees from Yale University in 1952 and 1956, respectively. During the academic year 1956–1957 he was an assistant professor of electrical engineering at Yale and also served as director of hydraulic servo valve research at Eastern Industries. During the summer of 1957 he joined North American Aviation as a senior research engineer in the company's Missile Development Division. He began his association with Purdue University as associate professor of electrical engineering in 1957, and in 1960 became professor of electrical engineering. In June 1961 he was appointed to his present position as director of Control and Information Systems Laboratory of Purdue's School of Electrical Engineering. His research activities have been in the areas of dual-mode control systems, nonlinear system stability theory, describing-function studies, adaptive control, and optimum control systems.

Dr. Gibson has published about 25 articles on automatic control, is co-author of the book *Control System Components* (1958), and is author of the book *Nonlinear Automatic Control* (1963). He is a member of Tau Beta Pi, Sigma Xi, ASME, and the American Society for Engineering Education.

V. I. Siforov (SM) was born May 31, 1904, in Moscow. He studied at the Institute of Electrical Engineering in Leningrad and received the degree in electrical engineering in 1929. In 1936 he received the degree of doctor of technical sciences, a doctorate that is awarded only after an individual has been active in his profession for a number of years. In 1938 he became professor and department head of the Leningrad Electrical Engineering Institute and later joined the Moscow Power Institute.

Dr. Siforov has been a corresponding member of the Academy of Science of the U.S.S.R. since 1953 and is also head of the laboratory at the Academy's Institute of Radio Engineering and Electronics. He is president of the Popov Society, an office he has held since 1954, and is also vice president of Committee 6 of the International Scientific Radio Union (URSI).



E. T. Mottram (SM) received the B.S. and M.E. degrees from Columbia University in 1927 and 1928, respectively. After a short assignment with the Western Electric Company, working on the installation and testing of special circuitry, he joined Bell Telephone Laboratories, where his early work had to do with the development of recording and reproducing equipment for sound motion pictures. Later he was concerned with developing instruments and techniques for recording and reproducing sound on disks and tape.

From 1939 to 1950 he was engaged in the development of airborne radio and radar equipment, electronic computers and bomb sights, and airborne homing missiles. In 1950 he became the director of transmission systems development, with responsibility for precision measuring equipment and a variety of television, radio, and wire communication systems. The projects on which he has worked have included submarine cable development, which has occupied an increasing proportion of his interest as more and larger-capacity systems have been developed and laid.