VOLUME 4 □ NUMBER 11 NOVEMBER 1967

IEEE spectrum

features

Departments: *please turn to the next page*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

the cover

Loran is one of the basic electronic navigations systems developed during the Second World War. It proved to be of inestimable assistance in guiding Allied warships and planes in dangerous and unfamiliar waters. Here shown is a stylized representation of the hyperbolic lines that form an essential part of a loran chart. For the study of this and other military electronic systems during World War II, see the feature beginning on page 56.

Spectral lines

Lets' support basic research. We live in an age in which there is scarcely anything that we do—personally, professionally, nationally, internationally—that is not dominated by applications of scientific knowledge that finds its roots in undirected, basic research; in knowledge that was first obtained by dedicated men who searched for fundamental truth with little idea of how it might be used. Beyond that, our hope for future health and prosperity, even survival, depends on the application of knowledge and understanding that we do not yet possess. How else will we feed the expanding multitudes, replenish depleting resources, clean a polluted environment, and protect our children from a barbaric enemy? Even more important, how will we rehabilitate our impoverished minorities?

Few will deny the oft-quoted statement of H. G. Wells that "Human history becomes more and more a race between education and catastrophe." Yet we constantly read and hear questions about the cost of basic research (the ultimate form of education). Is basic research justified? Does it contribute adequately to technology?[1] What is its cost effectiveness? And worse, we see a system of support that requires detailed proposals and expressed objectives to justify expenditures for work of an exploratory nature, perverting scientists to exaggeration and perfidy[2] who would better be used in high intellectual endeavor.

Instead of questioning how much (or little) should be spent on fundamental research,[3] we should be organizing our budgetary and our educational facilities to be sure that those few who have the capability to do first-rate research are not inhibited by mere financial, organizational, or environmental considerations. How many *research* scientists are there? Are there as many as $^1/_{100}$ of one percent of the population? How many might there be with ability and proclivities for a high level of scientific work if talent were not diverted by poor educational and social environments? Not more than a few tenths of one percent of the population, certainly. And what fraction of our budget goes to truly basic research?

How many potentially great minds are diverted because they are not intellectually challenged during the formative years? How many are lost because of social conditions that lead them away from intellectually stimulating activities? How many are lost because of the financial and social deprivations of student life, uninspired teaching, and financially poor universities? And how many, trained and ready, are lost to scientific achievement by the necessity of showing cost effectiveness in immediate results, and of generating proposals to satisfy bureaucrats, or are hampered by security restrictions?

Thus it is that our most valuable resource is dissipated. The talent that survives these impediments is tough, it is inspired, it is productive; but is it sufficient? And is it doing what it *might* do? In this regard, Warren Weaver has said: "The most important factor limiting the development of science in our country is the inadequacy of our supply of able and well trained scientists. . . . and the most serious need. . . . is at the top. We desperately need more outstanding individuals with real originality and imagination."[4]

Instead of arguing about what percentage of R & D expenditures should go to pure, or undirected, research, or whether political dissidents should get government support,[5] or where innovations originate,[1] our political representatives should be looking for means whereby science can be supported as *science*, rather than as a weapon system or a means to outshine communists, or as a cure for cancer. They should be searching for scientists to support, rather than for support for scientists.

Does this sound like a plea for anarchy in research? Certainly good judgment and uncommon common sense must be used in deployment of manpower and money for scientific endeavor, but that is not an excuse for enforcing subterfuge or exaggeration to obtain support, or pushing cost effectiveness ideas onto an activity whose ultimate utilitarian effectiveness is probably at least a generation away; whose immediate, and not least important, effect is the generation of an atmosphere in which technology thrives and students in all its branches are nurtured.

We agree that project-oriented work, so-called applied research or advanced development, requires a businesslike assessment, and must be judged by stated objectives and observable results. How to do this is quite another problem. Surely, basic research cannot be measured with the same yardstick as applied science. Our scientific leaders know pretty well what and who makes good research without the quality controls that apply to engineering and production.

The effectiveness of pure research has been amply demonstrated; it is up to us as a society to show ourselves worthy of the efforts of the talented manpower that can do first-rate research. *C. C. Cutler*

1. "First interim report on Project Hindsight," summary, June 30, 1966, revised Oct. 13, 1966, No. AD 642-400, Clearing House for Federal Scientific and Technical Information, Springfield, Va. Also, *Science*, vol. 156, pp. 1571–1577, June 23, 1967.

2. Altgilbers, T. E., "Proposal perfidy, a poverty syndrome," *IEEE Trans. Aerospace and Electronic Systems*, vol. AES-3, Jan. 1967.

3. "Secret research: tightrope act on Capital Hill," *Science*, vol. 156, p. 1718, June 30, 1967.

4. Goals for Americans, The Report of the President's Committee on National Goals. Englewood Cliffs, N.J.: Prentice-Hall, 1960.

5. "Small and NSF: a new dispute erupts," *Science*, vol. 157, pp. 1285, 1536–1539, Sept. 15, 1967.

Authors

Radio communication in the sea (page 42)

Richard K. Moore (F) is presently serving as director for the Remote Sensing Laboratory in the Center for Research in Engineering Science at the University of Kansas, Lawrence. Since 1962, he has also held the Black and Veatch professorship of electrical engineering at the university.

He received the bachelor of science degree from Washington University, St. Louis, in 1943, and the Ph.D. degree from Cornell University, Ithaca, in 1951. He was affiliated with the Radio Corporation of America, Camden, N.J., as a test equipment engineer for a short period of time and, subsequently, he served as an electronics and communications officer in the United States Navy.

He was next engaged in research, teaching, and graduate study at Cornell and Washington Universities and, in 1951, he became a member of the technical staff of the Sandia Corporation. There he was engaged in studies concerning radar return at near-vertical incidence. In 1955 Dr. Moore assumed responsibility as chairman of the Electrical Engineering Department of the University of New Mexico, where he supervised research in undersea communication and radar return.

Professor Moore's work in undersea radio communications started with his thesis research at Cornell. However, his primary efforts, since joining the University of Kansas, have been devoted to the field of radar.

He was active in the Albuquerque–Los Alamos Section of the IRE, and has participated in both IRE–IEEE and ASEE activities at the national level. He is the author of the book *Traveling Wave Engineering*, and of numerous technical papers and reports.

Disaster control coordination for large interconnected systems (page 52)

K. L. Hicks (SM), presently serving as a senior electrical engineer with the Stone and Webster Engineering Corporation, Boston, Mass., was in charge of the recent Stone and Webster study of the Northeast interconnection. The study was authorized by the 27 electric utilities in New York and New England that were affected by the power system failure of November 9, 1965. Throughout his five years as a member of the company's Electrical Division, Mr. Hicks has been concerned with power system planning and with the development of digital programs to solve problems associated with the reliability and economics of large transmission networks and power pools.

He received the bachelor of science degree in electrical engineering from Purdue University in 1949, and the master of science degree in electrical engineering from the same school in 1950. For the next 12 years, he worked in the System Planning and Special Studies Section at Sargent & Lundy, Engineers, Chicago, Ill. There he participated in many economic and technical studies for power systems, including a power pool study for the Iowa companies and stability studies of the Atomic Energy Commissions' gaseous diffusion plants at Oak Ridge, Tenn., Paducah, Ky., and Portsmouth, Ohio.

Mr. Hicks is presently an active member of both the IEEE Protective Devices Committee and the IEEE System Engineering Committee.

World War II: Electronics and the U.S. Navy (page 56)

Gordon D. Friedlander has been a staff writer for IEEE SPECTRUM for the past four years. This is his second historical article on military electronics, with particular emphasis on naval applications during the Second World War. A biographical sketch of Mr. Friedlander appears on page 111 of the February 1965 issue.

In the preparation of the two-part article, starting in this issue, and his feature piece "World War II radar: The yellow-green eye" (IEEE SPECTRUM, May 1966), Mr. Friedlander received very helpful technical background information, historical records, and excellent action photographs from **Rear Admiral Ernest M. Eller,** USN (Ret.), Director of Naval History for the Department of the Navy.

Admiral Eller was graduated from the U.S. Naval Academy in 1925, and he holds the M.S. degree in psychology from George Washington University. During World War II, he served with distinction in various staff and command assignments with the U.S. Pacific Fleet. The Legion of Merit with Combat "V" and the American Defense Service Medal are among his service awards. He is the author of numerous wartime technical reports and naval papers, and a frequent contributor to the *U.S. Naval Institute Proceedings.*

Information theory and cybernetics (page 75)

H. Marko (SM) has been professor of Communications Techniques at the Technische Hochschule, Munich, F.R. Germany, since 1962. He is also director of the school's Institute for Communications Techniques and, since 1965, he has been a member of the Administrative Committees of the Society for Communications Techniques (VDI) and of the German Society for Cybernetics.

He received the diploma in communications techniques and the Dr. Ing. degree from the Technische Hochschule Stuttgart, Germany, in 1951 and 1953 respectively. In 1953 he joined the Standard Elektrik Lorenz AG as a development engineer. There he founded and directed the Department of Fundamentals for Transmission Techniques, and he has worked on projects concerning telegraphic and radio techniques.

In 1959 he became chairman of the Committee on Information and Systems Theory of the Society for Communications Techniques within the Verband Deutscher Ingenieure. He has lectured at the Technische Hochschule Stuttgart and at Technische Hochschule Karlsruhe.

Computer-aided design (page 84)

Ronald A. Siders is a consultant in the manufacturing area on the staff of the management consulting firm of Booz, Allen & Hamilton, Inc., Cleveland, Ohio.

He was graduated from the Pennsylvania State University in 1959 with a bachelor of science degree in earth sciences, and was awarded the master of business administration degree with distinction from the Harvard Business School in 1966. Between 1962 and 1964, as a member of the Programming Systems Division of Honeywell Inc., he was engaged in work on the development of Fortran compilers and numerical control programming systems. He also served as a member of the Honeywell staff during the summer of 1965.

Since joining Booz, Allen & Hamilton, Inc., Mr. Siders has conducted a variety of assignments in computer systems design and installation, organization studies, and executive compensation. He is the lead author of *Computer Graphics—A Revolution in Design.*

Future goals of engineering in biology and medicine (page 93)

Nilo Lindgren A biographical sketch of Mr. Lindgren appears on page 196 of the March 1965 issue.

Radio communication in the sea

Oceanography and the nuclear submarine have spurred interest in very-low-frequency propagation in the sea. A comprehensive survey is offered of what is feasible in the field of underwater communications

Richard K. Moore

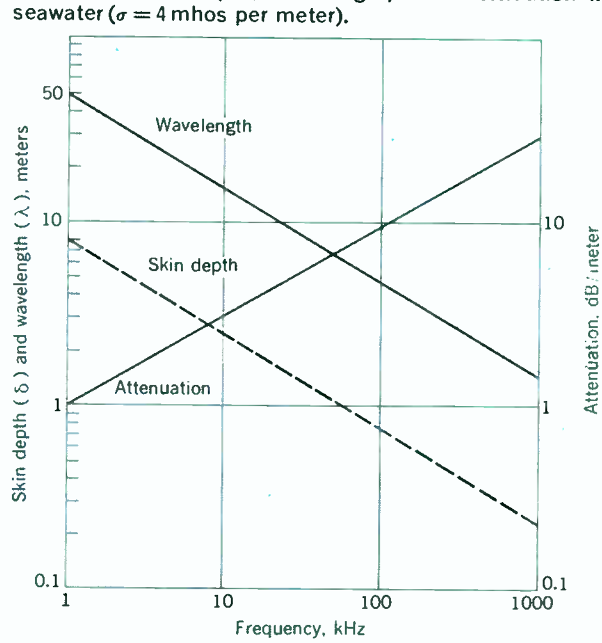
University of Kansas

Because of the electromagnetic properties of seawater, very-low-frequency communication systems are used. Surface-to-submarine, submarine-to-surface, and submarine-to-submarine propagation, as well as antennas and noise, are considered. It is shown that seawater is a good conductor and that atmospheric noise is generally more of a problem in sea communications than thermal noise. Communication ranges are limited by the effects of depth attenuation and atmospheric noise. Ranges in the tens of kilometers are possible for antennas within 5 meters of the surface; bandwidth must be less than 1 Hz and tens of kilowatts of power are required.

Since submarines and radio communication became practical at nearly the same time, it is not surprising that attempts to use radio for communication with and between submerged submarines have been made almost from the beginning of radio. At various times, experiments have been performed to show that submerged antennas are superior even for communication in air—but the results, when properly evaluated, have always been that submerged antennas are much worse. The advent in the 1950s of nuclear submarines that did not have to charge batteries gave a new impetus to the study of submarine communications since opportunities for communication with a submarine, on the surface for battery charging (or near the surface and snorkeling), now became rare. During that period also, communication between subterranean installations became a problem, and subterranean communications studies were initiated. Today, as exploitation of resources on the ocean bottom appears likely to boom, the communication problem once again becomes important.

Seawater is a poor medium for communications because it is a good conductor. Point-to-point communication within the sea is limited to such short ranges by attenuation that even for distances of a few kilometers the waves travel to the surface, are refracted, travel along the surface, and are re-refracted into the sea. Very-low frequencies (VLF) must be used to achieve any significant

FIGURE 1. Skin depth, wavelength, and attenuation in seawater ($\sigma = 4$ mhos per meter).



penetration into the sea, and information rates for submarine communication are therefore very low. Furthermore, atmospheric noise is extremely high at these low frequencies, necessitating high power even for low data rates. Actually, communication with men on the moon will be far easier than communication with a deeply submerged submarine less than 100 km away.

Electromagnetic properties of seawater

The electromagnetic properties of seawater limit communication ranges. To see this, consider a plane wave traveling in the sea, with electric field E given by

$$E = E_0 e^{j\omega t - \gamma z} \tag{1}$$

The medium properties are contained in the propagation constant γ:

$$\gamma = \sqrt{j\omega\mu(\sigma + j\omega\epsilon)} = \alpha + j\beta = \frac{1}{\delta} + j\beta \tag{2}$$

If a high enough frequency were used that the displacement current greatly exceeded conduction current, the frequency-independent value of α would become

$$\alpha = \frac{1}{\delta} = \frac{\sigma}{2}\sqrt{\frac{\mu}{\epsilon}} \tag{3}$$

A typical conductivity σ is 4 mhos per meter and permittivity ϵ is 81 ϵ_0, resulting in an α of 84 nepers per meter. The skin depth δ is therefore only 1.19 cm, and the penetration at such frequencies is essentially negligible. For this reason, low frequencies must be used in any attempt at communication through the sea.

At low frequencies, where displacement current may be neglected, the propagation constant is approximately

$$\gamma = \sqrt{j\omega\mu\sigma} = \sqrt{\frac{\omega\mu\sigma}{2}}(1 + j)$$
$$= \alpha + j\alpha = \frac{1}{\delta} + j\frac{1}{\delta} = \beta + j\beta \tag{4}$$

The amplitude of the wave therefore decreases with the distance in proportion to $e^{-z/\delta}$.

Figure 1 shows the variation of skin depth with frequency for the typical conductivity value of 4 mhos per meter. Because attenuation for one skin depth is 8.7 dB, no communication system can stand a path many skin depths long; at 10 kHz the attenuation is 87 dB for a distance of only 25 meters. Even at a frequency as low as 100 Hz, the distance for 87-dB attenuation is 250 meters, in addition to spreading attenuation and refraction losses.

Because phase and attenuation constants are the same in a conducting medium, a skin depth is also the distance for a radian of phase shift, and the wavelength is just 2π skin depths; hence wavelength is also given in Fig. 1. The vast difference between wavelength in sea and in air at low frequencies is often not fully appreciated. Note that at 10 kHz, the wavelength in air is 30 km, whereas it is only 15.7 meters in the sea. In a conducting medium, the wavelength is the distance for 2π nepers, or 55 dB of attenuation.

Sea conductivity, normally considered to be 4 mhos per meter, is in fact both temperature and salinity dependent. Thus σ varies from less than 2 in the cool and not-very-saline Arctic to 8 or more in the warm and highly saline Red Sea. Attenuation in the Arctic is therefore less than in temperate zones, so communication is feasible at greater depths or ranges. In land-locked hot seas, however, the situation is reversed.

Electromagnetic communication experiments in the sea were conducted as early as the turn of the century.[1] During World War I, both the German and Allied navies conducted experiments with low-frequency submerged reception, using low frequencies to transmit to submerged submarines.[2-5]

The theory of communication with submerged terminals was not developed until much later,[6] but the necessary tools were provided by Sommerfeld in 1909[7] in his classic paper dealing with propagation over the earth and sea. The history of above-sea propagation theory need not be related here; it covers, of course, hundreds of papers over the years since 1909. The theoretical work on the submerged propagation problem has been summarized, with historical notes, in a monograph by Banõs[8] in which numerous contributions by Wait are mentioned, along with Banõs' early work and that of McCracken.

Coupling of circuit and wave in a conducting medium has occupied numerous workers since the beginnings of experimental work. The early papers referred to were as concerned (or more so) with antennas as with propagation. Braun's early experiments with separated electrodes have been repeated, in one form or another, many times since then. Bouthillon's loop antenna, as with other loops used on submarines since World War I, was a magnetic dipole. Willoughby and Lowell's loop antenna, however, was in fact an electric antenna; they neglected to take into account the decrease in wavelength from air to sea, making their antenna many wavelengths in circumference with current distributions different from those assumed.

Much excellent work on the submerged antenna problem was done by the late O. Norgorden of the U.S. Naval Research Laboratory. Unfortunately, his reports are not widely available. Moore, in a 1951 dissertation, outlined problems of antenna analysis in conducting media, treating both small loops and insulated dipoles. The differences between antenna analysis in air and water were abstracted in a paper in 1963.[9] In the late 1940s, Tai published a number of reports in the Harvard University Cruft Laboratory series dealing with the submerged antenna problem. Wait has made numerous contributions to the theory of submerged antennas; King and Iizuka have published a series of papers detailing careful measurements of impedance for several types of submerged antennas. Other contributors too numerous to record here have added to both the theoretical and experimental literature on this subject since 1960. A number of pertinent papers may be found in a special issue of IEEE Transactions on Antennas and Propagation, published in May 1963 (vol. AP-11).

The great concern in VLF propagation in recent years has been the combined result of submarine and subterranean communication interest. There have been many papers on propagation mechanisms and atmospheric noise. Anyone seriously interested in *long-distance* submarine–subterranean communication should search this recent literature.

Propagation

Communication with submerged stations involves propagation to and above the surface, even when both terminals are submerged, owing to the extremely high attenuation for direct waves between two submerged stations. We may therefore consider the following cases of communication with submerged terminals:

1. Surface to submarine.
2. Submarine to surface.
3. Submarine to submarine (via surface).

Case 3 is a combination of cases 1 and 2.

Practical transmitting antennas in the air at VLF produce vertically polarized waves. Although these waves have curved fronts, for the purpose of treating their refraction into the sea the waves may be considered plane, an adequate approximation in any local region remote from the source. If the sea were perfectly conducting, only a vertical component of electric field would be possible in the air, as in Fig. 2(A). Because the intrinsic impedance for waves in sea is much less than that for waves in air, the actual situation is not much different from that of

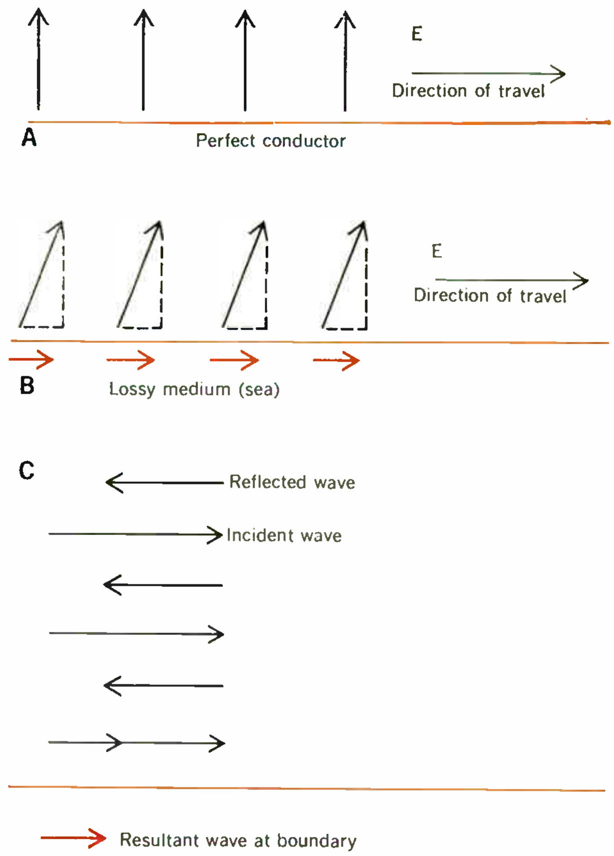


FIGURE 2. Waves in air over sea. A—Wave over perfectly conducting boundary. B—Wave over lossy medium. C—Normally incident wave near air-sea boundary.

FIGURE 3. Mode of travel of waves from transmitter that is submerged. A—From submarine to surface. B—From submarine to submarine.

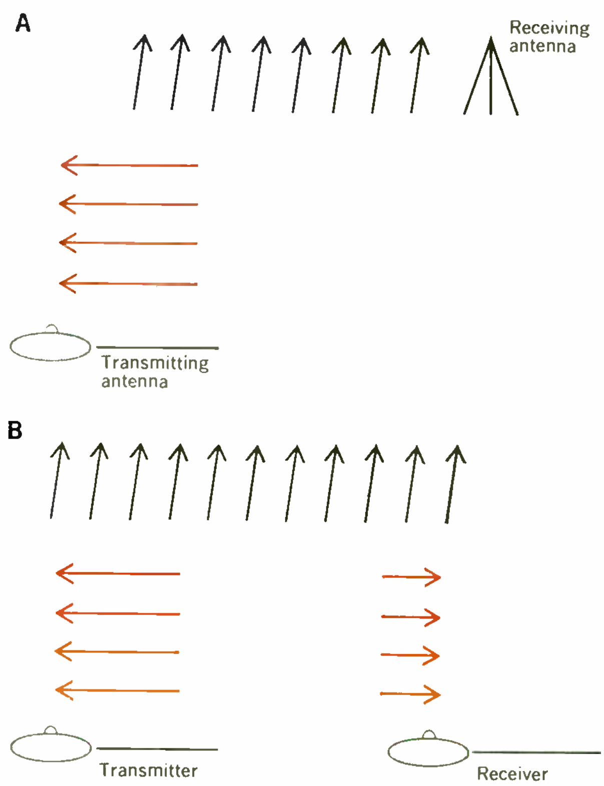


Fig. 2(A). In Fig. 2(B) a slight forward tilt of the electric field vector is indicated. The horizontal component is due to energy propagating into the sea; because it is slight, only a small part of the total power transmitted past an incremental area on the surface is diverted into the sea. If we were concerned with communication in air, we would consider this to be a loss; because we are concerned with propagation into the sea, this component is the one in which we are interested. If we could somehow launch from the ionosphere a vertically traveling horizontally polarized wave having the same incident electric field strength, as in Fig. 2(C), the reflection coefficient for the wave in the sea would be the same size as for the horizontally traveling vertically polarized wave.

Assume we launch a plane wave from beneath the surface of the sea and transmit it vertically; a small part of it would be transmitted through the surface and travel up in the air as a horizontally polarized wave. The same fraction crosses the boundary as is transmitted into the sea for a downcoming horizontally polarized wave. If this ideal plane-wave source were to transmit up at a slightly nonvertical angle, however, the wave would be refracted into a vertically polarized, almost horizontally traveling, wave; the discontinuity is so great that the angle of refraction is almost 90 degrees, even for very small angles of incidence. With practical horizontal antennas, which must be close to the surface because of attenuation in the sea, most of the energy is radiated at angles such that this refraction occurs (see Fig. 3). The wave set up above the surface is the same as that of a quadrupole radiator located above the surface and made up of two side-by-side antiphased vertical dipoles. Hence, at long distances it may be treated like a wave radiated in air, and the energy that leaks back into the water may be described by the mechanism of Fig. 2(B). Figure 3(B) shows this "up-over-and-down" mechanism for communication between two submerged antennas.

Submerged vertical dipoles, because of the null in their pattern in the direction of the dipole, are not as effective as horizontal dipoles. Because there are no strong components of the radiation directly upward, most of the energy must travel a longer path through water and also suffer a greater refraction loss. We therefore treat here only horizontal dipoles (electric and magnetic).

Surface to submarine—mathematical form

For a wave traveling in the y direction in air, the electric ($\mathbf{E}_a$) and magnetic ($\mathbf{H}_a$) fields are given by

$$\begin{aligned} \mathbf{E}_a &= \mathbf{1}_z E_0 e^{j(\omega t - \beta y)} f(\text{distance}) \\ \mathbf{H}_a &= \mathbf{1}_x \frac{E_0}{\eta_a} e^{j(\omega t - \beta y)} f(\text{distance}) \end{aligned} \tag{5}$$

where f(distance) is a suitable function of distance. Although a pure plane wave would vary with distance only exponentially, the wave over the earth represented locally here by plane components may have other distance factors; these are the subject of numerous papers and books on propagation in air. The form of this variation need not concern us here. Unit vectors $\mathbf{1}_z$ and $\mathbf{1}_x$ are in the vertical and transverse directions respectively. Intrinsic impedance η takes its air values, as indicated by subscript α; since phase-shift constant β is generally used only with its value in air, the subscript is usually omitted.

The value of $\mathbf{E}_a$ given in (5) is an approximation for a perfectly conducting sea. To obtain the horizontal component of $\mathbf{E}$, not present with perfect conductivity, note that the magnetic field is tangent to the air–sea boundary; because there is no permeability difference, it must be the same on both sides:

$$\mathbf{H}_s = \mathbf{H}_a \text{ at air–sea boundary} \tag{6}$$

The horizontal component of $\mathbf{E}$ in the sea (and in air as well) is related to $\mathbf{H}_s$ by

$$\mathbf{E}_s = \mathbf{1}_y \eta_s \mathbf{H}_a \tag{7}$$

Writing this in terms of E_0 and giving the complete expression for a point at depth d_r (whose z coordinate is $-z_r$), we find the expression for the field in the sea:

$$\mathbf{E}_s = \mathbf{1}_y \frac{\eta_s}{\eta_a} E_0 e^{j(\omega t - \beta x - \alpha d_r)} e^{-\alpha d_r} \tag{8}$$

The ratio of the intrinsic impedances in sea and air, the refraction-loss ratio, is very important; it is given by

$$\frac{\eta_s}{\eta_a} = \sqrt{\frac{j\omega\epsilon_0}{\sigma}} = \sqrt{\frac{j}{g}} \tag{9}$$

where the significant ratio of conduction current to air-dielectric displacement current is labeled g. Practical values of g are quite large; for example, it is 7.2×10^6 at 10 kHz. Thus the refracted $\mathbf{E}$ (or the horizontal component of $\mathbf{E}$ in air) is less than 1/1000 of the value of the magnitude of $\mathbf{E}$ for the wave in air. Although the magnetic field is the same in sea and air, the Poynting vector in sea is reduced by the refraction ratio.

Because the magnetic field is unaffected by refraction, one might be led to believe that a magnetic receiving antenna is inherently superior to an electric receiving antenna in this case. This is not true, however, for antennas of comparable size extract roughly the same power from the wave whether they are primarily sensitive to the electric or magnetic fields; the Poynting-vector value determines how much signal can be received for a given size antenna, not the relative value of the E- or H-field alone.

Submarine to surface—mathematical form

The fields of a submerged dipole may be obtained by modifying the Sommerfeld cylindrical-coordinate boundary value problem approach. The resulting expressions are, in general, quite complex, and have been well reported in the literature.[6,8] A few expressions will be given here to indicate the type of variation in the far field.

Baños identifies four regions: (1) asymptotic range, (2) intermediate range, (3) near-field range, and (4) vicinity of vertical axis. His intermediate and asymptotic ranges (jointly called "far" by Moore and Blair) are at distances significantly greater than the free-space wavelength/2π. In these regions the field is fairly simple, and is easily represented in local vicinities by a plane wave, as in the case of a surface antenna. In regions 3 and 4 the situation is more complicated (as it is for an antenna in air), and distance variations of different E and H components are not alike. Consequently, we restrict ourselves here to the far-field region, where the field is

$$\mathbf{E}_a = \mathbf{1}_z \frac{60\beta_a I l e^{j(\omega t - \beta_a r)}}{r} e^{-jd_t/\delta} F\left(1 - \frac{j}{\rho}\right) \times \left[\frac{-\sqrt{-j}}{\sqrt{g}} e^{-d_t/\delta} \cos\phi\right] \tag{10}$$

where I is antenna current, l is effective length of antenna, r is horizontal distance, d_t is depth of transmitting antenna, ϕ is the angle, measured in horizontal plane, from a zero along direction of submerged horizontal antenna, F is ground-wave distance attenuation factor, and ρ is the horizontal distance in units of free-space wavelength/2π.

The factor in square brackets is due to the source being submerged; the remainder of the equation is the same as for a vertical dipole in air. A readily available reference for the ground wave in air can be found in Jordan.[10] Except for the submerged-antenna factor, Eq. (10) is the same as Jordan's Eq. (16–13) with suitable approximations made for high conductivity and for a dipole at the surface. F is a complex function of the "numerical distance" ρ; it is plotted in Jordan and elsewhere.

The submerged-source factor contains four components

$$(-\sqrt{-j}) \qquad \left(\frac{1}{\sqrt{g}}\right) \qquad (e^{-d_t/\delta}) \qquad (\cos\phi) \qquad (11)$$

Phase — Refraction — Depth attenuation — Quadrupole

Note that the refraction factor for the wave *leaving* the sea is the same as that for a wave *entering* the sea. The cosine variation is that for a quadrupole; hence, the submerged horizontal antenna may be thought of as equivalent to a surface vertical quadrupole. The depth factor is the same as for a wave entering the sea. Presence of depth attenuation, the same as for a wave traveling vertically from antenna to surface, indicates that the mode of transmission for any other path between antenna and surface would be longer, with greater attenuation.

In the near-field region the depth attenuation is the same. Distance variation is different, and other components become significant—each component having its own refraction factor and angular variation.

Submarine to submarine—mathematical form

When both terminals are submerged, the near- and far-field regions may be described by one set of electric field expressions. Although these expressions are given with the ground-wave-attenuation factor included, practical ranges for two submerged terminals are so small that F may be considered always unity. These field expressions are given in the form most suitable for far-field use:

$$E_r = \frac{60\beta Il}{r} F\left\{1 - \frac{j}{F\rho}(2F - 1) - \frac{2}{\rho^2}\right\} \times \left[\frac{1}{g} e^{-(d_t+d_r)/\delta} \cos\phi\right] \qquad (12)$$

$$E_\phi = \frac{120\beta Il}{r} F\left\{\frac{1 + F}{2\rho} - \frac{jF}{2\rho^2}\right\}\left[\frac{1}{g} e^{-(d_t+d_r)/\delta} \sin\phi\right] \qquad (13)$$

In the far-field region, E_r is predominant because $1/\rho$ is small compared with unity. This means that the field of a horizontal dipole is best received by another submerged horizontal dipole aligned with the first, end to end. For a direct path, this would of course be a poor alignment, for the patterns of the dipoles would have nulls in each other's direction; with the up-over-and-down mode, however, this alignment is best.

In (12), quadrupole angular variation is evident, and

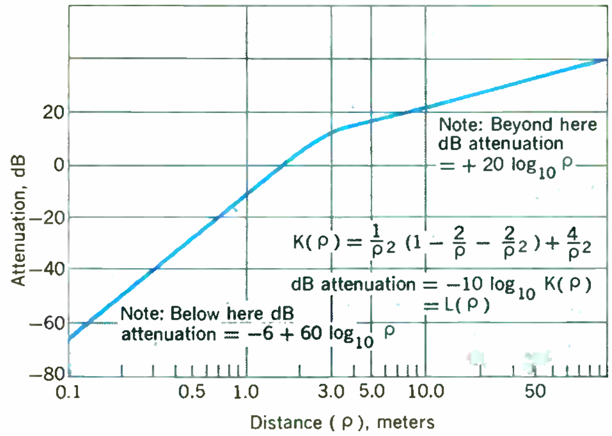


FIGURE 4. Field variations for an electric antenna.

the refraction factor is squared because it occurs both on emergence of the wave into the air and on re-refraction into the sea. Depth attenuation is the sum of the attenuations (product of the exponentials) for the upward transmitted wave and the downward received wave. Its form shows the presence of the up-over-and-down mode.

At shorter ranges ($\rho < 1$) the two components are comparable in size, with E_ϕ reaching a maximum broadside to the antenna. Because of its direction, it also corresponds with a receiving antenna parallel to the transmitting antenna. In the near-field region, the distance variation is as the inverse cube, as for a static dipole. Figure 4 presents an example of field variations, showing the transition from near- to far-field regions. It is self-explanatory.

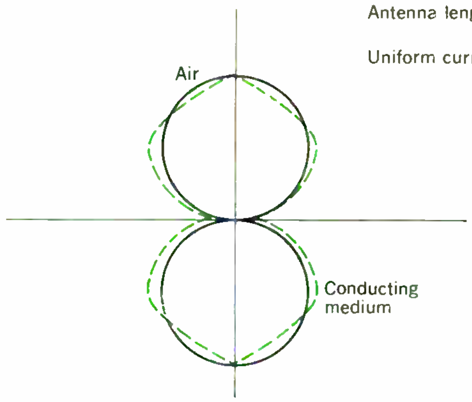
Although it is usually easier to build effective electric-dipole antennas, there are times when magnetic dipoles may be more appropriate. Fields of magnetic dipoles exhibit most of the characteristics already described for electric-dipole fields. The strongest component in the sea of the electric field of a horizontal magnetic dipole, in the far-field region, is radial; the depth attenuation factors are the same, indicating the same up-over-and-down mode of propagation; and the refraction factors are the same. Maximum radiation, however, is (in the far field) at right angles to the dipole, that is, in the plane of the loop that normally is used for a magnetic dipole. Fields beneath the surface for submerged transmitting horizontal magnetic dipoles are given by

$$E_r = \frac{60\beta^2 ISN}{r} F\left\{1 - \frac{j}{\rho} - \frac{1}{\rho^2}\right\} \times \left[-\sqrt{j}\,\frac{1}{\sqrt{g}} e^{-(d_t+d_r)/\delta} \sin\phi\right] \qquad (14)$$

$$E_\phi = \frac{120\beta^2 ISN}{r} F\left\{1 - \frac{j}{\rho} - \frac{j}{\rho^2}\right\} \times \left[-\sqrt{-j}\,\frac{1}{g} e^{-(d_t+d_r)/\delta} \cos\phi\right] \qquad (15)$$

$$H_\phi = \frac{60\beta^2 ISN}{\eta_0 r} F\left\{1 - \frac{j}{\rho}\right\}\left[e^{-(d_t+d_r)/\delta} \sin\phi\right] \qquad (16)$$

$$H_r = \frac{j120\beta ISN}{\eta_0 r} F\left\{\frac{1}{\rho} - \frac{j}{\rho^2}\right\}\left[e^{-(d_t+d_r)/\delta} \cos\phi\right] \qquad (17)$$



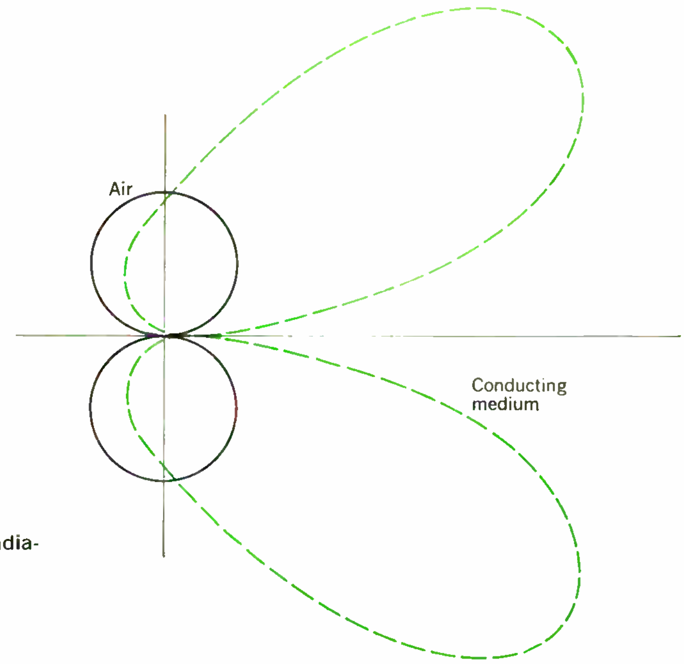


FIGURE 5. Effect of coordinate origin on radiation pattern.

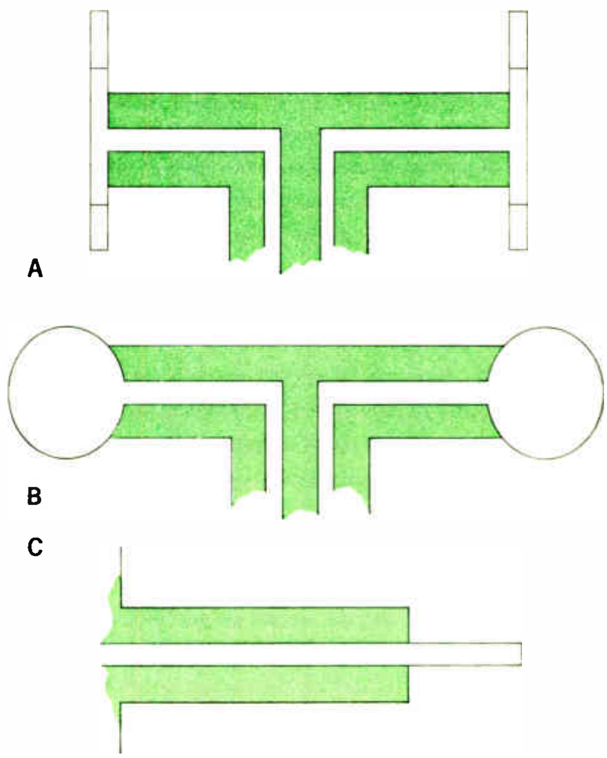


FIGURE 6. Submerged insulated antennas. A—Flat-plate termination. B—Spherical termination. C—Trailing-wire configuration.

The dipole is represented by a planar loop of area S, with N turns carrying current I.

Antennas

Waves are coupled to current in circuits by antennas. All of the preceding propagation expressions have been predicated on dipole sources. In air, more elaborate sources are not feasible at the low frequencies necessary for submarine communication. Nevertheless, these electric and magnetic dipoles in the sea take on interesting forms different from those common in air, primarily because attenuation and phase shift are equal in the conducting medium; in air, attenuation between the parts of an antenna is negligible.

The differences between antennas in air and in conducting media have been discussed in detail earlier by Moore.[9] An illustration of problems that arise is shown in Fig. 5. In air, because of no attenuation, antenna patterns or polar diagrams are normally used to illustrate directional properties: the location of the origin of coordinates used in calculating these diagrams is immaterial as long as it is somewhere near the antenna. When one attempts to make a similar calculation for an antenna in the sea, he discovers that the choice of an origin of coordinates is all-important, for the pattern depends on it, as shown in the illustration. The reason is that attenuation between one side of the antenna and the other is so great that the major contribution to the field at any point is primarily due to the nearest part of the antenna. The net result is that antenna patterns are meaningless in conducting media.

Electric antennas in conducting media may take the same form as short electric antennas in air, but the rapid decay of antenna current along a conductor suggests that there must be a better approach. The approach commonly used is to send current through insulated wires to electrodes or contacts within the seawater. A current then flows between these contacts and radiates. Antennas in air really do much the same, except conduction current flows in the sea whereas displacement current flows in air; further, insulation is unnecessary in air.

Figure 6 shows several ways to make an electric antenna for the sea. A dipole with flat plates to contact the water is shown in Fig. 6(A). If the plates are large compared with the insulation diameter, which is small relative to wavelength in the sea, a current is set up in the sea equal and opposite to that in the wire, and the electric fields in sea and wire are parallel. This antenna may be analyzed as a coaxial transmission line, with the seawater forming the outer conductor. Another way to terminate this "transmission line" is shown in Fig. 6(B); spheres and cylinders have both been used in the configuration.

The antenna form of Fig. 6(C) is commonly used as a trailing wire. It was called "short-circuit coaxial" by Moore, with the bare end serving to short-circuit the "transmission line." Because a trailing wire can be very long, this is often the most practical form for a submarine antenna. The antenna was independently analyzed by Norgorden and by Moore. The wavelength in the coaxial antenna is intermediate between that in sea and that in air; consequently, an antenna that is long in terms of sea wavelengths may still be a small fraction of a wavelength long as far as the coaxial transmission line mode is concerned. Any of the antennas of Fig. 6 may carry uniform currents even if they are much longer than a sea wavelength, whereas long antennas with uniform currents are not feasible in air.

If one assumes that any of the techniques shown in Fig. 6 short-circuits the equivalent coaxial transmission line, the input resistance R_{in} is given by

$$R_{in} = \omega\mu l/8 \tag{18}$$

where l is the length of the antenna. At first it seems surprising that the diameter of the antenna does not enter into the resistance. This is explained by noting that the current spreads out over a large cross section determined by the water wavelength, and the cross section of the insulated section detracts a negligible amount from the total cross-sectional area; however, the resistance is very low. For example, a 100-meter-long antenna at 10 kHz has a resistance of only about an ohm.

The reactance of an antenna of this type is given by

$$X_{in} = \frac{\omega\mu l}{2\pi}\left[0.116 + \ln(\delta/a\sqrt{2})\right] \tag{19}$$

where a is the conductor radius. Even though (δ/a) may be quite large, the reactance will not greatly exceed the resistance; tuning it out is relatively easy.

Loop antennas (magnetic dipoles) have often been used for submerged reception, and at times have been proposed for transmission. It can be shown that electric and magnetic antennas with comparable dimensions have about the same radiation properties; therefore, the trailing wire, which can be made quite long, is usually superior to the loop. The impedance of a loop can be made higher by using more turns; in many cases this is a significant advantage, offsetting the greater ease with which an electric antenna can be made large.

Apparently the first analysis of a submerged loop was due to Moore,[6] but the first published analysis in a journal was by Wait.[11] Moore, and later in more detail Kraichman,[12] treated a loop whose conductors are insulated, so there is no leakage current through the conducting sea. Wait considered the loop enclosed in a spherical insulating cavity; Kraichman showed that the insulating sphere has little effect on the radiation resistance provided the loop diameter is nearly as great as that of its surrounding insulating sphere. The effectiveness of a loop can be improved by loading it with high-permeability material; Williams[13] showed the advantage of loading with a high-permeability sphere.

The radiation resistance R_{rad}, for a loop of radius a, small compared with a wavelength in the sea, is

$$R_{rad} \approx \frac{4\omega\mu a}{3}\left(\frac{a}{\delta}\right)^2 N^2 \tag{20}$$

and the reactance x is

$$X = \omega\mu a\left[K(k) - 2 - \frac{\pi}{3}(\beta_s a)^3 + \frac{4}{15}(\beta_s a)^4 + \dots\right]N^2 \tag{21}$$

where $K(k)$ is an elliptic integral and N is the number of turns.

An electric dipole that is easier to match than the insulated antennas described earlier can be made by using a magnetic current loop, achieved in the form of a toroidal coil with a high-permeability core. This configuration can also be thought of as a transformer, with the current in the conducting sea being the one-turn secondary. Matching to this transformer may sometimes be easier, but, like the loop, it suffers because it cannot easily be made physically large. Considerable improvement in the performance of this antenna results from loading it with a long thin conductor.[14]

Arrays of submerged antennas are possible, but they do not result in increased directivity as do arrays in air. The attenuation between parts of the array, discussed previously, explains this. Why, then, should one consider arrays in conducting media? The reason is that this provides another method of solving the problem of matching low-impedance electric antennas. Because of the attenuation in the sea, antennas only a fraction of a sea wavelength apart have little interaction.

Noise and system considerations

The limiting factor in any communication system is noise. Submarine communications are subject not only to receiver noise, but to strong atmospheric noise, for low frequencies permit transmission of thunderstorm atmospherics to worldwide distances. In this section we shall compute the ranges possible for submarine communication under various conditions, emphasizing the limiting role played by atmospheric noise and by depth attenuation.

Receiver bandwidth determines the noise power that must be overcome by a communication system. Because of low carrier frequencies, the bandwidth for submarine communication is inherently limited; but an even more significant limitation is the requirement for narrow-band receivers if practical ranges are to be achieved. Thus voice bandwidths are out of the question. For many purposes the bandwidth required for noise reduction may be narrow even for code transmission. A 100-Hz bandwidth might seem minimal, but two orders of magnitude narrower are often needed for adequate noise performance, with a severely restricted communication rate.

Because noise power is proportional to bandwidth, the field expressions evaluated earlier should be combined with antenna radiation resistance figures to give ratios of transmitted to received power. The current in the transmitting antenna must be also expressed in terms of power to find this ratio.

When the insulated wire (short-circuited coaxial) antenna is used, the radiation resistance is given by Eq. (18). This allows the current to be expressed in terms of the power transmitted as

$$I = 2\sqrt{\frac{2W_t}{\omega\mu l_t}}$$

The resulting expression for the electric field above the surface in terms of transmitter power is obtained by substituting this value in (10):

$$E_a = \frac{120}{gr}\sqrt{2\sigma l_t W_t}\, e^{-d_t/\delta} F \cos\phi \qquad (22)$$

Here we have assumed $1/\rho \ll 1$. The field in the sea at a distance is just this surface field multiplied by

$$\frac{1}{\sqrt{g}} e^{-d_r/\delta}$$

the product of refraction and depth attenuation factors.

Our example to illustrate magnitudes involved in the submarine communication problem is for

Conductivity: 4 mhos/m
Power transmitted: 10 kW
Antenna length: 50 meters
Antenna depths: multiples of 2.5 meters
Receiver bandwidth: 1 Hz
Receiver noise figure: ideal
Frequency: 10 kHz

Using these figures, the field at the surface 20 km from the submerged antenna (depth 5 meters) is 0.225 μV/m—not a very large value for a 10-kW transmitter at such a short distance. The field at a receiving antenna also submerged 5 meters is only 1.13×10^{-11} V/m.

Determining the ratio of received to transmitted power for a receiving antenna in air requires knowledge of the antenna's impedance. The actual impedance of a vertical electric antenna in air is strongly influenced by ground properties in its vicinity. For our example, let us use the expression for the ideal short monopole above a perfectly conducting ground:

$$R_r = 40\pi^2 l_r^2/\lambda^2 \qquad (23)$$

Substituting this in (22), the power ratio between a submerged transmitter and a receiver in air is found to be

$$W_r = \frac{V^2}{4R_r} = \frac{(E_a l_r/2)^2}{4R_r} \qquad (24)$$

or

$$W_r = \frac{W_t l_t}{80\pi^2\sigma r^2} e^{-2d_t/\delta} F^2 \cos^2\phi$$

assuming that the receiver is matched to the receiving antenna. Note the interesting result that the length of the receiving antenna does not matter. Of course, that length determines the reactance that must be tuned and the low resistance that must be matched, so it really does matter!

Using (24) in the example we calculate that the received power is 7.2×10^{-9} watt. The noise in a 1-Hz bandwidth in an ideal receiver is 0.4×10^{-20} watt; the signal-to-noise ratio is 1.8×10^{12}. If only internal receiver noise were important, it is obvious that the submarine-to-surface path could be much longer than the 20 km chosen for the example. For instance, at a 100-km distance, the signal-to-noise ratio is still 7.2×10^{10}. Beyond that distance, one should use appropriate long-distance VLF propagation theory to describe attenuation, for the ionosphere can no longer be neglected.

The ratio of received to transmitted power for two *submerged* horizontal trailing wires may be calculated from (12) and (18) to be

$$\frac{W_r}{W_t} = \frac{(El_r)^2}{4l_t^2 R_r R_t} = \frac{4l_r l_t}{\pi^2 g^2 r^2} F^2 \cos^2\phi \times \left[1 - \frac{j}{F\rho}(2F - 1) - \frac{2}{\rho^2}\right]^2 e^{-2(d_r+d_t)/\delta} \qquad (25)$$

Using Eq. (25) and the parameters of the previously quoted example, except for zero depth, the signal-to-noise ratio is 1.25×10^5. When the sum of the depths of transmitting and receiving antennas is 5 meters, the ratio becomes 2280; when the sum is 10 meters, the ratio goes down to only 42.

If only thermal noise were important, communication would be much easier than in practice, for atmospheric noise is a more important factor influencing communication range than thermal noise. This is true because static from lightning discharges has a spectrum with its most important components in the range of frequencies also suitable for submarine communication. Also, propagation at these frequencies permits signals from lightning discharges to be propagated throughout the world.

Atmospheric noise has been measured at many points on earth by numerous observers. For the frequency range down to 10 kHz, the data have been compiled by an International Working Group of the CCIR (International Radio Consultative Committee) in the form of a pamphlet.[15] World maps with isonoise-field lines at 1 MHz are presented for four-hour time blocks throughout the day for each of four seasons, along with information allowing conversion of the 1-MHz information mapped for other frequencies.

At 1 MHz, diurnal variations as much as 50 dB are found at some stations, but at 10 kHz neither diurnal nor seasonal variations are very large, with extremes of about 12 dB for the total range. At frequencies below a few hundred kilohertz, noise level increases strongly with decreasing frequency. In fact, noise power varies from about $1/f^2$ to $1/f^6$ in the low-frequency range; a severe penalty is paid for going to low frequencies for tasks for which higher frequencies are suitable.

Because noise increases strongly with decreasing frequency and depth attenuation increases strongly with increasing frequency, clearly an optimum frequency exists for which the two effects balance out. Owing to the variability of noise, this optimum is not always the same even at a given location, and varies from location to location.

Let the signal-to-noise ratio in the receiver be S; then

$$S = W_r/N \qquad (26)$$

If N is given in terms of its frequency variation by

$$N = N_0/f^n \qquad (27)$$

the frequency variation for S/W_t for two submerged terminals may be found using (25):

$$S/W_t \propto f^{n+2} e^{-2d/\delta} \qquad (28)$$

Differentiating with respect to frequency and setting the derivative equal to zero yields the optimum frequency condition:

$$n + 2 = d/\delta \qquad (29)$$

That is, the optimum frequency makes the skin depth just equal the sum of the antenna depths divided by $(n + 2)$.

Communicating from beneath the surface to the surface requires that the signal from the submerged source must at least equal the noise level from atmospheric noise above the surface. Communication from one submerged terminal to another requires that the same condition be met. Owing to the submerged transmitter, the signal at the surface is refracted; the signal-to-noise ratio above the surface must be adequate if it is to be adequate at a submerged receiver.

These effects are best illustrated by examples, and several are presented in the ensuing discussion. Consider first the same situation in the preceding example: a field of 0.225 μV/m at a distance of 20 km from the transmitter. If this occurs during the 16–20-hour GMT time block at a point in mid-Pacific at 180° longitude and 40° north latitude during the spring, the noise voltage is 5.3 μV/m for the 1-Hz bandwidth postulated. This means that the signal-to-atmospheric-noise ratio is -27.4 dB, hardly adequate for communication. If only thermal noise had been considered, the signal-to-noise ratio would have been $+122.5$ dB, which is quite a difference.

Increasing the transmitter power to 50 kW and doubling the antenna length to 100 meters only raises the signal-to-atmospheric noise ratio S to -17.4 dB, still inadequate. If the transmitter is at a depth of only 2.5 meters, the signal is raised another 8.7 dB, no appreciable improvement. In fact, the noise power is so great that the signal-to-noise ratio is less than unity even if there is no depth attenuation, but only refraction loss at the surface. The noise is so intense that an antenna 20 meters high has a noise voltage of 0.5 mV for 100-Hz bandwidth.

Obviously, other factors must be considered. Actually, for this example, the signal-to-noise ratio can be improved by going either up or down in frequency. Going up in frequency reduces noise, as indicated in the CCIR report. Going down to about 3 kHz reduces noise because of increased absorption there.[16] The noise at 2 to 3 kHz, however, is only of the order of 10–20 dB lower than at 10 kHz; only larger reductions possible by increasing the frequency are considered in the remaining examples. For the examples that follow,

Range: 20 km
Power transmitted: 50 kW
Antenna length: 100 meters
Antenna depth: 2.5 meters and 5 meters
Receiver bandwidth: 1 Hz
Receiver noise figure: ideal

The first example is for the same mid-Pacific location (180°, 40°N) for the 12–16-hour GMT winter period. The technique for finding optimum frequency gives, in accordance with (29), $d/\delta = 4.9$, which corresponds to a frequency of 63 kHz for a depth of 5 meters. At 60 kHz the noise voltage is down to 23 dB below a microvolt, and the signal voltage is only 18.6 dB below a microvolt; the signal-to-atmospheric-noise ratio is 4.4 dB.

The effect of a lower frequency is shown by a 20-kHz example, for which the noise voltage is up to 0.95 μV and the signal is only up to 0.63 μV; S is -3.5 dB, an unsatisfactory level.

On the other hand, if the antenna depth is 2.5 meters instead of 5, the 60-kHz signal is increased by 20.7 dB; S at 60 kHz is 24.4 dB, making it possible to contemplate increased bandwidth, or power and antenna reduction. With the 2.5-meter depth at 20 kHz, the improvement is not too great, but sufficient to raise S to $+8.7$ dB, an acceptable value.

The second example is for an area adjacent to Alaska (165°W, 65°N) for the 12–16-hour GMT block in summer. Here the optimum frequency is (for 5 meters) almost 135 kHz. At 60 kHz the noise is 15 dB below a microvolt; S is -3.6 dB. For a 2.5-meter depth, however, the 60-kHz value for S is $+17.1$ dB.

A third example is for the same Alaskan location, but for the 16–20 time block in winter. In this case, the optimum frequency is 45 kHz, and S is $+4.5$ dB for a 5-meter depth. For a 2.5-meter depth, this increases to 22.9 dB.

Because the signal from the submerged transmitter must at least equal the noise above the surface where a submerged receiver is located, and further, the signal from a surface transmitter must also at least equal this noise, communication that is possible at all is feasible at least to a depth where the noise signal received from atmospherics equals the internal receiver noise. For a given frequency, therefore, we can calculate depths at which communication should be feasible for any source powerful enough to communicate by determining the depth required for attenuation and refraction losses to reduce the atmospheric noise to the receiver noise level. For the previous examples and for a unity-noise-figure receiver, these depths have been calculated as 3.06 meters at 60 kHz in mid-Pacific winter 12–16 GMT, and 7.1 meters at 45 kHz for Alaska winter 16–20 GMT. Of course, communication may be possible to a greater depth if the signal at the surface is much higher than the atmospheric noise there.

Common fallacies

Several fallacies appear from time to time in discussions of radio communication in the sea. Although some of these have been treated in this article where they have arisen, it is well to reiterate their fallacious nature.

Fallacy 1: Seawater is a dielectric at radio frequencies.

Discussion: The transition between dielectric and conductor occurs where conduction and displacement currents are equal. For seawater with conductivity of 4 mhos per meter, this occurs at 890 MHz; for all practical communication purposes, the sea is a good conductor.

Fallacy 2: Atmospheric noise can be neglected in radio communication through the sea.

Discussion: The preceding section has shown that atmospheric noise profoundly affects communication ranges and frequencies in the sea. The optimum frequency in the absence of atmospheric noise effects would be obtained by using Eq. (29) with $n = 0$—and it would be much lower than that found when noise is taken into account. Furthermore, noise is the limiting factor that requires large power for receiving antennas near the surface, rather than depth attenuation and refraction. Thus both surface and submerged transmitters must be much larger because of atmospheric noise than they would need to be otherwise. Also, atmospheric noise requires submerged transmitting antennas to be near the surface and large, whereas (25) would indicate that depth and length of transmitting and receiving antennas for submerged-to-submerged communication have the same effect.

Fallacy 3: The tangential component of E in air, and thus the downgoing E in water, can be increased by transmitting vertically downward with a high-altitude horizontal antenna.

Discussion: The tangential component of the field was shown to be the same for downcoming and horizontally traveling waves, because the reflection coefficient for downcoming waves is so near unity that most of the incident field is canceled by the reflected field in air.

Fallacy 4: Because a dc pulse propagates with less attenuation than an exponential pulse, it should be superior for communication in the sea.

Discussion: Propagation of electromagnetic waves in a conductor follows the same equation as heat conduction and diffusion. The highest point of a pulse that satisfies the diffusion equation is indeed attenuated more slowly than an exponential; the reason is that at longer distances the peak is due to lower-frequency components of the original pulse. Fourier analysis of the transmitted pulse shows frequency components over a wide range. The higher frequencies are attenuated in shorter distances; at longer distance the by-that-time sloppy pulse is primarily due to the lower of the original frequencies. Thus energy that went into the higher-frequency components has been wasted, and it would have been better not to transmit them at all.[17]

Fallacy 5: A large-loop antenna that is small compared with a long wavelength in air is also small compared with a wavelength in sea.

Discussion: The wavelength in sea is, as has been shown, much shorter than in air. A loop antenna the size of a submarine would therefore seem small compared with a wavelength in air, and could be treated there as a magnetic dipole. In the sea, on the other hand, such a loop is likely to be many wavelengths on a side; furthermore, attenuation along the antenna and between its top and bottom tend to make it appear more like an electric dipole located near its shallowest point.

Conclusions

Radio communication from, to, and between submerged antennas involves propagation paths through the highly conducting sea with segments vertically above the submerged antennas and a segment along the surface like that for a vertical antenna in air. In fact, a submerged horizontal electric dipole is equivalent in its fields to a weaker vertical quadrupole at the surface. Because of the strong attenuation in the sea, very low frequencies are required for such communication.

Antennas are quite different in the sea, owing to the effects of high conductivity. Both insulated wire electric antennas and loop magnetic antennas can be used successfully, but the radiation resistances are quite low and matching problems often exist.

Communication ranges are limited both by the effects of depth attenuation and by atmospheric noise. Because of these factors, communication from submerged transmitters is necessarily limited to rather short ranges, even when antennas are close to the surface. Ranges in the tens of kilometers are possible for antennas within 5 meters of the surface if bandwidths are restricted to the order of 1 Hz, transmitting trailing-wire antennas are used, and powers of tens of kilowatts are available. For communication from surface transmitters to submarines, much longer ranges are possible, requiring very large transmitters to overcome atmospheric noise. Because of atmospherics and attenuation, carrier frequencies most useful are in the tens of kilohertz range.

Communication with submerged terminals is possible provided sufficiently large power is coupled with sufficiently low communication rates. Because the ranges are restricted by fundamental attenuation considerations and by noise that must be considered an unchangeable part of the environment, significant breakthroughs are not to be expected in submarine radio communication. Nevertheless, both submarines and those exploring and exploiting the ocean bottom can use radio for essential communications; for two-way communication the transmitting antenna will ordinarily be quite close to the surface.

Direct communication between surface vessels and ocean-bottom working parties could, if need be, use extremely low-frequency carriers with the surface ship also using a submerged antenna if the path is sufficiently short that direct attenuation from one antenna to the other is not prohibitive. The submerged transmitting antenna from the surface ship would therefore only have to compete with atmospheric noise that had undergone refraction loss of the order of 60 dB and its own signal would not have to experience that loss; the power required would therefore be reduced accordingly. Presumably such communication would have to be at frequencies well below 10 kHz to overcome direct transmission loss. In most cases underwater sound appears superior for this application.

REFERENCES

1. Braun, F., *Drahtlose Telegraphie durch Wasser und Luft.* Leipzig: Veit and Co., 1901.
2. Taylor, A. H., "Subterranean and submarine antennas," *Proc. IRE*, vol. 7, pp. 337 and 559, 1919.
3. Batcher, R. R., "Loop antenna for submarines," *Wireless Age*, vol. 7, p. 28, 1920.
4. Willoughby, J. A., and Lowell, P. D., "Development of loop aerials for submarine radio communication," *Phys. Rev.*, vol. 14, p. 193, 1919.
5. Bouthillon, L., "Contribution a l'étude des radiocommunications sous-marines," *Rev. Gen. Elec.*, vol. 7, p. 696, 1920.
6. Moore, R. K., "Theory of radio communication between submerged submarines," Ph.D. dissertation, Cornell University, Ithaca, N.Y., 1951.
7. Sommerfeld, A., "Uber die ausbreitung elektromagnetischer wellen uber ein eben erde," *Ann. Phys.*, ser. 4, vol. 28, pp. 665–736, 1909.
8. Banõs, A., Jr., *Dipole Radiation in the Presence of a Conducting Half-Space.* New York: Pergamon, 1966.
9. Moore, R. K., "Effects of a surrounding conducting medium on antenna analyses," *IEEE Trans. Antennas and Propation*, vol. AP-11, pp. 216–222, May 1963.
10. Jordan, E. C., *Electromagnetic Waves and Radiating Systems.* Englewood Cliffs, N.J.: Prentice-Hall, 1950, Chap. 16.
11. Wait, J. R., "Insulated loop antenna immersed in a conducting medium," *J. Res. Nat'l Bur. Std. D*, vol. 59D, pp. 133–137, 1957.
12. Kraichman, M. D., "Impedance of a circular loop in an infinite conducting medium," *J. Res. Nat'l Bur Std. D*, vol. 66D, pp. 499–503, 1962.
13. Williams, R. H., "Insulated and loaded loop antenna immersed in a conducting medium," *J. Res. Nat'l Bur. Std. D*, vol. 69D, 1965.
14. Williams, R. H., Kelly, R. D., and Cowan, W. T., "Small prolate spheroidal antenna in a dissipative medium," *J. Res. Nat'l Bur. Std. D*, vol. 69D, pp. 1003–1010, 1965.
15. "World distribution and characteristics of atmospheric radio noise," CCIR Rept. 322, Documents of the Xth Plenary Assembly, Geneva, 1963, published by International Telecommunications Union, Geneva, 1964.
16. Maxwell, E. L., and Stone, D. L., "Natural noise fields from 1 cps to 100 kc," *IEEE Trans. Antennas and Propagation*, vol. AP-11, pp. 339–343, May 1963.
17. Moore, R. K., and Anderson, W. L., "Frequency spectra of transient electromagnetic pulses in a conducting medium," *IRE Trans. Antennas and Propagation*, vol. AP-9, pp. 603–607, Nov. 1960.

Disaster control coordination for large interconnected systems

Since widespread power failures have proved to be more than a remote possibility, new approaches to interconnection problems and coordination of disaster control procedures have evolved

K. L. Hicks *Stone and Webster Engineering Corporation*

A new philosophy for the analysis and design of large electric power interconnections has evolved following the Northeast power failure of November 9, 1965. Most contingencies that can cause widespread outages occur too infrequently to be included in the criteria for system design. Nevertheless, systems should be tested for combinations of events that cause system instability and separation so that the consequences of these unlikely occurrences can be evaluated and system designers can provide disaster control procedures to limit the extent and duration of system outages. The design and disaster control procedures should provide successive lines of defense against increasingly severe and unlikely events. Disaster control procedures can be coordinated so that governor action, load shedding, and system separation can be integrated according to time and frequency in such a way that maximum reliability and security will be provided.

Most of the power systems in the United States and Canada are part of the same interconnected system. More than 200 000 000 kW of generating capacity operates in synchronism, and the stored energy and spinning reserve of this network will replace large amounts of lost generation with hardly noticeable frequency deviations. If a generator anywhere in this interconnection trips off the line, its loss is immediately supplied from the remainder of the interconnection until the area in which the loss has occurred is able to increase generation to restore its tie lines to schedule.

Widespread outage in this large interconnection is invariably accompanied by separation of the affected area from the network. Traditionally, system design criteria have included design contingencies that the system must withstand without loss of load or damage to equipment. These contingencies include such events as permanent three-phase faults or loss of the most heavily loaded line. Other more severe combinations of contingencies may be included as design criteria, but there is a practical limit to the amount of money that can be spent to prevent system separation due to extremely unlikely events. Nevertheless, it is proposed that system designs be tested for these unlikely events that cause system separation so that the power system designer can specify and coordinate disaster control procedures to limit the extent and duration of system outages. Specific studies of system response to loss of generation will aid in the selection of critical frequencies to be considered in the coordination of these disaster control procedures.

System behavior during abnormal conditions

A study of the Northeast interconnection following the power failure of November 9, 1965, shows that any part of the interconnected system can be isolated by severe disturbances and left with rapidly decreasing frequency.[1] The behavior of certain essential elements of a power system during abnormal conditions of frequency and voltage will control the selection of basic parameters of frequency and time in the coordination of load shedding, and system separation.

Governor response. Following a sudden unbalance in load and generation, governor response is very effective in arresting the initial decay of frequency. The initial rate of change of frequency will be a function of the excess load on the system and the inertia of its generators and can be expressed as

$$R = \frac{60}{2H}\,\frac{L - G}{G} \tag{1}$$

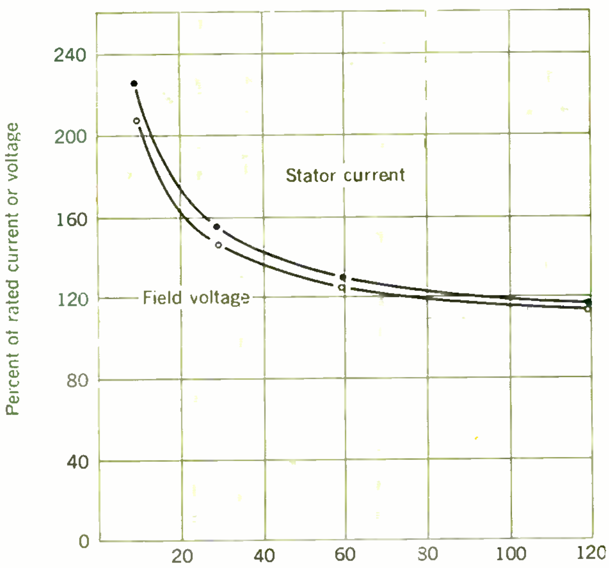


FIGURE 1. Generator short-time thermal capability. (From USA Standard C50.13—1965.)

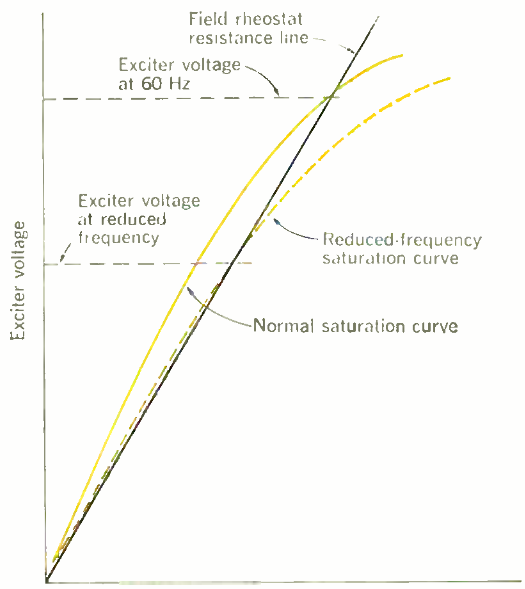


FIGURE 2. Self-excited exciter saturation at normal and reduced frequencies.

where

R = rate of change of frequency, hertz per second
H = average inertia constant of the generators, per unit
L = system load, megawatts
G = system generation, megawatts

If the unbalance in load and generation is expressed as

$$U = \frac{L - G}{G} \tag{2}$$

Eq. (1) reduces to

$$R = \frac{30U}{H} \tag{3}$$

For example, if a system that has separated from the interconnection is left with 9000 MW of generation and 10 000 MW of load, the initial rate of change of frequency will be

$$R = \frac{30}{4}\left(\frac{10\,000 - 9000}{9000}\right)$$

$$= 0.833 \text{ hertz per second}$$

assuming an H of 4.0 per unit for a representative system.

If there is sufficient spinning reserve in the remaining machines and if the reserve can be picked up quickly and sustained, this isolated system could theoretically survive the disturbance without loss of load; but consider what may happen to some of the equipment during this recovery period.

The initial rapid decay of frequency will be retarded within a few seconds by governor action; but this is a temporary holding action, and the additional mechanical output will be supplied at the expense of stored energy in the boilers. The firing rate of the boilers must catch up rapidly, or the system load must be reduced quickly, if this initial holding action is not to be abandoned.

Voltage regulators. Furthermore, an isolated system that has lost 10 percent of its generation will very likely experience low voltage and may require the voltage regulators of one or more generators to go to maximum boost to attempt to hold voltage. If the voltage regulator is in service, it will attempt to maintain generator terminal voltage; but unless something else happens to alleviate the low-voltage condition, the var output of the generator must be reduced again within about one minute, as shown in Fig. 1, to protect both the regulator and generator field from overloads.

If the regulator is out of service for maintenance or if its output must be reduced to prevent overloads after

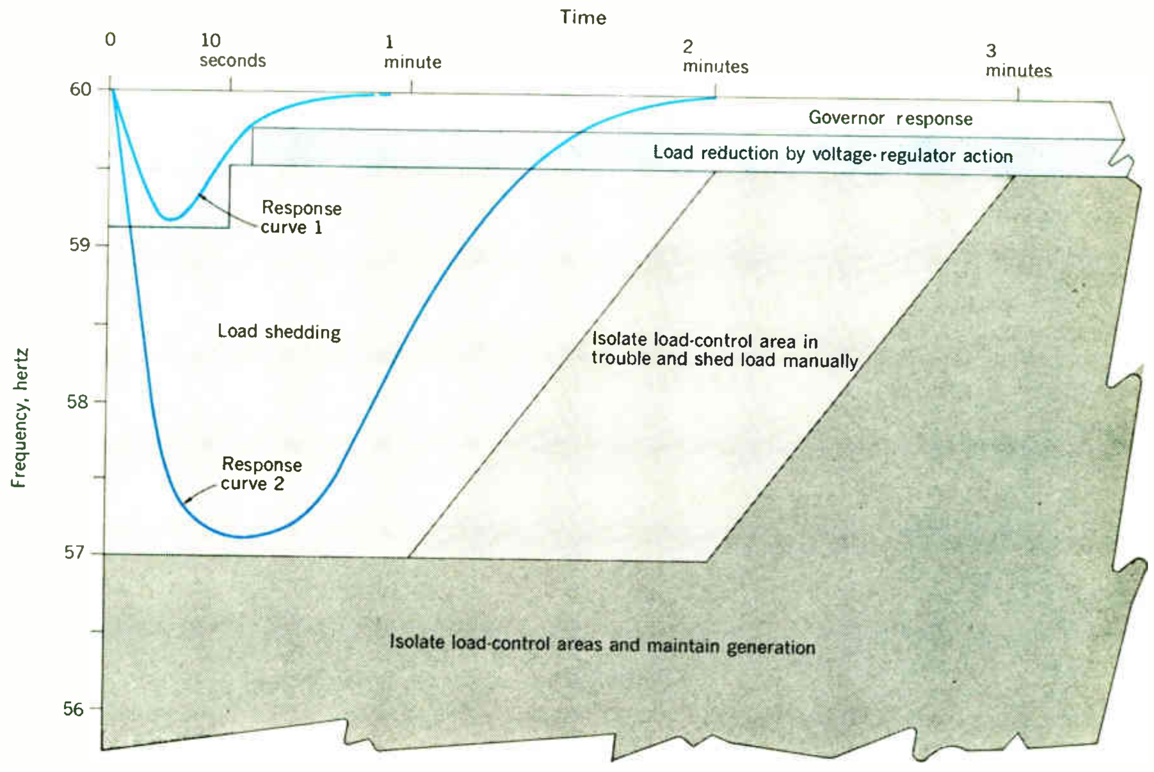


FIGURE 3. Coordination diagram for disaster control procedures.

one minute, the machine will very likely trip on loss of field. As shown in Fig. 2, the output voltage of a shunt-excited exciter operating without a regulator is determined by the intersection of the field-resistance line and saturation curve. At other than normal frequencies the shape of the saturation curve changes, and at one particular frequency the field-resistance line will coincide with the saturation curve, causing a complete collapse of the exciter voltage. A regulator would, in effect, shift the base of the field-resistance line to the right, stabilizing the exciter voltage. In any case, low voltage and low frequency both act to reduce the field voltage and the var output of the generator. As the exciter voltage drops, the effective internal voltage of the machine also drops and might eventually cause the machine to pull out of synchronism with the system. It is essential, therefore, to return frequency and voltage to normal as soon as possible to prevent a cascade tripping of units on loss of field or loss of synchronism.

Low voltage and low frequency not only affect boiler feed pumps and coal mills,[2] but also can cause a reduction in generator output and a reduction in the amount of spinning reserve. In short, a system operating at low frequency is in imminent danger of system collapse.

Coordinated disaster control measures

No matter how well a system is designed, the possibility of separation from the interconnection and the additional probability of system shutdown are always present. A careful analysis of the behavior and interaction of various system elements described suggests procedures to limit the extent and duration of such outages. Techniques include

1. Load shedding
2. Distribution of spinning reserve
3. System separation

These measures can be coordinated with respect to time and frequency in much the same way that protective relays are coordinated to provide backup for any type of system protection failure.

Figure 3 is a diagram coordinating disaster control procedures for a typical system that is part of a large interconnection. It is presented here to demonstrate the method rather than to provide specific rules for coordination. This diagram shows the complete range of possible states that might exist for a power system that has become separated from the interconnection. The range includes the time following a disturbance from zero to infinity and frequency from 60 Hz to zero, with appropriate action indicated in the proper time–frequency areas.

Three types of load shedding are included in this diagram: load reductions by underfrequency control of voltage regulators on substation transformers and distribution feeders, automatic load shedding by underfrequency relays, and manual load shedding.

Load reduction by voltage control has the advantage that it does not disconnect customers, but the amount of load reduction possible by this method is limited and the response is slow. There may be conditions, however, where the excess load is such that the initial decline of frequency will be arrested by voltage reduction before the frequency reaches the critical point at which load must be shed. In this respect, voltage reduction could prevent load shedding.

Automatic load shedding on underfrequency relays, on the other hand, is very fast and can be applied in any desired amount.

Manual load shedding has a place in the outer fringes of Fig. 3 and is applicable when conditions warrant any action necessary to maintain generators in service. Both manual load shedding and load reduction by voltage control may take place at normal frequency in the absence of a disturbance if the system load reaches a value in excess of the system generating capability minus spinning-reserve requirements, or if such load shedding is necessary to prevent overloading of lines, which might cause system separation.

The most critical value on the coordination diagram of Fig. 3 is the frequency at which automatic load shedding begins. This represents a line of demarcation between a system operating under design conditions and one entering the realm of emergency operation, in which disaster control measures are required. If the value selected is too high, load may be shed unnecessarily, but a low value may delay or prevent the recovery of this system. Ideally, this value should be selected just below the minimum transient frequency from which the system is certain to recover. Since any part of the system might be separated and the exact area included in the island is not predictable, the operating policy concerning the distribution and amount of spinning reserve is important. For example, if every machine is required to carry at least 5 percent spinning reserve, and if this reserve is sustained capability available within one minute, any area with 5 percent excess load could be expected to recover without load shedding. The minimum transient frequency that will result from a loss of 5 percent generation depends upon the inertia of the system and the response of the governors. The initial rate of change of frequency is easy to compute according to Eq. (3), but the minimum value of frequency that might be reached should be determined by testing a model of the system for its transient response to the loss of generation. An example of such a curve is labeled "Response curve 1" in Fig. 3.

If spinning reserve is not evenly distributed among all the units, or if the required time to achieve sustained output is excessive, the minimum frequency at which load shedding starts should be higher. If the system is part of the Canada–United States interconnection, the value might be about 59.5 Hz, since this would provide a reliable indication that the area has separated from the interconnection.

In any case, coordination is required between spinning-reserve policy and the selection of the threshold value of frequency for load shedding.

Once load shedding has begun, loads should be shed in increasing amounts as the frequency continues to drop until up to 50 percent of the load has been dropped. The system should be expected to recover for an excess of load roughly equal to the amount of load shedding provided. If the excess load exceeds that provided for, frequency will continue to drop and further action is needed as the next line of defense.

The selection of the frequency that is to trigger the next line of defense is a problem similar to the selection of the load-shedding threshold frequency. The minimum transient frequency from which the system can be expected to recover with the aid of automatic load shedding is a function of the amount of load to be shed, the characteristics and settings of the underfrequency relays, the response rate of the spinning reserve, and the inertia of the system. This frequency may also be determined by testing a detailed system model for transient response to loss of generation. Such a curve is labeled "Response curve 2" in Fig. 3.

In the region beyond load shedding, system separation provides the next line of defense against complete system shutdown. Assuming that there is more than one load-control area in the isolated system, load-control areas provide the most practical subdivision because the metered ties define a specific island to a dispatcher which can be monitored for area interchange and can be isolated by opening ties at metering points. As indicated in Fig. 3, this drastic step is indicated if the frequency drops below the frequency that indicates an excess load too great to be handled by load shedding.

If the frequency remains too long in the critical range from which the system is expected to recover by automatic load shedding, area separation is again indicated. In this case, there may be time to locate that particular area in trouble and isolate it before it becomes necessary to isolate all load-control areas. As shown in Fig. 3, the time allowed for this action is a function of frequency.

In the general case, the dispatcher in the area in trouble will recognize the condition by the heavy flow into his area while other areas will be experiencing nearly normal net interchanges. He may have time to shed load manually before his ties to other areas are opened.

In any event, when the frequency has been in the critical zone long enough that the system is in danger of wholesale loss of generation due to excitation failure or generator overloads, the operator in each load-control area should open his ties and perform manual load shedding or generator separation as necessary to maintain generation in service without damage to equipment.

The load-control areas and the isolated system should be returned to normal frequency and resynchronized with the interconnection and the load picked up as generation becomes available.

Conclusions

A method for coordinating load shedding, spinning-reserve rules, and area separation has been presented. Time–frequency diagrams can be employed to coordinate the various disaster control procedures in much the same way relay coordination limits customer outage to the minimum necessary. The characteristics of generator exciters and other equipment suggest appropriate time limitations.

The choice of disaster control procedures need not be restricted to those discussed here to make this approach valid.

Full text of a paper presented at the IEEE Summer Power Meeting, Portland, Oreg., July 9–14, 1967. Recommended for publication by the Power System Engineering Committee of the IEEE Power Group.

REFERENCES

1. "Northeast power failure—November 9, 10, 1965," House of Representatives Committee on Interstate and Foreign Commerce Report (Part 2), Serial No. 89–54, U.S. Government Printing Office.

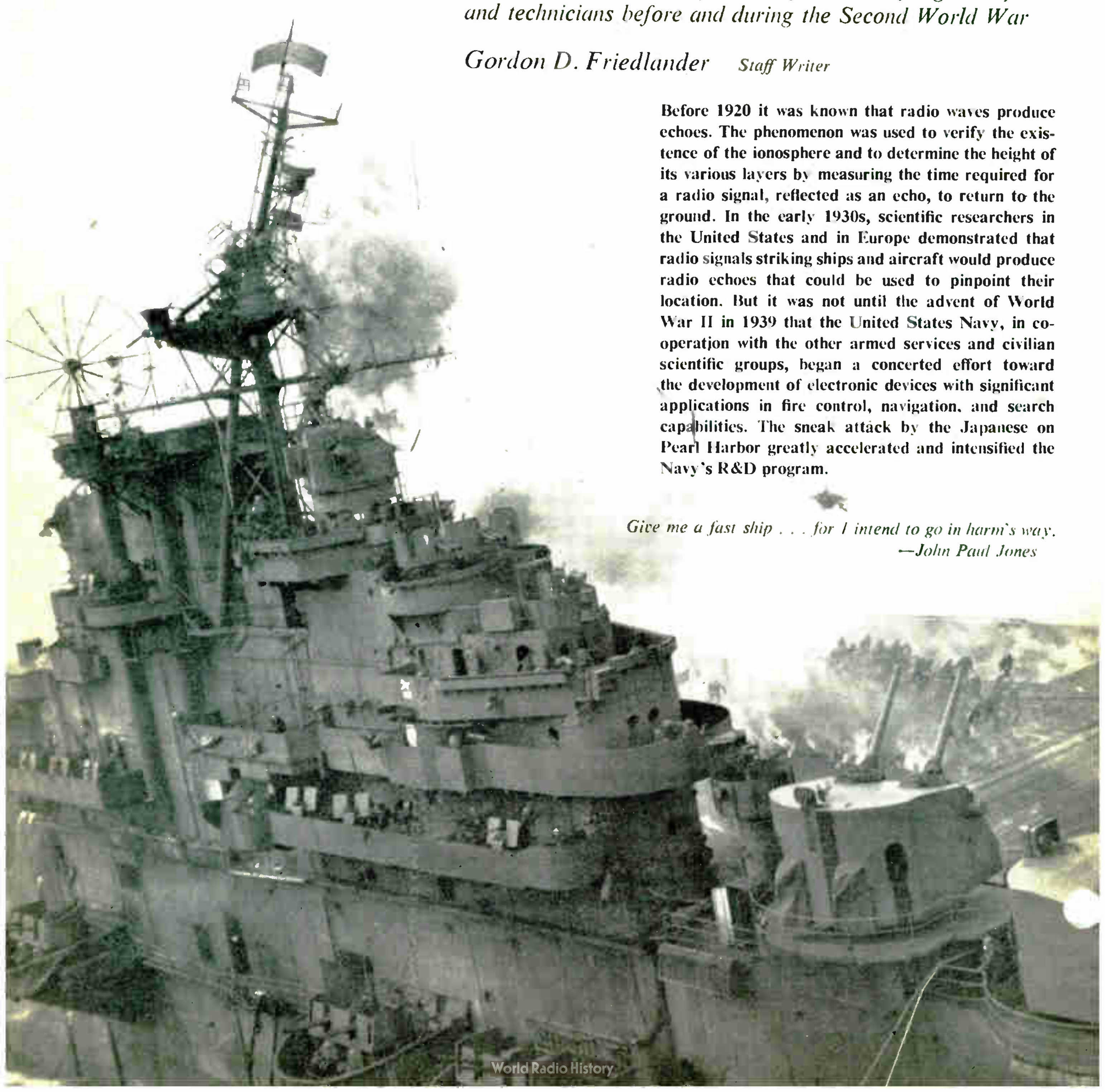
2. "Effect of low frequency and low voltage on thermal plant capability and load relief during power system emergencies," *Trans. AIEE*, vol. 73, pt. III-B, pp. 1628–1688, 1954.

World War II: Electronics and the U.S. Navy

Radar, sonar, loran, and infrared techniques

In the inexorable paths of fate and destiny, it was fortuitous that both the Germans and the Japanese made far less than optimum use of their top scientists, engineers, and technicians before and during the Second World War

Gordon D. Friedlander *Staff Writer*



Before 1920 it was known that radio waves produce echoes. The phenomenon was used to verify the existence of the ionosphere and to determine the height of its various layers by measuring the time required for a radio signal, reflected as an echo, to return to the ground. In the early 1930s, scientific researchers in the United States and in Europe demonstrated that radio signals striking ships and aircraft would produce radio echoes that could be used to pinpoint their location. But it was not until the advent of World War II in 1939 that the United States Navy, in cooperation with the other armed services and civilian scientific groups, began a concerted effort toward the development of electronic devices with significant applications in fire control, navigation, and search capabilities. The sneak attack by the Japanese on Pearl Harbor greatly accelerated and intensified the Navy's R&D program.

Give me a fast ship . . . for I intend to go in harm's way.
—John Paul Jones

An historical prologue

In the 1930s, the United States Navy, working within a very restricted naval budget (and agains twidespread public apathy), valiantly attempted to modernize its fleet of capital ships—the mighty "battlewagons" (BBs) that were the nucleus of all preatomic-age navies. The U.S. Navy's battleships of this period, like those of Great Britain's Royal Navy, were originally constructed just before or after World War I. And despite extensive modernization and rebuilding that included the installation of dual-purpose (capable of high-angle antiaircraft fire) secondary batteries and electronic gear, they were still relatively slow and cumbersome vessels. A bright spot in the overall naval picture of the United States in the twenties and thirties was the construction of several first-rate heavy and light cruisers and the famous "gold plater" (*Farragut* class) destroyers, built under the limitations of the London Naval Conference.

In 1939, spurred by the danger of United States' involvement in the European war, Congress released the purse strings and authorized sufficient funds for a major program of warship construction in all categories. With the commissioning of *U.S.S. North Carolina* (BB-55) and *U.S.S. Washington* (BB-56) in 1941, the Navy began its wartime acquisition of a powerful and up-to-date battleship force.

Yet, even these fast superdreadnoughts and the subsequent capital ships of the *U.S.S. South Dakota* (BB-57) and *Iowa* (BB-61) classes, rated 35 000 and 45 000 tonnes respectively, were constructed with restrictive limitations. To negotiate the locks of the Panama Canal, no capital ship could be built with a maximum beam (breadth) exceeding 108 feet (33 meters), nor an overall length of more than 900 feet (302 meters). The German *Kriegsmarine*, however, not proscribed by such dimensional parameters, proceeded to build its mammoth, 52 000-tonne *Bismarck* and *Tirpitz* with a maximum beam of about 125 feet (38 meters) to provide a steadier gun platform for the main battery rifles, plus protective armor and inner hull structure that was unusual for that era. Even the new 31 000-tonne battle cruisers *Scharnhorst* and *Gneisenau* were extremely powerful and lethal examples of first-line warships.

The 'ABDA Command'—flotilla without eyes. At the time of the Japanese attack on Pearl Harbor, operational search radar equipment had been installed on seven battleships, six cruisers, six aircraft carriers, and two fleet auxiliary vessels. The daring assault on Pearl Harbor crippled the battleship strength of the U.S. Pacific Fleet with the loss of two capital ships and serious damage to several more. In the ensuing dark days of the winter of 1941–1942, during which time U.S. bases in the Philippines were overrun by the Japanese invasion, the remaining units of the Pacific Fleet—cruisers, aircraft carriers, and destroyers—were scattered from the Aleutians to Australasia. After the fall of Singapore in January 1942, some of these Allied warships—U.S., British, Dutch, and Australian—were forced to fall back on Surabaya, the great Netherlands naval base in Java. Here, one of the most unusual and bizarre, yet tragically heroic, naval episodes of World War II had its inception.

In the climate of a desperate and deteriorating naval situation, the ABDA Command was hastily formed in January 1942, with Admiral Thomas C. Hart, USN, as commander of the Allied naval components. Admiral Hart, however, was not on the immediate scene in Java. The Surabaya naval base and all Dutch naval forces were under the command of Vice Admiral C. E. L. Helfrich, Royal Netherlands Navy, and his deputy, Rear Admiral K. W. F. M. Doorman, RNN, was the commander of the now reduced Dutch squadron of two cruisers and three destroyers. Thus these men were the ranking Allied naval officers. In February 1942, Admiral Helfrich took over the naval command.

The combined striking force included the United States' cruiser *Houston* (already battered in a previous action), the British heavy cruiser *Exeter*, the Dutch cruisers *De Ruyter* and *Java*, and the Australian cruiser *Perth*; the American "flush deck" (World War I vintage) destroyers *John D. Edwards* (Fig. 1), *Paul Jones*, *John D. Ford*, *Alden*, and *Pope*; the British destroyers *Electra*, *Jupiter* and *Encounter*; and the Netherlands destroyers *Evertsen*, *Witte de With*, and *Kortenaer*. Eight other U.S. destroyers and a Dutch cruiser had either previously been sunk in Javanese waters or were rendered *hors de combat* by a series of unfortunate maneuvering and dry-docking accidents. And, ironically, the Australian heavy cruiser *Hobart*, although undamaged by battle or accident, had too little bunker fuel to be committed in the forthcoming great naval action.

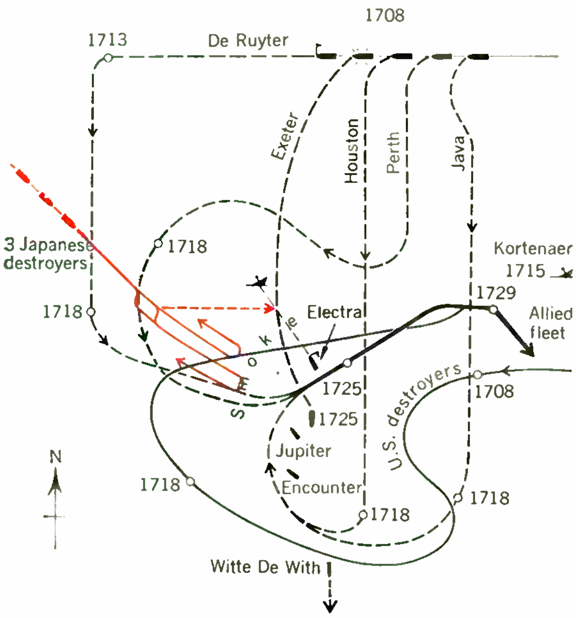
Finally, no ship in the ABDA Command was equipped with any form of radar gear and there was absolutely no air cover either for reconnaissance or offensive action against the approaching Japanese forces.

The Battle of the Java Sea: 27 February 1942. Doorman knew, through sketchy Allied intelligence reports, that a vast Japanese invasion force under Vice Admiral Kondo, consisting of four battleships, five aircraft carriers, eight heavy cruisers, four light cruisers, 39 destroyers, and 56 troop transports and freighters, was bearing down the Karimata and Makassar Straits toward the Java Sea in a three-pronged thrust aimed at eastern, central, and western Java. Let the reader be assured: at this point in time, the basic Japanese strategy of envelopment, and naval tactics, was superb not only in concept but also in execution.

At 2200 hours on 26 February, Admiral Doorman in the flagship *De Ruyter* cleared Surabaya Strait to commit the entire ABDA Striking Force, without air support and adequate intelligence information, directly into the teeth of the Japanese dragon and its vastly superior naval strength.

Both Doorman and Captain A. H. Rooks, USN, commanding officer of *Houston*, reasoned that the only hope

(Left) Direct hit! This remarkable photo was taken at the instant an aerial bomb exploded on the radar mast of the aircraft carrier **U.S.S. Franklin (CV-13)** during the bombing and kamikaze action of 19 March 1945 in the Pacific. Note the basket-type air-search antenna at the left and the fire-control radar antennas on the damaged structure.



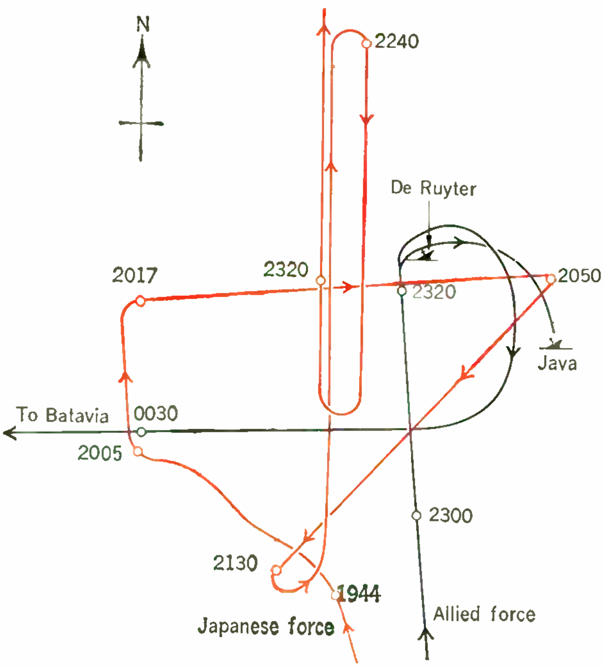


FIGURE 1. The World War I-vintage flush-deck destroyer **U.S.S. John D. Edwards** (DD-216), a veteran of the Battle of the Java Sea, shown at Mare Island Navy Yard on 9 September 1942, after extensive repairs and refitting. Note the "bedspring"-type radar grid antenna on the foremast.

FIGURE 2. First phase of the Battle of the Java Sea, 27 February 1942, showing the maneuvers of the ABDA Striking Force and enemy elements between 1708 and 1729 hours (5:08 P.M. and 5:29 P.M. local time). In this brief action, two Allied destroyers were sunk and **H.M.S. Exeter** (cruiser) was put out of action. Track of Japanese vessels is shown in color.

FIGURE 3. Final phase of the Battle of the Java Sea between 1944 and 2400 hours on 27 February, during which time the Netherlands cruisers **De Ruyter** and **Java** exploded and went down. Track of the Japanese naval force is shown in color.

of saving Java lay in a desperate bid to break through the Japanese screening forces to destroy the troopships and supply vessels. During the night of 26–27 February, the Striking Force steamed north, east, and west in a vain search for the enemy. At dawn, the Allies were discovered by planes from the Japanese carriers; the Battle of the Java Sea, two phases of which are shown in Figs. 2 and 3, was joined.

During that fateful day, one Allied ship after another was either sunk or put out of action by Japanese aerial bombs, naval gunfire, or torpedoes. *Exeter* was crippled and subsequently sank; *Houston* was badly damaged; *Kortenaer* and *Electra* went down. In the words of the noted chronicler of United States Naval history, Rear Admiral Samuel Eliot Morison, USNR (Ret.), commenting on this action: "Admiral Doorman . . . had to play a fatal game of blindman's buff.[1] He had no planes of his own, and intelligence . . . never reached him in time. The convoy [troopships and freighters] . . . was now retiring northward; but Doorman never knew how closely he had approached it . . .

"Actually the Netherlands commander was steering al-

FIGURE 4. Members of the National Defense Research Committee (NDRC) as of May 1945. Back row: Karl T. Compton, president of M.I.T.; Roger Adams, University of Illinois; Conway P. Coe, Commissioner of Patents; and Irvin Stewart, secretary, Office of Scientific Research and Development. Front row: Frank B. Jewett, president of National Academy of Science; Rear Admiral Julius A. Furer, USN, Coordinator of Research and Development, Navy; James B. Conant, president of Harvard University; and Richard C. Tolman, California Institute of Technology.

most directly toward the enemy convoy. If only he had air reconnaissance he might have evaded [Rear Admiral] Takagi and struck the fatal blow. But the dice were loaded against him."

But Doorman did not know the position of the enemy convoy, nor how much Japanese firepower lay between him and his objective. As a final gamble, he ordered a reversal of his flotilla's course to run hard by the Javanese coast in the hope of intercepting the enemy troopships. Here cruel fate once again intervened. Due to faulty communications* and lack of international signal coordination, his order was misinterpreted by some ships under his command and these units failed to execute the change in course. As the decimated ABDA force proceeded along the shoal water, it blundered into a recently laid Dutch mine field. This resulted in the loss of H.M. destroyer *Jupiter*. At 2320 that night, *De Ruyter* and *Java* were trapped in a wide spread of Japanese torpedoes; both ships exploded, and went down. But let Admiral Morison conclude this sad and heartbreaking chapter:

"Brave Doorman, before he lost contact with *Perth* and *Houston*, ordered them not to stand by but to retire to Batavia . . . In one afternoon and evening, half the ships of Admiral Doorman's Striking Force had been destroyed; and the admiral had gone down with his flagship. The Japanese had not lost a single ship, and only one destroyer was badly damaged. The convoy was untouched."

Pursuing Morison's observations one step further, if sonar or radar tracking gear had been installed aboard any of the ABDA ships, the overwhelming odds would have been somewhat reduced. But it would be another four months before the Battle of Midway, in which the radar-equipped naval aircraft of a fast carrier task force of the U.S. Navy could reverse the tides of war.

*Admiral Morison relates that some of the Dutch warships signaled with bright blue searchlights and Aldis lamps, thereby making them easy targets for the enemy. Apparently, infrared techniques were unknown to some Allied navies at that stage of the war.

Enemy miscalculations—and American resilience

When Germany surrendered in May 1945, Prof. Wilhelm Osenberg of the Technische Hochschule at Hanover said: "Germany lost the war because of incomplete mobilization and utilization of scientific brains." Today, the weight of the evidence indicates, despite the feverish acceleration of German military research in the last two years of the war (1943–1945), that Hitler and his generals were initially convinced of Germany's ability to win the war quickly by means of the improved weapons and systems that had been developed between 1933 and 1939. Hence, during the period of the great German "blitzkrieg" offensives and victories (1939–1942), advanced scientific research for military applications was considered to be a waste of time and money!

In Japan, the scientific organization for military research, before and during the war, was even worse than that in Germany. It is estimated that Japan's use of her academic and research scientists was only ten percent of the effective potential.

The United States, despite the disaster at Pearl Harbor and the swift succession of Japanese victories in that empire's occupation of all of Southeast Asia, Indonesia, and the island chains of the South Pacific, was still intact as the world's most productive and powerful industrial nation. Its scientific, technological, and manufacturing resources were so vast that, even after a series of sharp and staggering blows against its overseas armed forces and naval might, it could still come from behind to recoup its losses and seize the initiative from a militaristic tripartite alliance that had overrun and enslaved a large chunk of the free world.

NDRC and OSRD

Immediately after the outbreak of World War II in September 1939, Dr. Frank B. Jewett, president of the National Academy of Science, proposed to the United States Government that a special research group be formed to ensure the full utilization of the nation's scientific potential in the event of our military involvement in the European conflict. On 24 June 1940, the Council on National Defense established the National Defense Research Committee (NDRC), with Dr. Vannevar Bush as its chairman. In its charter, the NDRC was charged with the correlation and support of scientific R&D on all mechanisms and devices for ground, naval surface, and submarine warfare.[2] To accomplish this, the Committee (Fig. 4) would "aid and supplement the experimental and research activities of the War and Navy Departments, and . . . utilize wherever possible existing Government facilities, such as the National Bureau of Standards."

After operating for several months as as adjunct to the Council of National Defense, NDRC was placed, by Executive Order of 28 June 1941, under direct control of the newly created Office of Scientific Research and Development (OSRD). Dr. Bush was named director of this supra-agency. Thus, by the establishment of NDRC and OSRD, the gaps existing in the master plan for the optimum use of American science in World War II were filled.

One of the top-priority problems, which emphasized the critical need for a coordinating link between civilian science and the Navy, was that of effectively combatting the rising toll of Allied shipping from German U-boat action. Fortunately, the overriding urgency of this situation was fully appreciated both by the top naval brass and the first-echelon civilian scientists, and a crash program was initiated for the development of fully operational search and fire-control radar. This effort soon encompassed the evolving ancillary technologies of sonar and loran.

Naval radar and its development*

In June 1900, Nikola Tesla, the great Yugoslavian–American electrical engineer, wrote:

* The concluding installment of this two-part feature will deal with the subjects of acoustical and homing torpedoes, proximity fuzes, and other weapons in which electronics played a major role.

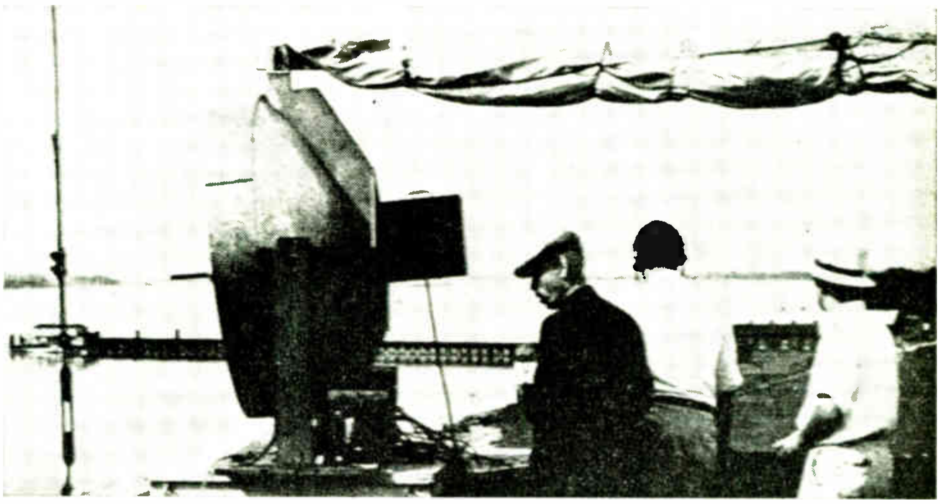


FIGURE 5. Official U.S. Navy photograph, taken in 1937, showing Dr. A. Hoyt Taylor (left) and two associates, Dr. Claude Cleeton and John P. Hagen, operating the equipment which was soon to become the first shipboard radar in the U.S. Navy.

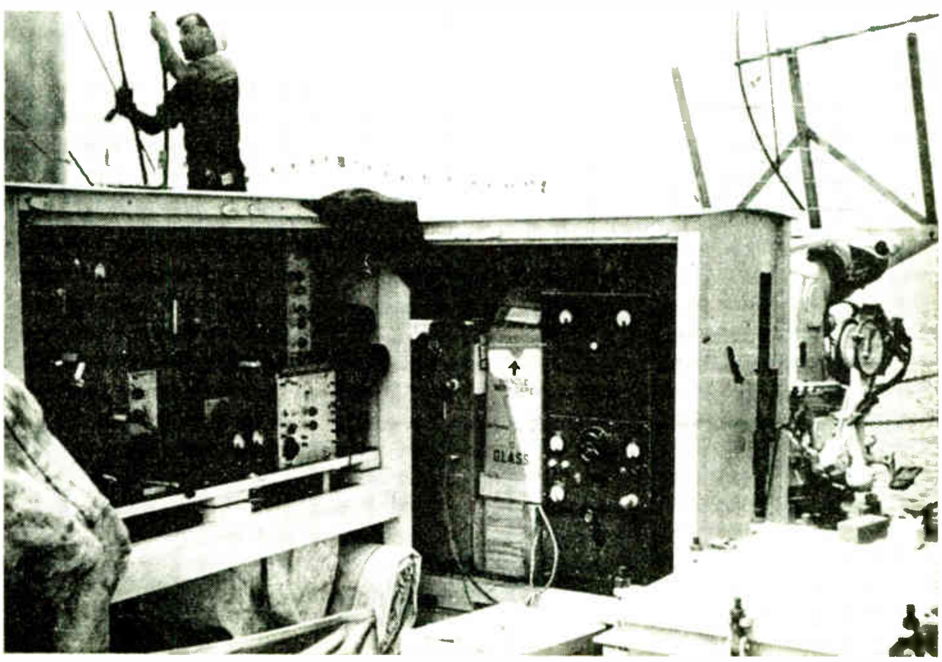


FIGURE 6. The 200-MHz search radar, the first shipboard equipment (1937), being installed atop a 5-inch (127-mm) gun mount on **U.S.S. Leary.**

"When we raise the voice and hear an echo in reply, we know that the sound of the voice must have reached a distant wall or boundary, and must have been reflected from the same. Exactly as the sound, so an electrical wave is reflected, and the same evidence which is afforded by an echo is offered by an electrical phenomenon known as a stationary wave—that is, a wave with fixed nodal and ventral regions.

"Instead of sending sound vibrations toward the remote boundaries of the earth, and instead of the wall the earth has replied. In place of an echo I have obtained a stationary electrical wave—a wave reflected from afar . . ."

That was Tesla—the visionary and the prophet—at his best, predicting one of the most significant electronic systems of the future.

Twenty-two years after Tesla's article appeared in *Century*, Guglielmo Marconi, in a lecture before the Institute of Radio Engineers in June 1922, told his audience that the possible application of reflected radio waves would be of great value in marine navigation. Although the word "radar" was then an unknown acronym in the lexicon of radio, Marconi still expressed himself most articulately when he said:[3]

"It seems to me that it should be possible to design apparatus by means of which a ship could radiate or project a divergent beam of electric rays in any desired direction, which rays, if coming across . . . another steamer or ship, would be reflected back to a receiver screened from the local transmitter on the sending ship, and thereby immediately reveal the presence of and bearing of the other ship in fog or thick weather . . .

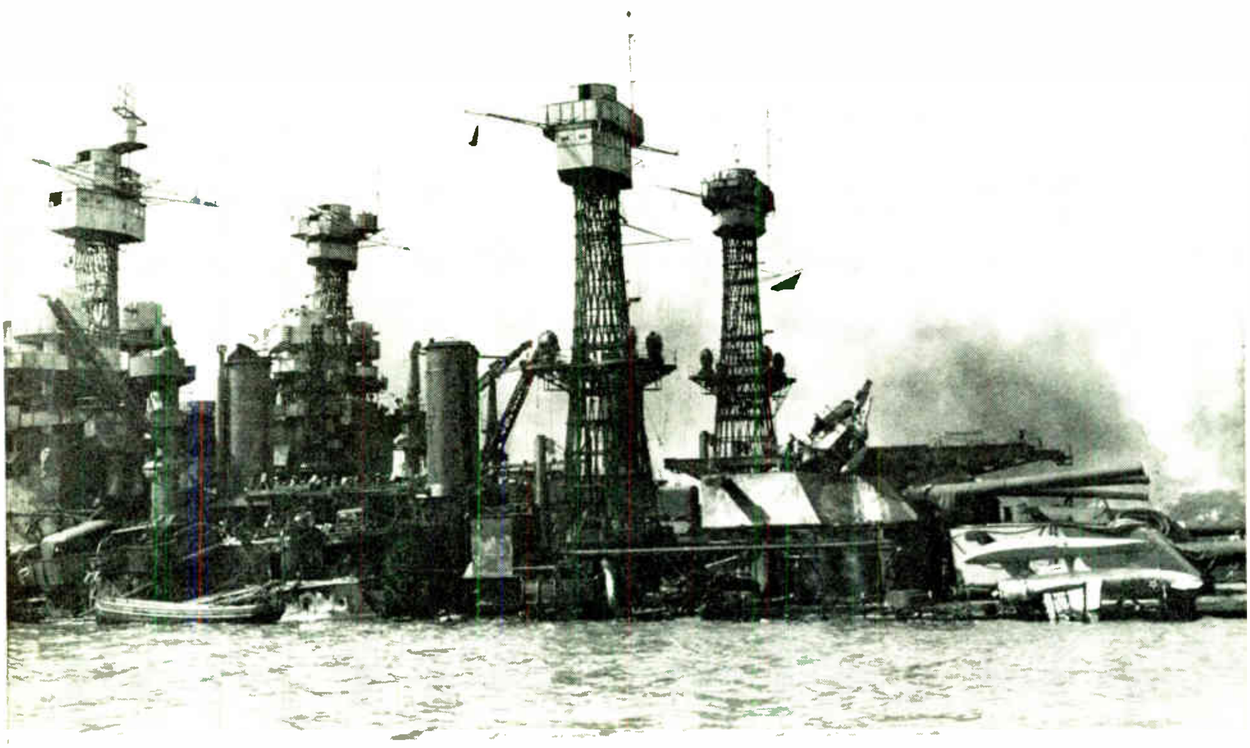
"I have brought these results and ideas to your notice as I feel . . . that the study of short electric waves, although sadly neglected practically all through the history of wireless, is still likely to develop in many unexpected directions, and open up new fields of profitable research."

The 'Research Period.' In what seemed to be coincidental pursuit of Marconi's recommendations, the U.S. Naval Research Laboratory (NRL) initiated in that same year the Research Period (1922–1939), an era of intensive experimentation for determining the optimum radio frequencies for radar operation.[4] At the outset, this empirical research was based only on the certainties that

1. Radio waves travel at the sensibly constant speed of 186 000 mi/s (300 000 000 m/s).
2. The bearing of the target can be ascertained by making use of the directional characteristics of the radar antenna.
3. The same antenna can be used for both transmission and reception.

Actually, Marconi's theory was verified at the Naval Research Laboratory in September 1922 by A. Hoyt Taylor and Leo C. Young, who noted a distortion of radio signals caused by their reflection from vessels steaming by the Laboratory on the Potomac River. Subsequent experiments were often hampered by difficulties in obtaining special vacuum tubes and other electronic components suitable for operation at high frequencies. Also, appropriations of research funds never seemed to be adequate, and

FIGURE 7. Day of Infamy! In the foreground, the heavily damaged battleship **U.S.S. West Virginia** (BB-48) is shown settled on the bottom of Pearl Harbor. Note the early search radar grid antenna atop the vessel's "basket" mainmast at upper left.



NRL personnel were often upset by the tricky and unpredictable behavior of high-frequency radio waves.

In June 1930, further experimentation at the NRL indicated that radio signals reflected from an aircraft could be detected from a ground station. When this information was verified, the Navy's Bureau of Engineering (today's Ship Systems Command) directed the NRL, in January 1931, to "investigate the use of radio to detect the presence of enemy vessels and aircraft."

In the summer of 1932, NRL had built an experimental, continuous-wave radio detection device that could detect an airplane, under optimum conditions, at a range of 50 miles (80.5 km). And, in that year, all NRL information on radio detection was made available to the U.S. Army Signal Corps for parallel and independent further research.

The year 1934 witnessed the Navy's experimentation with pulse-type radio detection. Due to the rapid progress in the radio manufacturing industry, technologically improved and readily available electronic components permitted the development, in 1936, of an 80-MHz pulse-type radar with an aircraft detection range of 45 miles (72.4 km); Fig. 5. And, in 1937, a 200-MHz search radar (Fig. 6) was installed atop a 5-inch (12.7-cm) gun mount aboard the destroyer *U.S.S. Leary* for sea tests and trials. In the following year, improvements worked into the set permitted bearings and ranges to be taken on airplanes 100 miles (161 km) distant.

By 1939, the research stage in radar R & D was drawing to a close as commercial development and production of sets on a mass basis became feasible. Operational tests in that year of the latest NRL radar aboard the battleship *New York* were so encouraging that production models were installed on the battleship *California*, the aircraft carrier *Yorktown*, and the heavy cruisers *Northampton*, *Pensacola*, *Chester*, and *Chicago*.

The 'Introductory Period': 1940–1942. In 1941, before the attack on Pearl Harbor, additional improved radar sets, with an 80-km aircraft detection range, were placed aboard the battleships *Texas*, *Pennsylvania*, *West Virginia* (Fig. 7), *North Carolina*, and *Washington*; the heavy cruiser *Augusta*; the aircraft carriers *Lexington*, *Saratoga*, *Ranger*, *Enterprise*, and *Wasp*; the light cruisers *Albemarle* and *Cincinnati*; and the seaplane tender *Curtis*. These equipments performed so well in subsequent wartime action that "honorable discharges" were given in 1945 to those sets aboard ships that survived enemy action. Unfortunately, however, only some of these ships participated in the critical defensive actions against the Japanese naval juggernaut during the opening three months of the United States' active involvement in the war.

The Introductory Period was marked by a determined effort to improve the performance of radar as dictated by urgent wartime operational needs. Naval personnel became acquainted with radar, discovered its potential—and then demanded more equipment than could be produced!

The year 1942 began with radar as a proved naval weapon. As far as the general public was concerned, the use of radar as a "secret weapon" came to light early in that year when an incident involving radar on the morning of the Pearl Harbor attack was revealed. The Roberts Commission report related the experience of an Army Signal Corps sergeant on that fateful day:

". . . at about 7:02 A.M., a noncommissioned officer discovered [on the oscilloscope] what he thought was a large flight of planes slightly east and north of Oahu, at a distance of about 130 miles [209 km]. He reported this fact at 7:20 A.M. to a Lieutenant of the Army who was at the central information center, having been detailed there to familiarize himself with the operation of the system. This inexperienced Lieutenant, having information that certain United States planes might be in the vicinity at the time, assumed that the planes in question were friendly planes, and took no action with respect to them. The recording of the observation made indicated that these airplanes were tracked toward the island and then lost . . ."

Although specialization in design for different radar applications was pursued in the development of detection equipment for sea-surface as well as air search, and great progress was being made in the evolution of a cathode-ray-tube visual readout,[5] friendly planes could not be distinguished from enemy aircraft until the advent of—

'IFF'

It was now evident that some technique in radar had to be found to distinguish friend from foe. To meet this need, "beeper" equipment was designed to operate in conjunction with a master surface radar, which could "interrogate" transponders (reply equipment) either carried aloft in friendly planes or at sea in Allied surface vessels. The coded reply "blip" would then be matched to the corresponding radarscope targets (Fig. 8) to indicate nonhostile craft. This equipment was called IFF (identification friend or foe).

During this period, the Navy recognized that equipping not only the United States' fleet but also the Allied naval and air forces with IFF was a matter of top priority. Working in close collaboration with NDRC and British

FIGURE 8. A typical World War II operator's console of a more advanced Navy air-search radar set. The oscilloscope readout may be seen at the upper center, and the radarscope (the "yellow-green eye") is at lower center of the photograph.

scientists, the Navy designed prototype models that were quickly put into mass production. By the fall of 1942, the delivery rate of airborne IFF equipment was 15 500 sets per month, and a proportionate quantity of ground and seaborne apparatus was also being built and installed. A fair percentage of the airborne devices was supplied to the U.S. Army Air Force, and both types of installations were furnished to the British.

The Combined Research Group was formed in early 1943 at the NRL to push R & D work on an improved IFF system.

Radar at Coral Sea and Midway: May–June 1942

After overrunning Southeast Asia, the Netherlands East Indies, and the major islands of Oceania in a series of lightning amphibious thrusts in the four months following the United States' declaration of war, the Imperial Navy and naval air force now concentrated on taking its ultimate target—Australia.

At the end of April, Vice Admiral Shigeyoshi Inouye, commander-in-chief of the Fourth Fleet, and Vice Admiral Takeo Takagi, commanding the Carrier Striking Force, moved south from Rabaul with two heavy carriers (*Zuikaku* and *Shokaku*), one light carrier (*Shoho*), six heavy cruisers, one light cruiser, 14 destroyers, 12 troop transports, numerous fleet auxiliary vessels, and seven submarines.

The Allied fleet (Task Force 17), under Rear Admiral F. J. Fletcher, USN, marshalled to intercept the Japanese thrusts against Port Moresby, Tulagi—and eventually the Queensland peninsula of the island continent—included the aircraft carriers *Yorktown* and *Lexington*; six U.S. heavy cruisers, plus the Royal Australian Navy cruisers *Hobart* and *Australia*; 13 destroyers; and two fleet oilers (*Neosho* and *Tippecanoe*).

Carrier battle: 7–8 May. According to Morison,[6] the number of naval aircraft involved was nearly equal: 121 Japanese, 122 American. Although the Japanese had the edge in fighter planes and torpedo bombers, *none of the enemy carriers had either radar or homing devices at this time*. In the series of dogfights high above the carriers of both navies, American radiotelephone communications at times inadvertently jammed the frequencies used by the enemy aircraft, thereby preventing the pilots from getting a homing bearing on their carriers. In fact, some of the Japanese planes, without IFF equipment of course, mistakenly laid a course to land aboard *Yorktown*, using blinker signals for landing instructions! The *Yorktown* obligingly returned the signals and lured them into close-range antiaircraft gunfire.

This Japanese error revealed another shortcoming: the lack of infrared airborne and seaborne homing signal equipment. During the night action of 7 May, Rear Admiral Hara, commanding the two large enemy carriers, was obliged to turn on his searchlights to guide his returning planes; 11 of which splashed at sea.

Lexington, *Yorktown*, and the U.S. cruisers had operational radar gear, including IFF capability. The first electronic detection contact with the enemy came at 1930 hours on 7 May, when *Lexington*'s radar showed Japanese planes in orbit around what appeared to be an aircraft carrier 30 nautical miles* (56 km) distant. Figure 9 shows the tracks of the two big U.S. carriers, and relative positions of friendly cruisers and destroyers during the height of the battle on 8 May.

Quoting Admiral Morison: "The Battle of the Coral Sea [was] the first purely carrier-against-carrier naval battle in which all losses were inflicted by air action and no ship on either side [visually] sighted a surface enemy . . ."

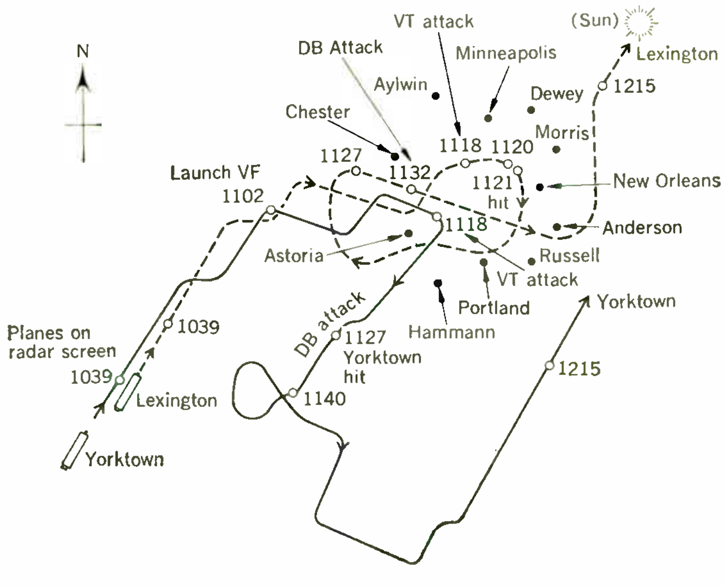
Coral Sea may be summed up as a Japanese tactical victory and a United States strategic victory. Although American battle losses (*Lexington*, *Neosho*, and destroyer *Sims*) were greater than those of the enemy (aircraft carrier *Shoho* sunk; carrier *Shokaku* heavily damaged), the primary Japanese objectives—the capture of Port Moresby and the invasion of Australia—were frustrated.

Finally, radar had played its first vital role as the "all-seeing eye" of tactical naval and air warfare.

'Fourth of June.' The big showdown and the turning point of the Japanese tide of victory occurred at Midway on 4 June 1942. Admiral Yamamoto, thwarted in his effort to push south to Australia, next committed the Imperial Japanese Combined Fleet, consisting of 11 battleships, four major carriers (*Akagi*, *Kaga*, *Hiryu*, and *Soryu*), and a light carrier, plus innumerable cruisers, destroyers, transports, and auxiliaries, in an all-out attempt to seize the Midway Islands. Again opposing the Japanese was Task Force 17, now reinforced by the carriers *Enterprise* and *Hornet* (in addition to *Yorktown*), several more cruisers and destroyers, and three tactical submarine patrol groups of Rear Admiral Raymond A. Spruance's Task Force 16.

In this engagement, however, although badly outnumbered, the American forces had a salient strategic advantage over Yamamoto: shorter supply and communication lines from a home base (Midway) to the battle site. As

FIGURE 9. Opening phase, Battle of the Coral Sea, 8 May 1942, between 1039 and 1215 hours. As indicated on the chart, the aircraft carrier **U.S.S. Lexington's** radar detected the presence of airborne Japanese planes at 1039 that had just taken off from Imperial Navy carriers. Other names on the map pinpoint the positions of U.S. cruisers, carriers, and destroyers during this time interval.



*Note: Miles at sea are invariably expressed as "nautical miles." One nautical mile = 6080 feet, or 1.85 km.

tactical advantages, Midway could accommodate many more heavy bombers than an aircraft carrier—and could not be sunk; and two very effective search radar sets had been installed on the islands. Further, all U.S. aircraft carriers had radar, as did the majority of the cruisers. But most important of all, strategically and tactically, U.S. Naval Intelligence had succeeded in breaking the Imperial Navy's communications code (JN-25), thereby giving the top U.S. commanders advance knowledge of the enemy's intentions and objectives.

Ironically, however, it was a Navy *Catalina* (PBY) flying boat, not radar, that visually spotted the point of the Japanese Occupation Force columns almost 700 miles (1300 km) west of Midway on the morning of 3 June.

Midway was a swirling and complex series of violent aerial dogfights and aircraft versus surface vessel dive-bomber and torpedo-bomber attacks that covered thousands of square kilometers of sea area. By the morning of

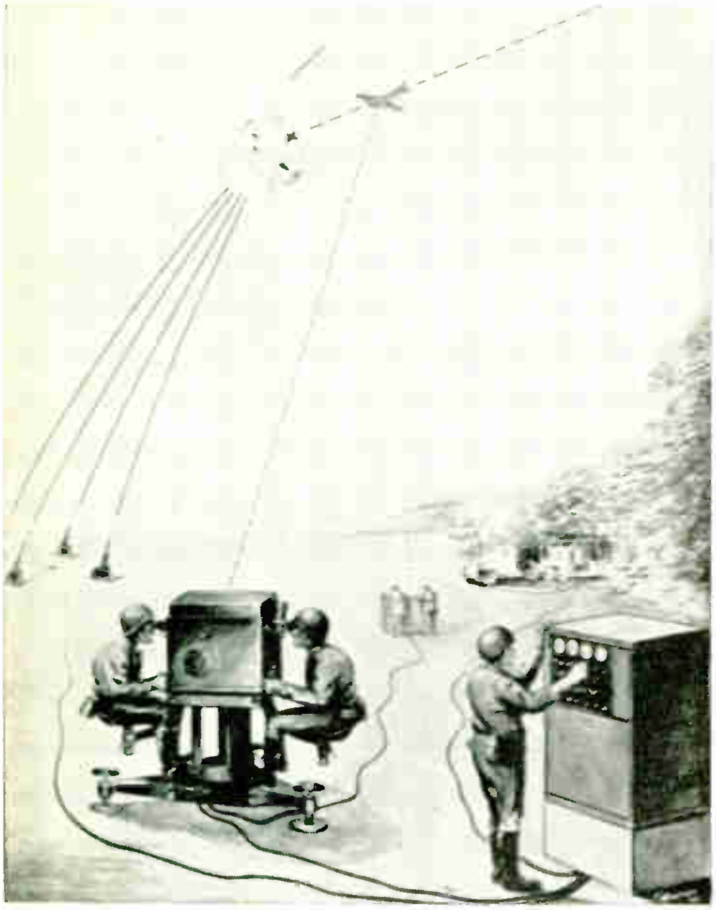
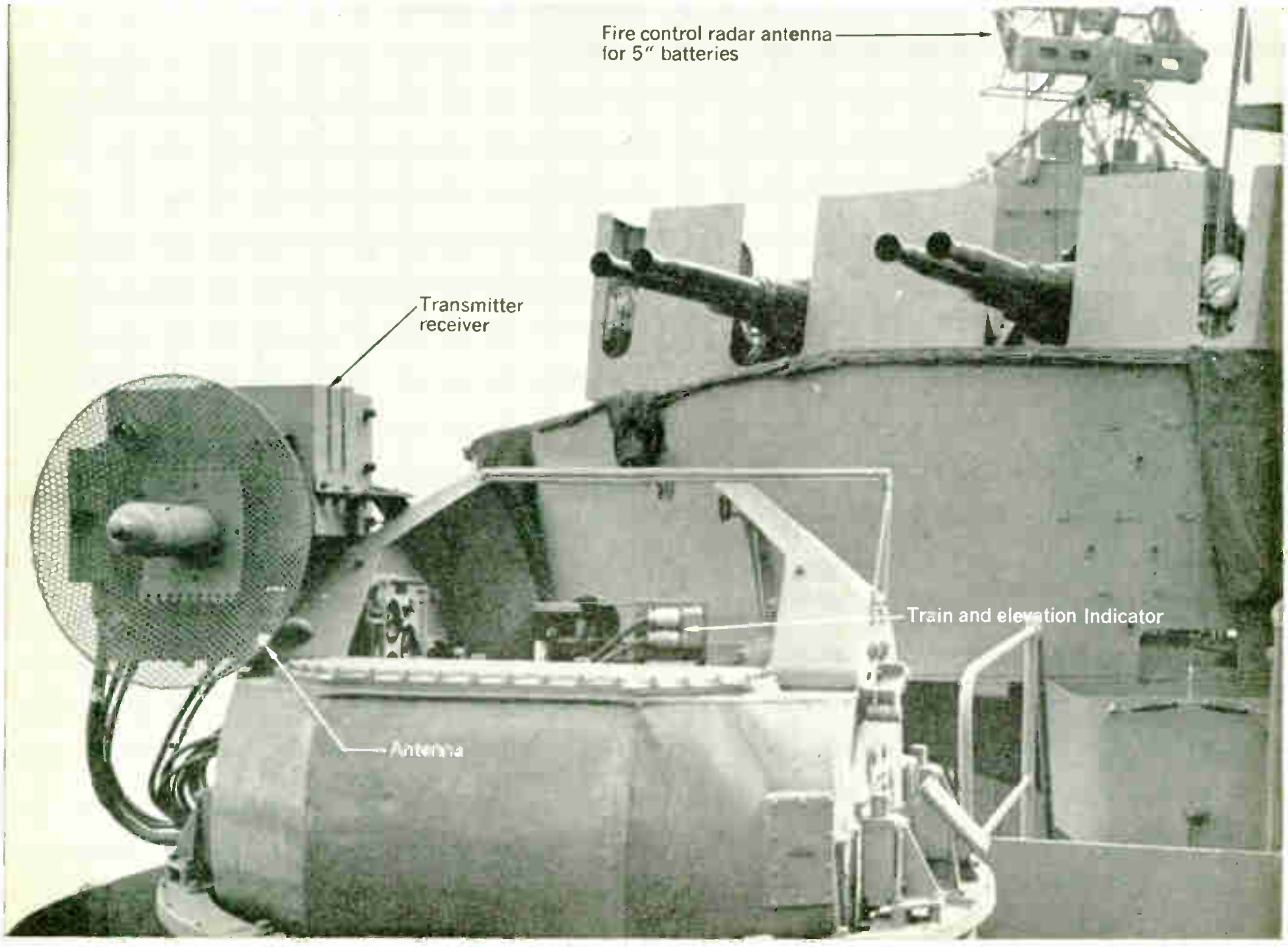


FIGURE 10. Diagram from a military manual showing radar antiaircraft fire-control equipment coupled to a Marine Corps antiaircraft battery in action against hostile bombing planes.

FIGURE 11. Fire-control radar (Mark 49 director and Mark 19 radar) coupled to a quadruple-mount 40-mm Bofors antiaircraft battery aboard battleship **U.S.S. New Jersey** (BB-62).



June 8, the decision was apparent. Japan had lost all four of her heavy carriers, a heavy cruiser, and Rear Admiral Yamaguchi (who went down in *Hiryu*). United States' naval losses were the carrier *Yorktown* and the destroyer *Hammann*. Our aircraft losses to the faster and more maneuverable Japanese carrier planes—and some very accurate enemy antiaircraft fire—were unusually severe. One squadron, designated VT-8 ("Torpedo 8"), of 15 planes was completely wiped out, and two other squadrons suffered more than 50 percent casualties.

Midway was the second great naval encounter in which aircraft carried almost the entire action. Strategically, it forced the Japanese from the offensive role to a defensive posture for the first time in the war. But some of the shortcomings of radar were also revealed. For example, seaborne radar proved to be of little use: high-flying enemy planes from carrier *Hiryu* were only detected by *Yorktown*'s radar at a distance of less than 40 miles (74 km)—less than 8 minutes' flying time. Furthermore, erratic and inaccurate antiaircraft fire from U.S. warships demonstrated the urgent need for electronically directed fire control.

Fire-control and height-finding radar

Yorktown's problem at Midway in directing fighter planes in a short-notice "scramble" to intercept Japanese bombers was eventually solved, late in 1943, when the Navy installed its first height-finding radar on a major carrier. Many subsequent sets of this type were sent to Great Britain.

The problem of more accurate antiaircraft fire control for land-based batteries was partially solved as early as 1941 by scientists and engineers of the Bell Telephone Laboratories and the Western Electric Company. Essentially this apparatus consisted of an on-line optical tracking and measuring instrument that sequentially determined the position of the aircraft in the three spatial coordinates—altitude, direction relative to antiaircraft battery, and bearing of plane. The device then furnished these data to an electrically coupled predictor, or computing machine (Fig. 10), which extrapolated the plane's speed, range, and future position. By this means, accurate and automatic gun "lead distances" ahead of the fast-moving aircraft were established so that the shell would effectively intercept the flying target.

Shortly thereafter, a significant improvement—the famous "Mark 14" sight—a gyroscopic lead-computing device, was developed by Prof. Charles S. Draper of M.I.T., working under a Navy contract with the Sperry Corporation. This Sperry–Draper sight permitted the battleship *South Dakota* to down 32 Japanese planes while defending the carrier *Enterprise* against furious aerial attacks during the Battle of the Santa Cruz Islands on 24 October 1942.

In 1943, the most outstanding achievement in antiaircraft fire control occurred when advanced height-finding radar sets (Fig. 11) were coupled to the computer, and proximity fuzes (a subject that will be discussed in the concluding installment of this article) were used with deadly effect in both land and sea operations.

The advent of the radarscope in 1943 (see "World War II radar: the yellow-green eye," IEEE SPECTRUM, pp. 62–71, May 1966) was the major breakthrough in electronic fire control for surface-to-surface naval actions. With this equipment—far more accurate and reliable than the conventional optical range finders—an enemy ship's distance, bearing, speed, and size could be accurately determined by visual readout, even though the target might be completely invisible, either in night action or under adverse weather conditions.

Action in the North Atlantic

While the war in the Pacific Theater of Operations waxed furious from the Aleutians to Australasia, the submarine menace in the Atlantic became more grimly acute as U-boats took an ever larger toll of merchant shipping, even in the coastal waters off the United States. Both the U.S. Navy and the Royal Navy entered World War II with good echo-ranging gear, but their principal antisubmarine ordnance—the depth charge, or "ashcan"—was virtually the same as that used in World War I, and relatively ineffective as a weapon against submarines of improved modern construction.

As early as 1939, the Navy, realizing that the interests of the United States paralleled those of Great Britain, enlisted the help of some prominent scientists, private companies, and universities for the development of more effective antisubmarine devices. Numerous experiments indicated that sound waves were the only effective technique for propagating signals through water.

Sonar and sonobuoys—ploy and counterploy. The descriptive acronym sonar (*SO*und *N*avigation *A*nd *R*anging) was coined to express more precisely the naval and military applications of underwater sound systems. Actually, the electronic principles used in sonar are analogous to those of radar. Knowing that sound travels about 4½ times faster in water (5000 ft/s or 1525 m/s) than it does in air, fairly accurate tracking and range determination of enemy submarines could be made from surface warships in pursuit. The sonar device consists of a transducer, located near the keel of the ship, which serves as both the transmitter and receiver of the acoustical signal. An oscillator, receiver, and amplifier generate and receive the electrical impulses to and from the transducer. A recorder is calibrated for direct reading in meters, fathoms, or feet. The frequency generally employed is in the low ultrasonic range (20 000–30 000 Hz). The readout is usually an oscilloscope display, coupled to an aural "pinger" that sounds both the transmission pulse and the return echo from the target.

In the war against U-boats, "barrier patrols" were set up to close the waist of the South Atlantic gap between Africa and Brazil. These air and sea search units were equipped either with radar or sonar, or both. Of course, the enemy submarines had sonar receivers that could intercept the frequencies of Allied underwater detection systems. When the U-boat commanders heard the "pings" coming stronger and faster, they estimated the moment of attack and crash-dived their submarines below the sonar beam. To counter this maneuver, the Allies employed a "creeping attack" by two destroyer-escort vessels; the first vessel maintained sonar contact at a fixed distance and ping interval while the second vessel was guided in at dead slow speed by radio or visual signal from the contact ship to a point where depth-charge attack would be most effective.

To foil this technique, the U-boats created water disturbances by expelling chemical "clouds" or bubble streams to reflect strong, false sonar echoes and confuse the pursuer into believing that a second submarine was

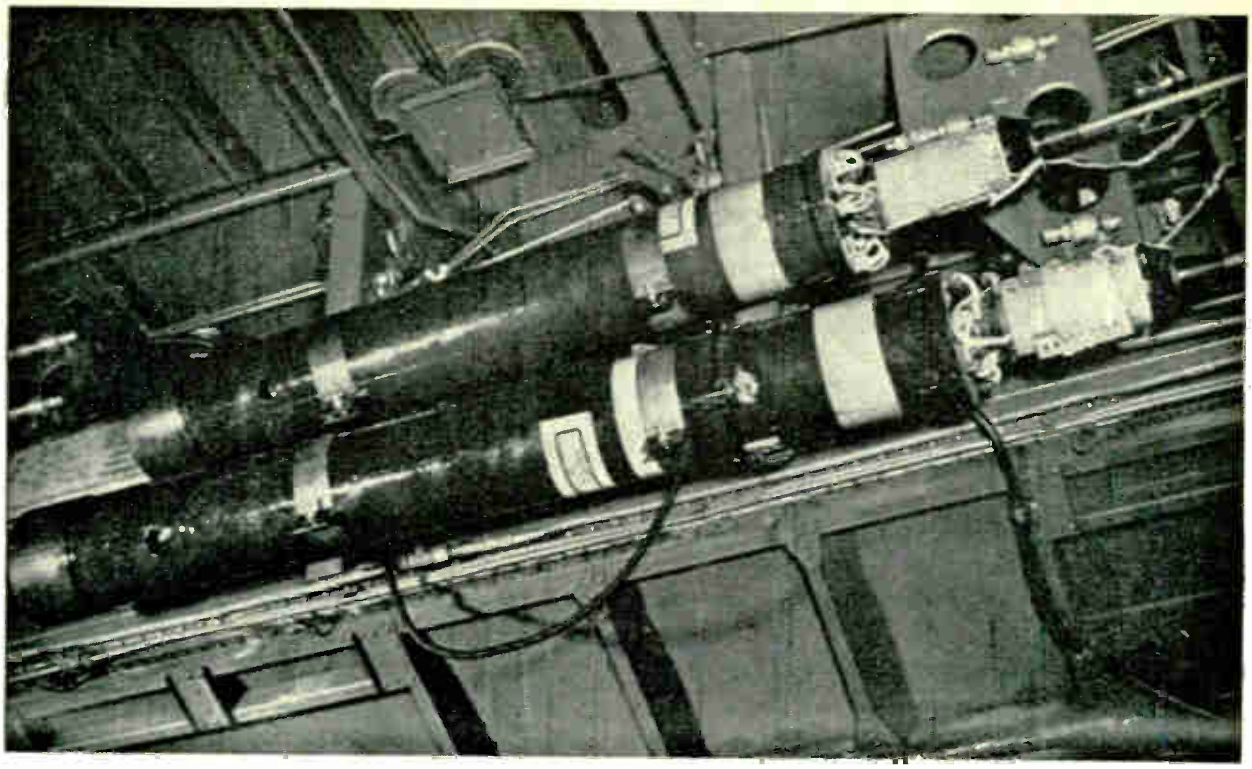


FIGURE 12. Sonobuoys in bomb rack mount aboard a Navy plane. These devices, dropped into the sea by parachutes, contained hydrophones to pick up the noises produced by the propulsion machinery of submerged enemy underseas craft.

FIGURE 13. Air-search and surface-beamed rotating radar antennas aboard the submarine **U.S.S. Clamagore** (SS-343). Many metallurgical, mechanical, and electrical problems had to be overcome before reliability of this apparatus, subject to long periods of immersion, was assured.

present. The Allied sonar and "Asdic" (the Royal Navy version of underwater sound detection) operators, however, soon learned to recognize the "quavering ring" of a false echo and stopped the futile attacks on "simulated" enemy subsurface craft.

The U.S. Navy also effectively utilized "hunter-killer" groups of small escort aircraft carriers (CVEs) in this relentless campaign.

But what the German submarines needed most, particularly when running on the surface at night, was adequate warning of an enemy patrol plane's approach. This was obtained by equipping U-boats with radio search receivers. Radar transmits a short, powerful pulse of radiated energy, and when this strikes a target, only a very small fraction of the signal strength is reflected. This miniscule echo is detected and amplified at the originating radar apparatus by means of a very sensitive receiver. Therefore, a submarine equipped with an intercept search receiver could sense the approach of aircraft long before the submarine could be detected on the pursuing plane's radarscope. Thus the U-boats had ample time to dive and elude their potential attackers.

In wartime, every offensive or defensive measure inevitably produces a countermeasure. The Allied answer to German search receivers was the installation of 10-cm-wavelength sets in patrol aircraft and surface vessels. The German search receivers, designed to pick up transmission at 200 MHz, were useless for the interception of the new microwave frequencies. United States and British patrol planes, equipped with the new radar device, approached "unannounced" to sink many surfaced U-boats. The Germans, thrown into a dither by the rising toll of lost submarines, at first erroneously believed that the Allies had devised a highly efficient infrared detection system and made desperate efforts at countermeasures. It took them quite a while to figure out the Allies' advanced radar technique.

Other U.S. developments included destroyer escorts (DEs) equipped with 10-cm radar, sound gear for precise detection at close-in ranges, and the deadly "hedgehog" mounts whose 32 depth charge projectiles were released in a "ripple fire" sequence to form an impact circle that could readily rupture the pressure hull of a submerged U-boat.

And when enemy submarines in the North Atlantic used their radios to assemble their "wolf packs" to prey upon Allied convoys, the location of these U-boat groups could be plotted by American and British "huff-duff" (high-frequency radio direction-finding) stations. Then the radar-equipped land- and carrier-based bombers searched at the indicated ocean coordinates. If the U-boats were submerged when the planes arrived, the hunters dropped a number of parachute-borne expendable sonobuoys (Fig. 12) whose built-in hydrophones could pick up the noises produced by the submarines' propulsion machinery. These sounds activated the radio transmitter in the sonobuoys to alert nearby bombers or destroyers to move in for a sonar search-and-destroy mission.

A plaintive note—and a renewed menace. By the end of 1943, the Allies' destruction of U-boats became so great that the *Kriegsmarine*'s Grand Admiral Karl Dönitz expressed his chagrin in a letter dated 14 December to Dr. Karl Keupfmeuller, a noted German scientist:

"For some months past, the enemy has rendered the U-boat war ineffective. He has achieved this object, not through superior tactics or strategy, *but through his superiority in the field of science*; this finds its expression in the modern battle weapon—detection. By this means he has torn our sole offensive weapon in the war against the Anglo-Saxons from our hands. It is essential to victory that we make good our scientific disparity and thereby restore to the U-boat its fighting qualities. . ."

> July 30, 1935—With armed sentries guarding the road to the Navesink Lighthouse Station, the new mystery ray developed at the U.S. Signal Corps laboratories at Ft. Monmouth, N.J., was tested for the first time last night, under actual working conditions. In a score of tests the ray, which is said to be able to detect ships more than 50 miles off the coast, even though they are drifting without their motors running, successfully located a ship five miles at sea. The ease with which the cutter was located is believed to have convinced army officers that a valuable adjunct to coastal protection has been developed. It was revealed the ray would next be tested on high-flying aircraft.
>
> —*The New York Times*

But despite Dönitz' gloomy assessment, the *Kriegsmarine* renewed the Battle of the North Atlantic with more U-boats than ever before. These craft were not only equipped with radar to detect Allied planes, but also with more powerful A.A. guns for defensive action, and the "Naxos" search receiver capable of monitoring 10-cm radar transmission.

Perhaps the most remarkable and ingenious device developed by the Germans late in the war was the *snorkel*, a telescoping air intake and exhaust tube that permitted a submarine to run at high speed, at periscope depth, on diesel power—while simultaneously recharging its batteries! The snorkel proved to be a very difficult target for Allied planes to detect by radar. This novel device required the Allies to increase greatly their air and sea patrols, and to install airborne search receivers for the interception of, and "homing in" upon, German U-boat radar.

Radar on United States submarines

The installation of air-search and surface-beamed radar equipment aboard U.S. Navy submarines in 1943 (Fig. 13) greatly enhanced the offensive and defensive capabilities of these boats. Such installations, with antenna grids exposed to frequent and prolonged submergence in salt water, presented many metallurgical, mechanical, and electrical problems that involved corrosion, watertight integrity, and reliability of apparatus. Persistent R & D efforts, however, including several novel modifications, overcame not only these difficulties, but also eliminated—by means of a wide spectrum of diversified microwave frequencies—many of the problems associated with deliberate jamming radiation from enemy transmitters.

Japanese radar, the 'kamikazes,' and the picket ships

At the beginning of World War II Japanese radar was far behind that of the United States, Britain, or Germany. The Japanese radar, which made its debut in the Pacific Theater in 1942, was effectively neutralized by the U.S. Navy's electronic countermeasures. Thus, whenever Japanese aircraft approached our warships, their radar equipment became so completely jammed that the enemy pilots were obliged to use the less reliable and accurate visual methods for launching their bombs and torpedoes. In fact, the U.S. Navy's jamming tactics proved so confusing to the Japanese airmen that many of their attack runs were aborted.

In one surface action at night, an American submarine's search receiver intercepted radar signals from three enemy underseas craft. The U.S. boat stealthily ran these signals to their source and launched a successful torpedo attack against all three. The search receivers were also used by U.S. submarines, with excellent results, in the evasion of Japanese radar-equipped planes.

Nevertheless, Japanese radar, with some assistance from the Germans, steadily improved, and, by 1945, it presented a moderate threat to some of the U.S. naval operations in the Pacific.

As the tide of war turned relentlessly against the Japanese, they initiated their suicidal kamikaze (literally "divine wind") air attacks in 1944 in a desperate attempt to stop U.S. naval task forces and amphibious invasion fleets. By hurtling stripped dive bombers, with an extra-heavy payload of high explosives, directly onto the flight decks of aircraft carriers and into the superstructures of other warships, the enemy was able to inflict considerable damage upon a number of the major naval units (see title illustration, p. 56). Pitted against fanatics who were bent upon their own and their enemies' destruction, radar aboard warships singled out for this type of attack was no longer capable of providing an adequate lead-time warning for effective defensive action.

In an effort to counter this menace, U.S. Navy destroyers were stationed, much like outriders, on a distant radar picket perimeter to give early warning of the approach of the kamikaze planes before they reached the main force of heavy warships. Usually, the kamikazes struck fiercely at the picket ships. In one 80-minute action, the destroyer *Laffey* was attacked by 22 enemy planes. Her crew, by means of radar-coupled A.A. fire control, shot down nine, but eight others crashed the ship in suicide dives. *Laffey* was reduced to a shambles from stem to stern. But despite very heavy personnel casualties, prompt and efficient damage control measures permitted the badly damaged vessel to be towed to Okinawa for temporary repairs.

Electronic navigation—loran

Loran (*LO*ng *RA*nge *N*avigation) is one of several electronic navigation systems developed during World War II. The loran technique was proposed by Dr. A. L. Loomis, a physicist, back in 1940, and the Radiation Laboratory of M.I.T. subsequently undertook the initial R&D work on the system. In the spring of 1942, NDRC summoned Army and Navy representatives to evaluate the feasibility of loran for wartime applications.

Operational principles. Unlike radar, which utilizes the echo principle, loran is a "differential distance" system in which loran receiver-indicators aboard ship measure the time difference between the receipt of signals from two or more fixed-position land-based stations (Fig. 14). Since radio waves travel at a constant known speed, the distance differential can readily be determined to locate an airplane or surface vessel on a hyperbolic line. (The line of a hyperbola is described by the locus of points whose distance differential from the fixed points is constant.) The determination of a second hyperbolic "trace" from another pair of shore stations (Fig. 15) gives a positional "fix" on a ship or plane at the intersection of the two lines.

FIGURE 14. Basic principles of loran, in which differential distances can be determined by radio signals between two shore transmitting stations.

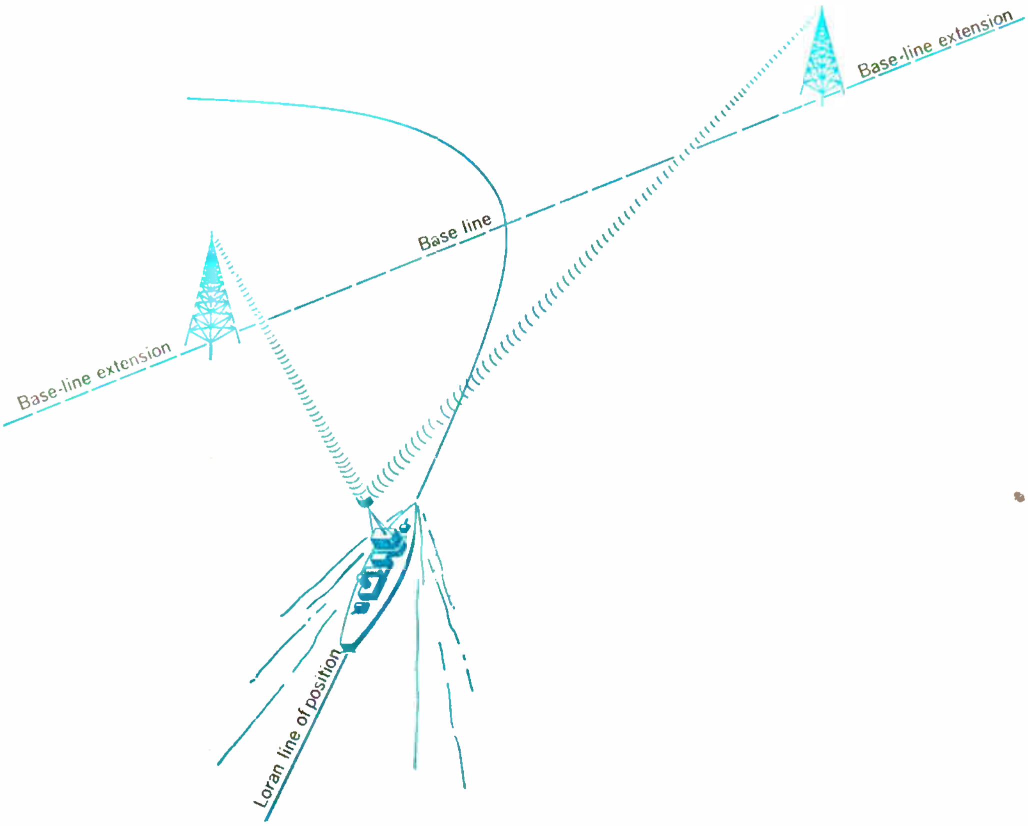
Basically, the loran equipment consists of a radio receiver that can identify the pulse signals, transmitted at regular intervals, from two loran shore stations, and can measure the time difference between receipt of the two transmissions. Today it is no longer necessary to plot the hyperbolic lines on a chart, since this has already been done on prepared charts (Fig. 15) issued by the U.S. Hydrographic Office. Thus, in an established loran system, three or four powerful transmitters, situated in strategic geographic relationship with each other, emit radio signals to produce a navigational "grid" for providing accurate positional information to planes and ships at sea. For example, a three-transmitter grid can cover about one million square miles (3 323 000 km^2) of sea area. During the war, Navy-produced loran equipment was installed on a worldwide basis to assist Allied ships and planes navigating in dangerous or unfamiliar waters.

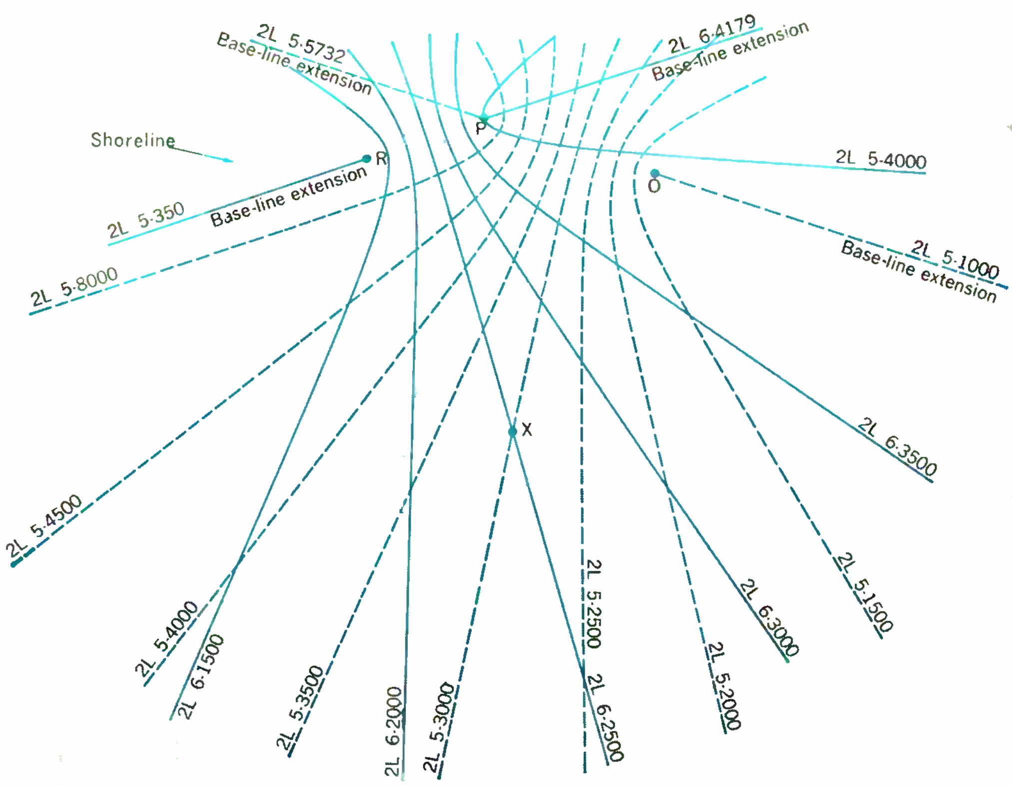
'GCA'

From the start of World War II it was obvious that U.S. Navy planes would have to operate from the icy mists of the arctic wastes to the steaming jungles of the tropics, and that, under frequent adverse weather conditions, a top-priority military requirement would be a flexible and reliable blind-landing system for combat planes.

As early as August 1941, it was logically concluded that if radar was sufficiently accurate to control A.A. gunfire, it could be applied to assist aircraft in landing by determining the plane's horizontal drift in compass degrees, and vertical deviation in feet, from a predetermined, theoretical "perfect glidepath" in space. The heart of the GCA (ground-controlled approach) system was ground-based radar equipment in which a radar operator, plotting the plane's position on the radarscope, "talks down" the pilot by means of precise vertical and horizontal maneuvering instructions that keep the aircraft on the theoretical glidepath until touchdown.

FIGURE 15. Typical loran chart, showing the hyperbolic lines, that is issued for worldwide sea lanes by the U.S. Hydrographic Office.


Base-line extension
Base line
Base-line extension
Loran line of position


2L 5-5732
Base-line extension
2L 6-4179
Base-line extension
Shoreline
P
R
2L 5-4000
O
2L 5-350
Base-line extension
2L 5-8000
2L 5-1000
Base-line extension
X
2L 6-3500
2L 5-4500
2L 5-2500
2L 6-3000
2L 5-1500
2L 5-4000
2L 6-1500
2L 5-3500
2L 6-2000
2L 5-3000
2L 6-2500
2L 5-2000

By contrast with most other electronic landing systems, GCA requires only that the airplane contain a standard radio communication set and a gyrocompass. All of the special GCA equipment—radar installation, communication sets, and control operator's position—was housed in or near the airfield control tower. Two radars were used: a 10-cm search set, and a 3-cm sector-scanning system. The search set was employed for area traffic control within a 30-mile (48-km) radius of the airport.

With a PPI (plan position indicator) radarscope, the azimuth (bearing) and range could be determined for any aircraft within the 48-km radius. The main function of the search system was to detect and identify individual planes coming into the landing zone, and by voice instruction to guide the pilot into an approximate alignment with the final descent path at a distance of 4–10 miles (6.5–16 km) from the airfield. In this way, the search radar led the plane into the area in which the 3-cm scanner beams operated and where the pilot received more precise positional information as to range, azimuth, and altitude from this equipment.

In March 1942, after three months of preliminary tests by the Radiation Laboratory Group, the GCA equipment was removed to the Naval Air Station at Quonset, R.I. Here, the first GCA instrument approach was made by using a Navy plane with its windshield "blacked-out." A short time later, two Navy *Catalina* flying boats made a successful GCA landing during a blinding snowstorm.

The production of GCA equipment in quantity was authorized in May 1943, and a Navy-operated school for the training of officers and enlisted personnel in GCA operation and equipment maintenance was established shortly thereafter. By August 1945, more than 25 GCA-equipped airfields were in use from North Africa to the South Pacific.

Infrared techniques

The U.S. Navy's interest in infrared techniques for signaling applications dates back to 1918, when the United States was a participant in World War I. At that time, German interception of U.S. wireless and visual "blinker" signals was serious enough to merit research into the use of invisible radiation systems. A mechanically modulated carbide lamp, equipped with an infrared filter and a selenium-type photocell receiver, was mounted aboard the battleship *Pennsylvania*. Although Morse code messages were transmitted over short distances at sea, the efficiency and reliability of the cumbersome apparatus left much to be desired.

Nothing further was done in this area until 1930, when the NRL developed a short-range ultraviolet signaling system that could not be visually detected. Meanwhile, The Radio Corporation of America (RCA) was working on an infrared picture tube in connection with its early experiments in television. This tube—vintage 1933—was the precursor of later types developed for the Navy. In 1935, a telescope with an infrared image tube for signalling and reconnaissance purposes was demonstrated, and, in 1939, RCA delivered to the Navy a complete infrared signal system based on the image-tube principle.

Research is accelerated. In 1940, with war clouds threatening to engulf the United States, the NDRC urgently pushed R&D work in the infrared field. More than 30 industrial firms and university research groups participated in projects that included the investigation of optical materials and gaseous sources, and the development of infrared storage phosphors, image tubes, bolometers, thermopiles, and photoconductive cells. By propagating infrared energy through the atmosphere, the Navy hoped to design short-range communication, navigation, fire control, and missile-guidance systems.

Research advanced so rapidly during the next two years that when American amphibious forces invaded North Africa in November 1942, operational infrared devices—filters, optics, and phosphors—were used for ship-to-ship and ship-to-shore communications.

An infrared test station was established at the mouth of Delaware Bay by the Navy's Bureau of Ships in 1943. This facility, together with subsequent stations at Patuxent and Solomon Island, Md., Little Creek, Va., and Fort Pierce, Fla., permitted the evaluation of experimental apparatus prior to full-scale production.

Prominent among the infrared devices developed by the Navy during the Second World War were

1. Apparatus for line-of-sight (reliable for distances to the horizon) voice and teletypewriter transmission.
2. Morse code infrared signal devices for use at night in enemy waters, during periods of enforced radio silence.
3. Homing and recognition kits that were provided to underwater demolition teams and frogmen, and beach raiding parties in Pacific Theater operations.
4. A device to give bearing and range information on Allied vessels to assist nighttime ship station-keeping during radio and radar blackouts.

Preview to part II: mark one—fire one!

The final installment, scheduled for publication next month, will explore the history of some of the electronically guided naval weapons—and defensive countermeasures—that were developed by both German and U.S. scientists. These ingenious ordnance devices played a major tactical role in the decisive naval operations of World War II.

Epilogue

The great-circle track through the endless black
Is traced by a radar beam,
While a pencil mark on the loran chart
Bisects the swift Gulf Stream.
—*from "Winter, North Atlantic"*

The author wishes to thank Rear Admiral Ernest M. Eller, USN (Ret.), Director of Naval History, Captain Linwood S. Howeth, USN (Ret.), Lieutenant (j.g.) J. D. Bogart, USN, and the Department of the Navy for their valuable assistance in the preparation of this article. All photographs are reproduced through the courtesy of the U.S. Navy Bureau of Ships and the National Archives.

REFERENCES

1. Morison, S. E., *The Rising Sun in the Pacific, 1931–April 1942.* Boston: Little, Brown, 1948, pp. 342–358.
2. Furer, J. A., "Naval research and development in World War II," *J. Am. Soc. Naval Eng.*, vol. 62, pp. 21–53, Feb. 1950.
3. Dunlap, O. E., Jr., *Radar.* New York: Harper and Row, 1948, pp. 10–11.
4. Harrison, C. W., Jr., and Blower, J. E., "Electronics—your future," *J. Am. Soc. Naval Eng.*, vol. 62, pp. 99–137, Feb. 1950.
5. Baxter, J. P., III, *Scientists Against Time,* Boston: Little, Brown, 1946.
6. Morison, S. E., *Coral Sea, Midway and Submarine Actions.* Boston: Little, Brown, 1954, p. 48.

BIBLIOGRAPHY

Lord, W., *Incredible Victory.* New York: Harper & Row, 1967.

Information theory and cybernetics

A generalization of Shannon's work, the theory of bidirectional communication, is presented for a better understanding of the communication processes in man

H. Marko Technical University of Munich

Following a review of terms used in information theory, conventional communications systems are examined. Functional models that lead to a statistical impulse generator are developed for a neuron and a fish. The problems of information generation, reduction, and storage in man are considered. The concept of bidirectional communication is advanced where the receiver as well as the transmitter is active and the theory of bidirectional communication is shown to be a generalization of Shannon's information theory.

According to Norbert Wiener, cybernetics is the science of regulation and control in organisms and machines. In today's terms, cybernetics can more generally be defined as the science of message transmission, processing, and the regulation and control of complex systems such as automatic devices and organisms. In terms of subject matter, according to this definition, cybernetics is not actually a new science because the areas of research include control and message processing. The important characteristic of cybernetics is that problems are treated by the same methods; cybernetics can therefore be considered as a bridge between sciences. In particular, the methods developed in communications technology are eminently suitable for describing complex systems. While communications deals with the transmission and transformation of signals, information theory deals with their message content—that is, the information transmitted by these signals. This article considers the application of information theory to cybernetics; applications in communications, neurophysiology, behavior physiology, and psychology will be discussed.

In keeping with the special methods of cybernetics, functional models will be used in describing processes. The use of models, which can be realized by electronic circuits and programmed in computers, augments the analytical description of the biology researcher with the synthetic description of the engineer. A synthetic description by a realizable model has the advantages of being quantitative and complete (which does not yet mean exactitude), and compels one to use an economic method of description because of the cost factor. It is true that there is the danger that the model may be an oversimplification of the physical system that it describes. During the establishment of such models, a continuous and close contact is therefore required between the communications engineer and biologist—an example of the bridge between sciences in cybernetics.

Information, transinformation, and channel capacity

Shannon's information theory (1948) permits a quantitative description of the information content of a message. If p_x denotes probability of a certain symbol x at the transmitter, information content I_x is defined by

$$I_x = \mathrm{ld}\,(1/p_x) \text{ bits}$$

where ld stands for the logarithm to base 2.

A highly unpredictable message is characterized by a large information content; a predictable message has small information content. The information content thus indicates the uncertainty of the message from the transmitter that is eliminated by its transmission to the receiver. The average information content of a message, entropy H_x, is obtained by forming an average value

$$H_x = \sum p_x \,\mathrm{ld}\,(1/p_x) \text{ bits} \qquad \text{for all } I_x$$

Entropy H_x has a maximum value when all messages are equally probable. It is reduced in the same proportion as the probability spread becomes more differentiated; for the extreme probability distribution where $p_x = 1$ for one message and zero for all others, $H_x = 0$.

These definitions are sufficient for undisturbed message transmissions. If disturbances are present, however, the definitions must be expanded and the transinformation content defined. Transinformation content $T_{x/y}$ is given by

$$T_{x|y} = \text{ld}\,(p_{x|y}/p_x)\ \text{bits}$$

Term $p_{x|y}$ denotes the conditional probability for the emergence of message x from the transmitter on receiving message y. For example, in an only slightly disturbed channel, $p_{x|y} \approx 1$ for $x = y$ (no error) and $p_{x|y} \ll 1$ for $x \neq y$ (error). In the first case $T_{x|y}$ is positive; in the second, negative. An erroneous or disturbed transmission results in negative transinformation at the receiver. The average transinformation T is given by

$$T = H_x - H_{x|y}\ \text{bits}$$

where H_x is the entropy previously defined and $H_{x|y}$ is the so-called equivocation, that is, the conditional entropy of the probability spread $p_{x|y}$ averaged over all value pairs of x and y. The average transinformation content cannot become negative; it is greatest when the equivocation disappears (undisturbed case). Normally, it is not the transinformation content itself that is significant, but the transinformation flow, measured in bits per second. The maximum value of transinformation flow for a given communication channel is the channel capacity C. Therefore,

$$C = (T/\tau)_{\max}\ \text{b/s}$$

where τ is average transmission time per communication.

Figure 1 shows typical signals and numerical examples for the channel capacity of three commonly used transmission processes. In (A) the channel is band limited; in (B) the channel is both band and amplitude limited; and (C) illustrates a binary channel. In the band-limited case, the expression for channel capacity was given by Shannon[1] as

$$C = B\ \text{ld}\,(1 + S/N)\ \text{b/s}$$

where B is the channel bandwidth and S and N are the signal and noise or interference, respectively, at the receiver output. The ratio S/N is also referred to as the signal-to-noise ratio. If a Gaussian amplitude distribution (white noise) is assumed for the interference, a Gaussian amplitude distribution is found to be optimal for the signal. The quoted figures in the examples of Fig. 1 for television, telephone, and telegraph channels refer to conventional systems; the channel capacity is greatest for television and least for telegraphy.

In the band-limited channel, the entire amplitude and time range are utilized within the limitations of the band. If amplitude limiting is introduced, as in a telegraph channel, only the zero crossovers of the band-limited signal may be transmitted and the channel capacity is reduced. A transmission can be made by varying the distance of the two zero crossovers, which results in arhythmic pulse-duration modulation. The writer computed the channel capacity for this case[2,3] and obtained

$$C = 2B\ \text{ld}\,(\bar{T}_x/\Delta T)\ \text{b/s}$$

Referring to Fig. 1(B), the minimum distance between two zero crossovers is T_0 and the extension of this distance, proportional to the signal value, is T_x. The term ΔT is the range of variation of the zero crossovers owing to interference. An exponential distribution of T_x with respect to its mean value $\bar{T}_x$ yields the condition for maximum channel capacity. Because of amplitude limiting, the channel capacity is reduced from 300 to 200 b/s.

For the binary channel, besides band and amplitude limiting, equidistant scanning in time is performed that

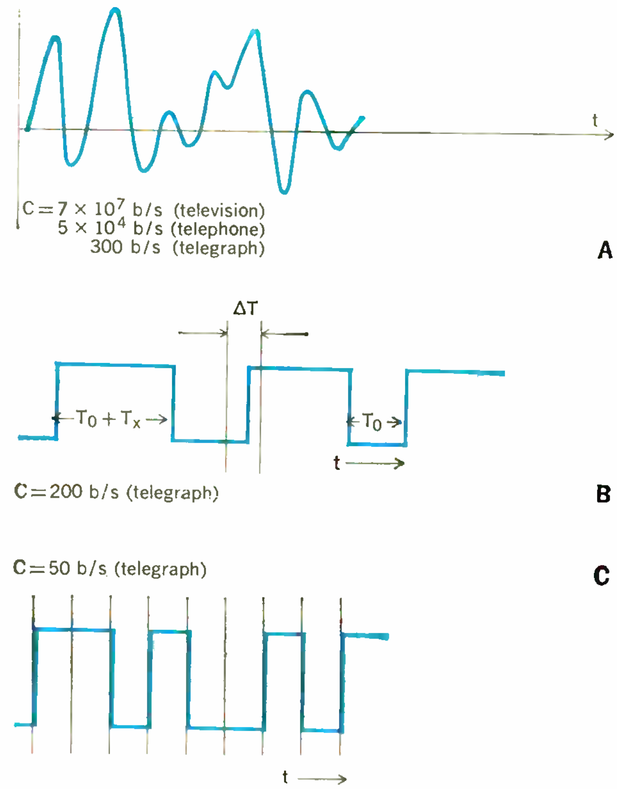


FIGURE 1. Basic communication systems. A—Band limited. B—Band and amplitude limited. C—Binary channel.

results in the telegraph signal of Fig. 1(C). Here, neglecting the probability of errors (a reasonable assumption for practical systems), channel capacity is a function of bandwidth and is given by the expression of Küpfmüller–Nyquist as

$$C = 2B\ \text{b/s}$$

For the telegraph channel, $C = 50$ b/s, which is one quarter of that for the band- and amplitude-limited case.

Almost all communications systems are based on the methods of Fig. 1, although in some systems the channel capacity is even further reduced. Roughly, the first method corresponds to amplitude modulation (AM), the second to frequency or phase modulation (FM or PM), and the third to pulse code modulation (PCM).

In considering the three basic methods of signal transmission, interference was assumed to exist as superimposed noise. In the range of very-high frequencies or low temperatures, however, interference arises from the quantum structure of the communications channel.[4] Figure 2 illustrates a mechanical model of a quantum receiver, where the separate quanta are represented by tiny balls. The position of a slide at the transmitter controls the intensity of quantum flow—that is, the wave amplitude. Because of the statistical nature of quanta arriving at the receiver (Poisson process), a statistical variation of the received signal occurs that makes the assignment of x and y uncertain; variations exist even without the presence of external interference.

The quantum receiver can be pictured as a quantum counter that counts quanta during a time interval, Δt. The statistical coupling between x and y permits the de-

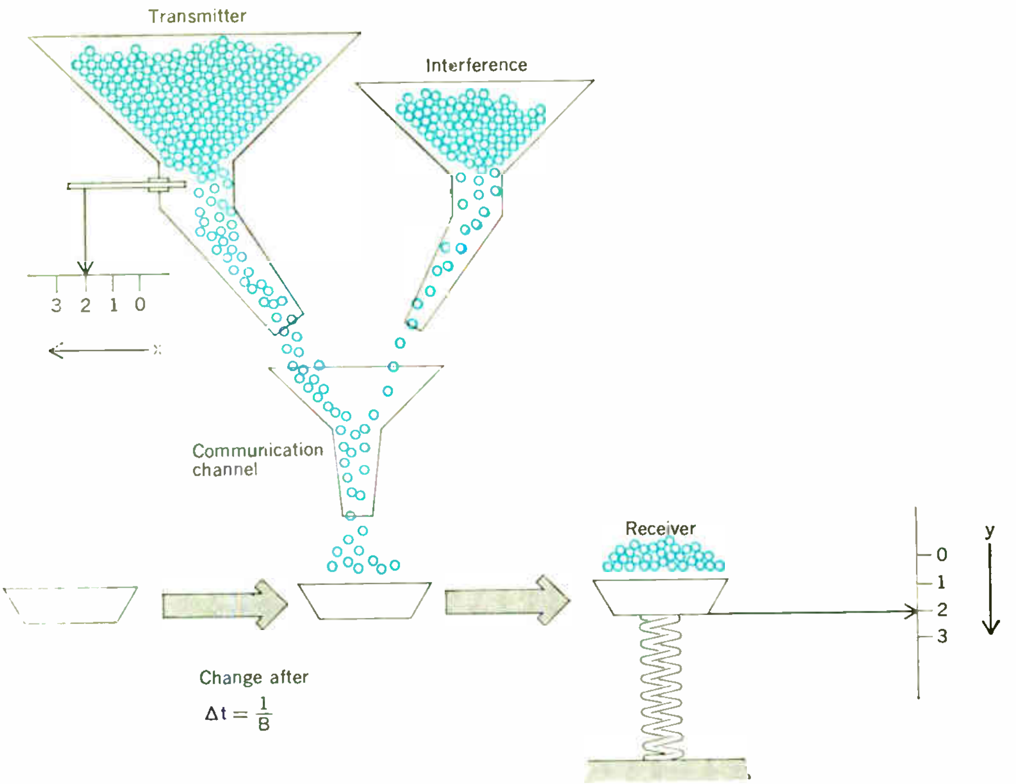
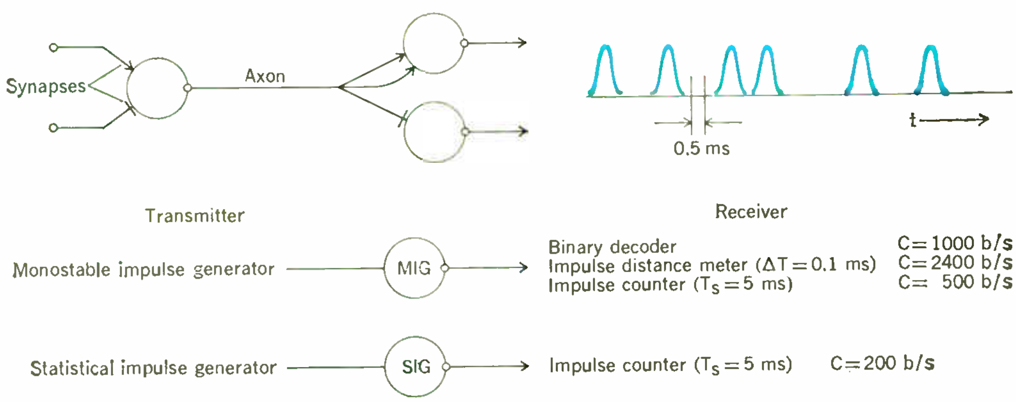


FIGURE 2. Model of a quantum receiver.

FIGURE 3. Channel capacity of a neuron.



termination of channel capacity. The concept of the quantum receiver seems to be of particular interest in the regions of very-high frequencies, low temperatures, and low signal energies—conditions encountered in space communications.

Functional model of a neuron

A sketch of a neuron, consisting of a cell body and axon, is given in Fig. 3. The axon terminates at stimulating (→) or inhibiting (⊣) synapses on the cell bodies of succeeding neurons. It is known that signal transmission in the neuron occurs by pulse frequency modulation, where the impulses have a duration of approximately 0.5 ms. One can picture various mechanisms for the transfer of information among neurons; for example, the neuron can be pictured as a monostable impulse generator (MIG), which produces a maximum pulse frequency of 1 kHz. If a binary decoder is used as a receiver, the resultant channel capacity is 1000 b/s, corresponding to the method of Fig. 1(C). For a receiver using the principle of Fig. 1(B) with impulses having a time spread of $\Delta t = 0.1$ ms, the channel capacity is 2400 b/s. Considering the receiver as a pulse counter that counts the impulses over the synaptic integration time of 5 ms, the channel capacity is 500 b/s.

In all these examples, the transmitter was considered as a monostable impulse generator that produces an impulse as soon as its input voltage exceeds the synaptic threshold value. If one regards the transmitter as a statistical impulse generator (SIG), which, similar to the quantum receiver of Fig. 2, produces a statistical sequence of impulses with a mean rate proportional to the input

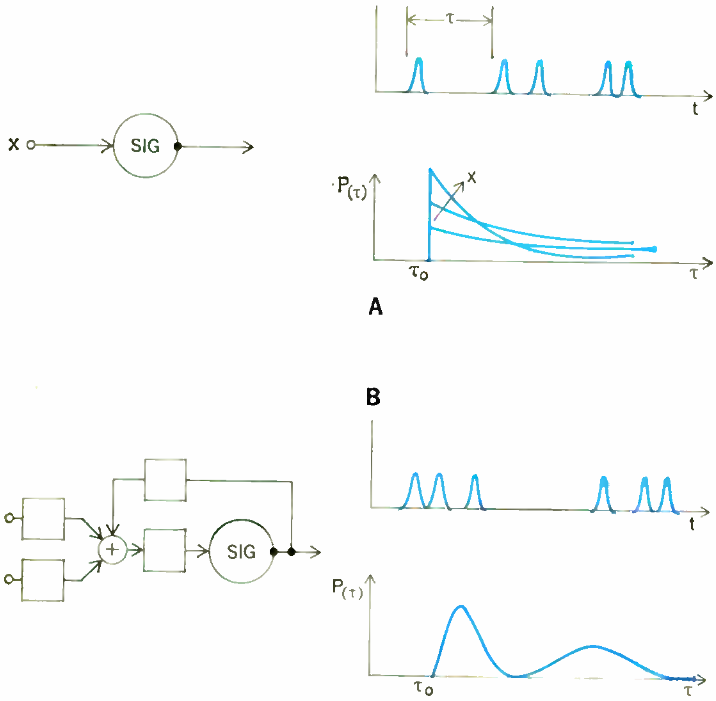
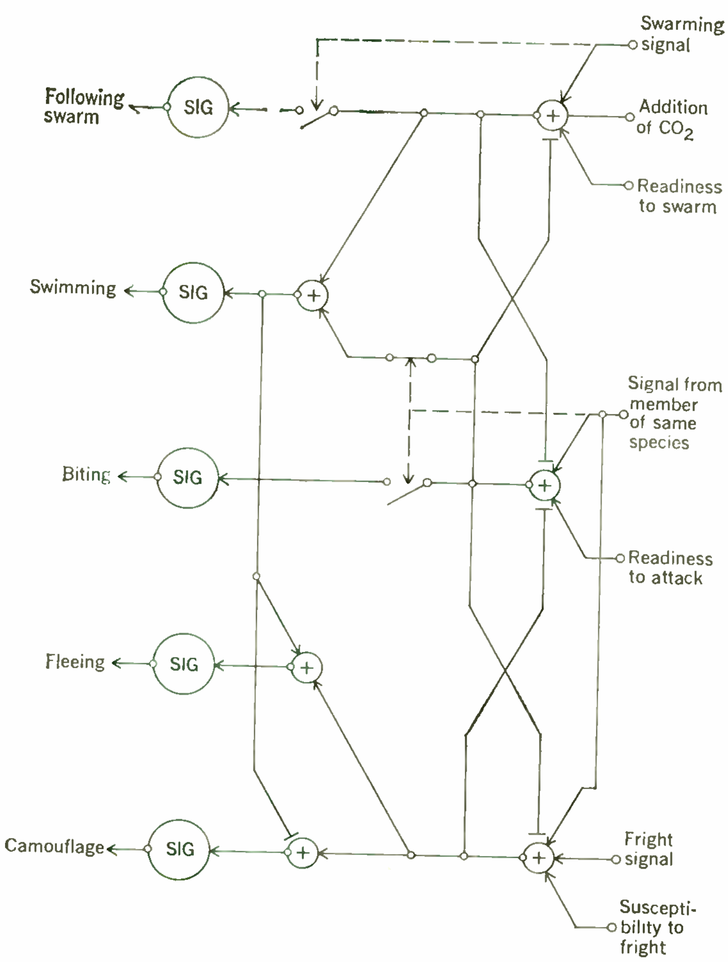


FIGURE 4. Generalized Poisson process. A—Statistical impulse generator. B—Statistical impulse generator with linear transmission elements and feedback.

FIGURE 5. Behavior model of a fish.



signal, the channel capacity is 200 b/s. For this case, the receiver is assumed to be a pulse counter with an integration period of 5 ms.

The sequences registered by neurons in many cases exhibit a statistic with varying impulse intervals, similar to a Poisson process for spontaneous activity. This suggests that information transmission occurs in accordance with the principle of the statistical impulse generator and is governed by a principle similar to the quantum receiver. This demonstrates the analogy in the descriptive methods of engineering and biological systems.

The pulse sequence and interval statistics of a pulse generator that delivers a Poisson process is shown in Fig. 4(A). For such a generator, the interval distribution follows an exponential law, where a dead interval τ_0 takes into consideration that a minimum distance must be maintained. Input signal x varies the mean value of the exponential distribution, reducing the mean distance for a large input signal.

Figure 4(B) illustrates how the statistics of such a Poisson process can be altered by the addition of feedback and linear transmission elements. In this manner, a generalized Poisson process is obtained whose interval distance can, for example, be the bimodal distribution with two maxima shown in the figure. The effect of positive feedback increases the probability for the next pulse immediately after the occurrence of the preceding one. By adding more linear elements at the synaptic inputs, the dynamic behavior (adaptation) can be studied. It is relatively simple to reproduce such statistical impulse generators electronically, realizing any desired given pulse interval distribution.

Functional model for the behavior of fish

The electronic functional model for the behavior of a fish species (Ciclidae) has been described in the literature.[5] Separate instinctive actions, such as following the swarm, swimming, biting, fleeing, and camouflaging, are represented by a stochastic impulse sequence. The frequency of instinctive actions depends on the strength of the impulse signals and so-called receptivities (see Fig. 5). For example, if a signal from a member of the same species is received, the frequency of biting will depend on the readiness for attack. This condition is represented on the model by closing a switch at the input of the corresponding pulse generator. In the absence of a signal from a member of the same species, the readiness for attack expresses itself in stronger swimming (appetence behavior). The frequency of biting is reduced further by increasing the susceptibility to fright or readiness for swarming (antagonistic behavior patterns).

The described functional model depicts the interconnections between separate instinctive actions and readiness (receptivity) relatively well, as verified from statistical studies. Assuming the generalized Poisson process of Fig. 4, the model for the statistical impulse generator resembles the functional model of the neuron; the time scale, however, was stretched because instinctive actions take seconds instead of milliseconds for the neuron.

Functional model for information processing in man

Figure 6 is an example of the application of information theory to a study of the processing of information in man. On the left of the figure, the number of receptors is given for various sensory organs; for example, 2×10^8 rods in the eye's retina and 3×10^4 hair cells in Corti's organ. The number of nerve fibers leading from the receptors is 2×10^6 in the optic nerve and 2×10^4 in the acoustic nerve. Küpfmüller[6] estimated the numerical values for maximum information flow for the eye and ear as 5×10^7 b/s and 4×10^4 b/s, respectively. The figure for the eye was found by estimating the number of resolvable picture dots, the possible shades of gray, and the maximum speed of sequential dots, with a cinematographic limit frequency of 16 Hz. In accordance with information theory, a comparison with the figures quoted in Fig. 1 for television and telephone channels shows that these channels are well adapted to our receptors.

Information flow for various conscious activities of man can also be estimated. Küpfmüller found that counting required 3 b/s; reading, 40 b/s; hearing, 40 b/s; and piano playing, 20 b/s. Because the value of 50 b/s is not exceeded, this value is the maximum channel capacity for the conscious output in man. The value of 50 b/s indicates that there is an apparent reduction of information in the central nervous system. Such reduction is necessary for a good reason: not all impressions of details of the external world need penetrate into our consciousness. A good deal of this information is superfluous or redundant, and therefore can be eliminated; otherwise our memory would be unable to store the entire flow of sensory information. If one multiplies 50 b/s by our average life expectancy of 70 years, the information quantity received is about 10^{10} bits; it is therefore assumed that the storage capacity of our memory is of this order of magnitude.

When considering information processes in man, two questions arise. The first has to do with how the central nervous system reduces information—that is, how it is able to differentiate information that is important for our consciousness from the unimportant. This is a problem in the formation of variables that also exists in sign recognition by machines. The second question is concerned with the generation of information. If one assumes that man is also a source of information, a mechanism for generating information must exist.

Several simple examples of information processes are illustrated in Fig. 7; a system with one input is shown in (A), and (B) applies to a two-input system. Information reduction with one input is possible only in a nonlinear system. Realizable linear systems of finite size cannot suppress completely a frequency band, no matter how small, even with infinite damping. At best, suppression is possible only at a number of discrete frequency points that do not contain any information. Except at these points, and including a delay, the frequency response can be reversed by a complementary network. The picture changes if a noise source follows a filter, making the damped frequency band noisy; but this becomes a system with two inputs.

Figure 7(A) shows basic nonlinear characteristic curves for commutation, threshold, and limiting. The information-theoretical efficiency η is defined as the ratio of information content at the output to input (a symmetrical distribution of input values is assumed). In commutation, one bit (corresponding to polarity) is lost; in threshold, one half is destroyed; limiting produces only one bit. The various efficiencies that are quoted in Fig. 7(B) are based on these conditions.

As an example of a system with two inputs, consider the first four rules of arithmetic with whole numbers from 0 to 9. Assuming the numbers exist with equal probability, an efficiency is defined as the ratio of output information to the sum of information at the two inputs. In adding two numbers, the known output information designated by I_3 will be less than the sum of the known inputs, $I_1 + I_2$. Assuming the total is 6, any combination of $0 + 6, 1 + 5, 2 + 4, 3 + 3, 4 + 2, 5 + 1$, or $6 + 0$ yields the same value for I_3; information has thus been destroyed by addition. Efficiency $\eta = 0.6$ for addition (or subtraction), where the resulting distribution is triangular in shape, so that the totals 0 and 18 have the smallest probability and 9 the greatest. Similar computations apply to multiplication and division with efficiencies of 0.73 and 0.8 respectively.

Küpfmüller and Jenik[7] have shown that simply switched neurons can add, subtract, multiply, and divide—important indications for the mechanism of information reduction. It is certain that the formation of

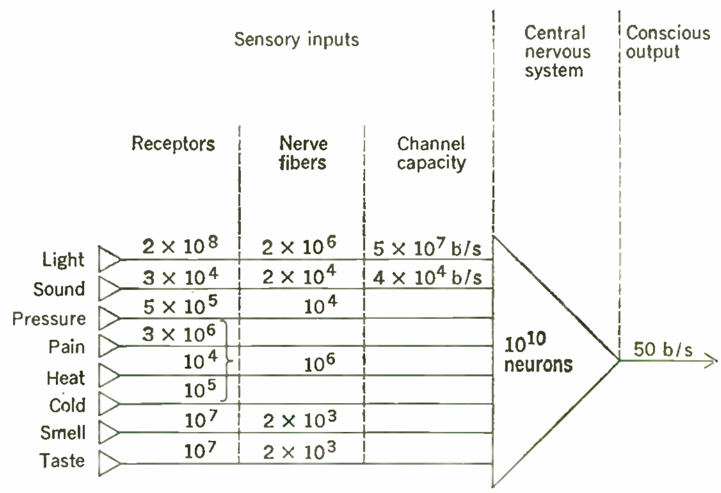
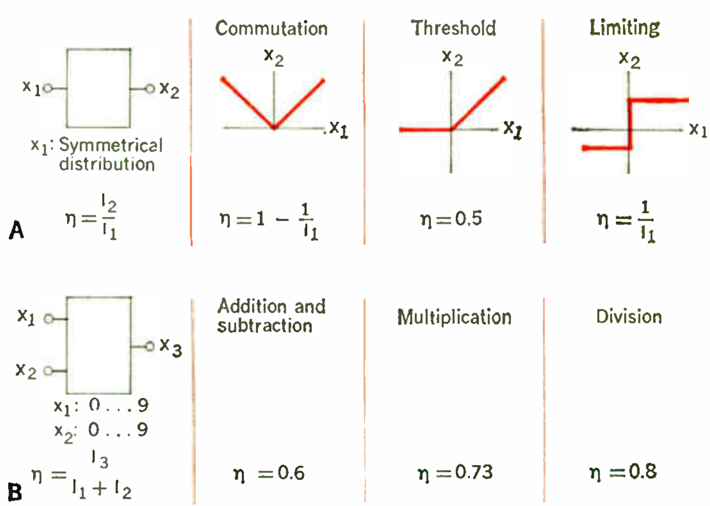


FIGURE 6. Information processing in man.

FIGURE 7. Examples of information reduction. A—Single input. B—Two inputs.



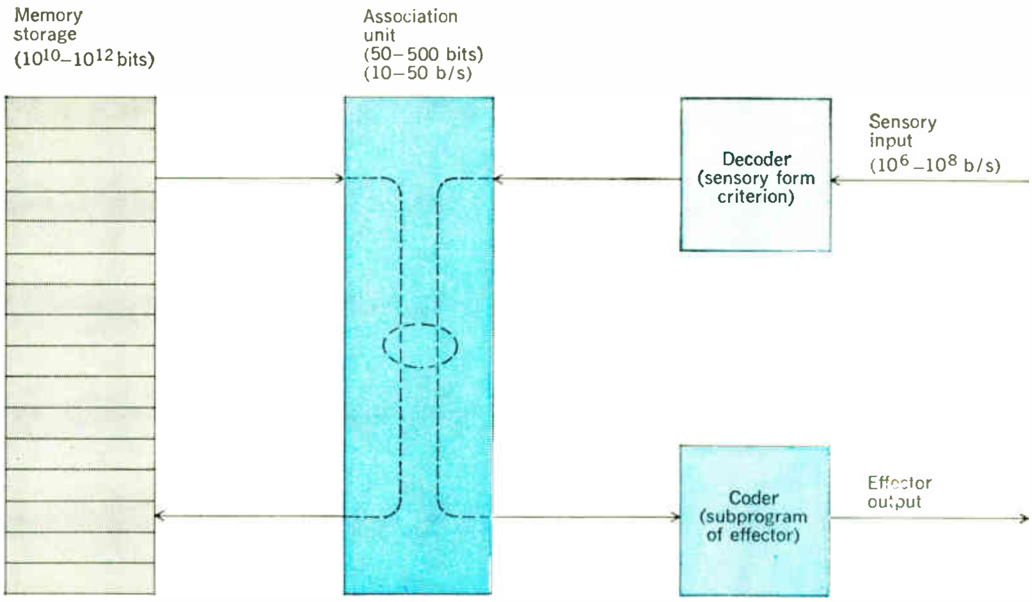
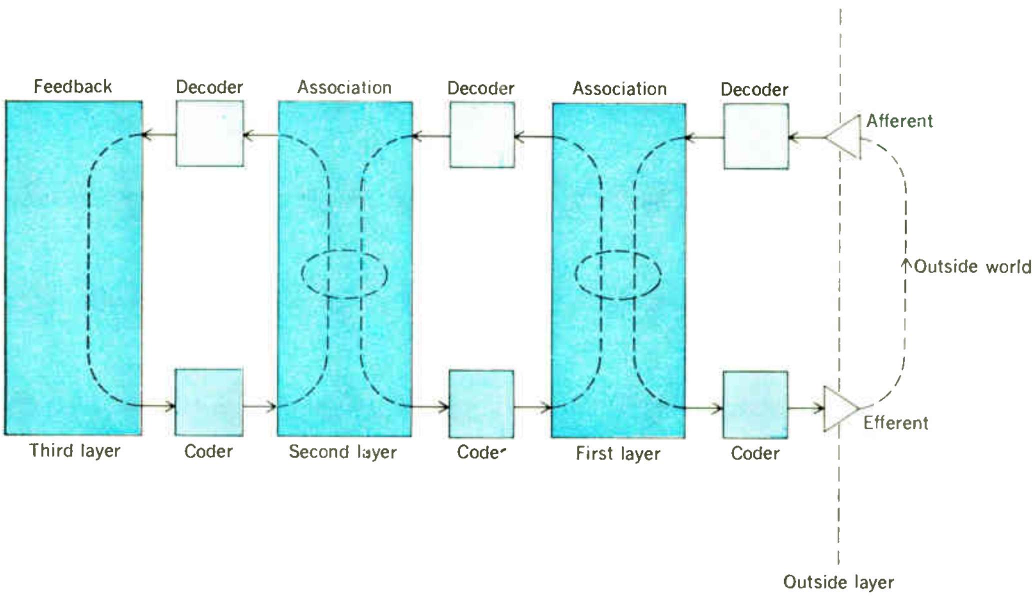


FIGURE 8. Functional diagram for generating information.

FIGURE 9. Layer model for processing information in man.



threshold values and limiting also play a significant role in the nervous system.

Concept of bidirectional communication

The theory of bidirectional communication[8,9] is an amplification of Shannon's information theory with his definitions utilized wherever possible. Although Shannon's theory describes a unilaterally directed information connection (communication channel), the theory discussed here deals with bidirectional communication. The borderline case of a unidirectional transinformation results in Shannon's communication channel. Contrary to Shannon, the receiver of the communication is an active statistical generator and no longer differs from the transmitter.

The theory was developed to describe the creation of information and its processing in man, as well as its transmission between people. Its principal characteristics are:

1. The information transmitter and receiver in man are identical. He has a finite supply of information that is substantially different from that of any other man. This results in a stochastic process that is characteristic for him; entropy is the parameter used for the characterization.

2. Transmission of information between two people is said to be a stochastic synchronization. The entropy is formed of free and dependent parts that constitute the received transinformation.

Figure 8 is a functional schematic for the genera-

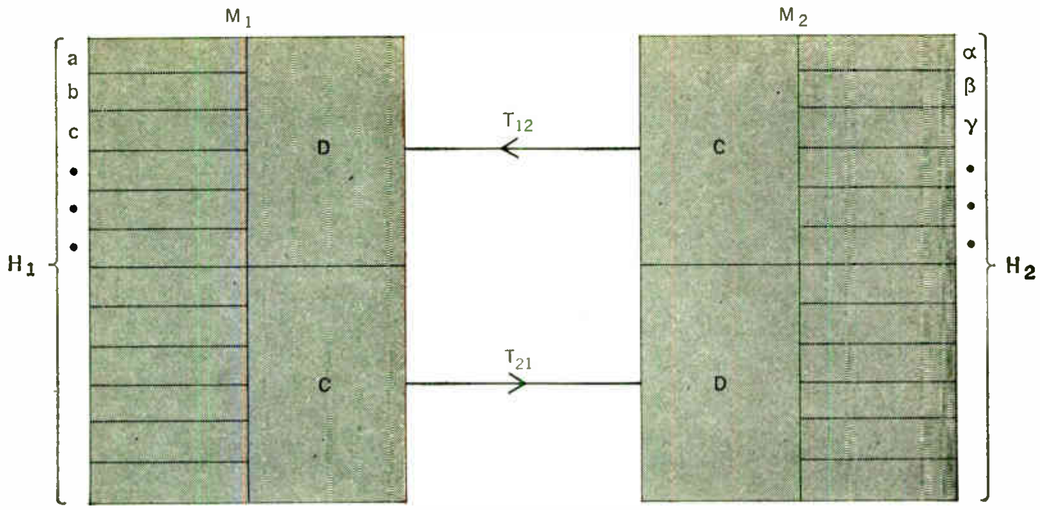


FIGURE 10. Diagram for communication between two people.

tion of information. On the left is the memory storage which stores 10^{10} to 10^{12} bits of information for a man. The association unit corresponds to a short-period memory; according to rough estimates[10] it may have a storage capacity from 50 to 500 bits. The association unit is connected with memory storage by a feedback loop. Depending on the information contained in the association unit at a given time, a new piece of information is selected from storage memory and transferred to the association unit. In this manner, a continuous process of subsequent messages occurs. In principle, the process is stochastic and not deterministic because it arises from feedback. Spälti and others[11,12] have shown that a feedback system, such as a sine-wave oscillator, operates in principle as a stochastic process.

The message contained in the association unit at any instant can be considered as the conscious message. The process goes on even if no signal arrives at the sensory input, corresponding to pure thought and independent of the external world. In succession, various messages in memory storage are sent to the association unit, which is made conscious of the messages. This process occurs at a speed that corresponds to the conscious information capacity, 10 to 50 b/s. If a sensory input is present, the resulting information of 10^6 to 10^8 b/s must be reduced to an information flow of 10 to 50 b/s. Neuronal interconnections, similar to the structure of the perceptron,[13] provide such information reduction.

Messages from the sensory input, after reduction, also arrive at the association unit and initiate a process which one may call the stochastic synchronization of the inner process. This means that the probability of the occurrence of the next message entering the association unit no longer depends only on the message already there, but also on the incoming message. The stronger the interconnection, the better the synchronization of the internal process with the external; in general, however, the synchronization has only a statistical character. One should note that the association unit, through the subprogram of the effectors (organs of response), also controls the effector output. The essential feature is the presence of an internal feedback process that is synchronized by an external process in a stochastic manner.

The following description of the model refers to information arriving in the association unit—that is, the information that becomes conscious. The definition of the quantity of information is difficult in practical applications; conscious and subconscious events cannot be sharply separated at all, but blend with each other. Figure 9 shows a layer-form extension of the model of Fig. 8 with several association units and internal feedback that takes this into account. Progressing toward higher internal layers, increasing information reduction occurs by decoding. Stochastic synchronization occurs in the association units with coupling between the statistical processes, where the innermost process has the lowest information flow. Such a structure could have been formed in man by a progressive coupling between the afferent and efferent paths. The first layer corresponds to a coupling in the sense of reflexes, from which, progressing inward, feedback processes develop.

The important characteristic of both models is the presence of one or more internal feedback paths that lead to stochastic processes. If one just considers the innermost processes, the description that follows is applicable to both model concepts. (The difficulty in defining information supply remains, and this arises in almost all applications of information theory to biology.) The basic model used contains only those function groups that are necessary for describing the generation of information and communication in man. It does not claim morphological equivalence, but its structure or function could lend insight to neurophysiological or experimental psychological studies.

Figure 10 describes the transmission of information between a pair of people, designated as M_1 and M_2. The functional model of Fig. 8 is represented in simplified form; D denotes the sensory input with decoding and C the effector output with coding. Because the available quantity of information is different in M_1 and M_2, separate message elements are indicated as a, b, c, . . . for M_1 and α, β, γ, . . . for M_2. The entropy of the two processes is identified as H_1 and H_2. The directional mean transinformation in the direction of $M_2 \rightarrow M_1$ is

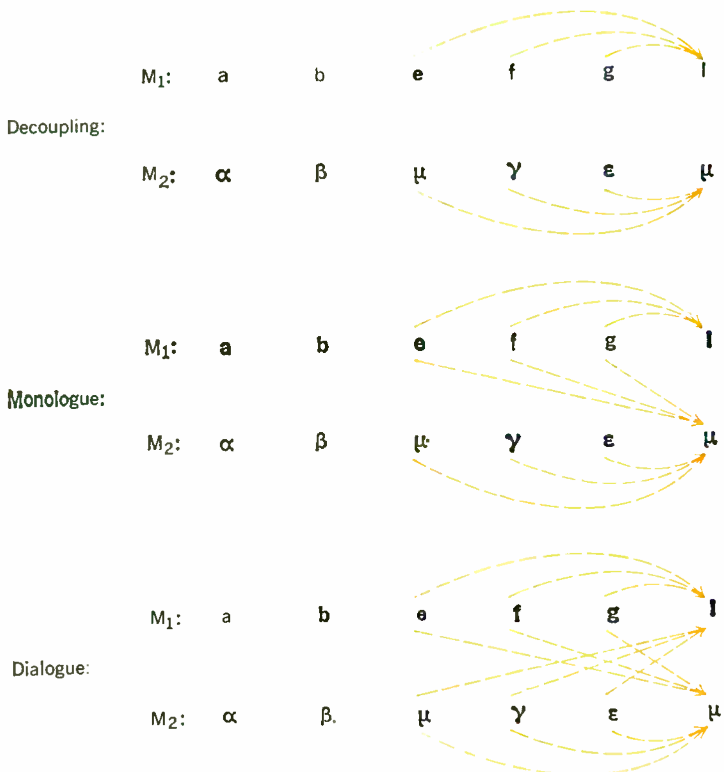


FIGURE 11. Time sequence of the stochastic processes of M_1 and M_2 and their stochastic coupling.

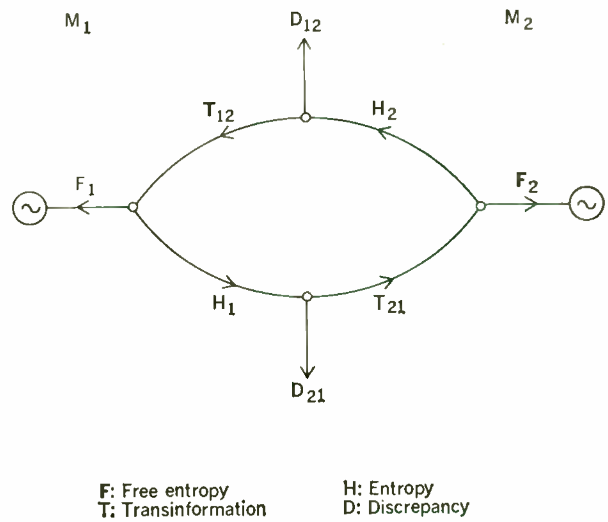


FIGURE 12. Information-flow diagram for communication, dialogue state of operation ($M_1 \rightleftarrows M_2$).

T_{12}; in the direction of $M_1 \rightarrow M_2$ it is T_{21}. (The first index of T refers to the receiver, the second to the transmitter.)

Figure 11 illustrates three stochastic processes of M_1 and M_2 and their statistical intercoupling (indicated by dashed lines). The choice of present message elements (symbols) depends on the past symbols of its own and foreign (external) processes. The intercoupling shown in Fig. 11 extends to three past symbols for each case and can be described by a third-order Markoff process. The length of the statistical intercoupling is determined by the capacity of the association storage unit where both processes are stored. The simplest case is a Markoff process of the first order, where only one symbol of its process and one foreign symbol are involved.

The following discussion applies to any stationary process.[9] Three possible states of operation, shown in Fig. 11, are defined:

1. Decoupling, $M_1 | M_2$. The communication connection in both directions has been interrupted; therefore $T_{12} = T_{21} = 0$. The two stochastic processes are independent of each other.

2. Monologue, $M_1 \rightarrow M_2$ or $M_2 \rightarrow M_1$. One communication is interrupted; for example, the connection from M_2 to M_1 is opened and $T_{12} = 0$. The process from M_1 to M_2 occurs and $T_{21} \neq 0$.

3. Dialogue, $M_1 \rightleftarrows M_2$. The communication connection is established in both directions, each process influencing the other; generally, T_{12} and T_{21} are not zero. This is the most general condition, and contains all special cases, including identically constructed systems and also the case where the synchronization is not stochastic but deterministic.

The case where the synchronization is deterministic could also be called suggestion. However, a deterministic synchronization is only possible in one direction. (A simultaneous suggestion between two people in both directions is impossible.)

Entropy H_1 and H_2 and transinformation T_{12} and T_{21} have been defined. It is required now to define some additional parameters:

$$F_1 = H_1 - T_{12} \text{ bits} \qquad F_2 = H_2 - T_{21} \text{ bits}$$

The total information of M_1 is H_1 and T_{12} is the information M_1 receives from M_2; the difference F_1 is the free entropy and represents the amount of information created by M_1 itself. The same kind of definition holds for F_2.

Discrepancy parameters D_{12} and D_{21} are defined as

$$D_{12} = H_2 - T_{12} \text{ bits} \qquad D_{21} = H_1 - D_{21} \text{ bits}$$

The use of these parameters is shown in the information-flow diagram of Fig. 12. From the equations for F_1 and F_2, it is seen that the total entropy consists of freely created information (free entropy) and the received transinformation. Further, knowing the entropy, the discrepancy required to act in the direction of transinformation can be determined.

Figure 13(A) shows an information-flow diagram for the monologue, $M_2 \rightarrow M_1$. This case corresponds to the unidirectional channel of Shannon, where D_{12} is the equivocation and F_1 the irrelevance. A diagram for the degenerated case of decoupling is given in Fig. 13(B).

Taking the ratio of transinformation received to the entropy yields another useful parameter, the stochastic degree of synchronization (psychologically, the degree of perception)

$$\sigma_1 = T_{12}/H_1 \qquad \sigma_2 = T_{21}/H_2$$

Parameters σ_1 and σ_2 may vary between 0 and 1. The condition $\sigma_1 = 1$ represents a suggestion; the free entropy disappears and the entropy is determined solely by the transinformation received. One finds that either $\sigma_1 = 1$ or $\sigma_2 = 1$, but both cases cannot occur simultaneously; that is, a simultaneous suggestion in both directions is as stated before.

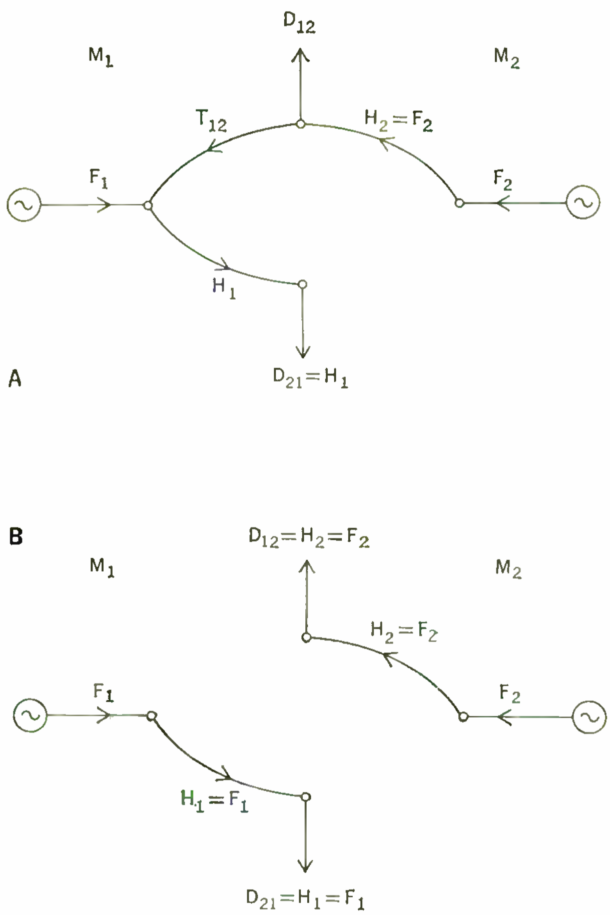


FIGURE 13. Information-flow diagrams. A—Monologue, $M_2 \rightarrow M_1$. B—Decoupling, $M_1|M_2$.

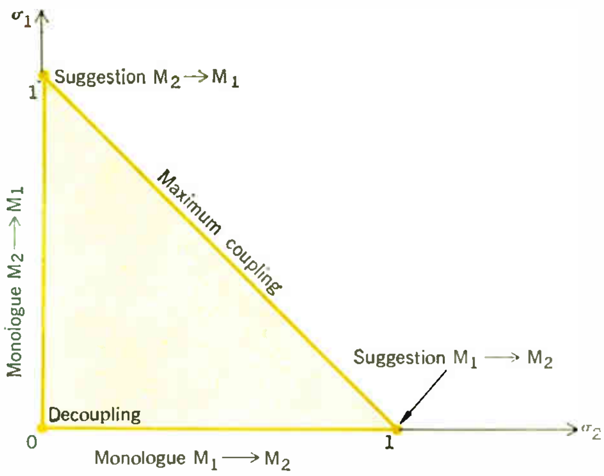


FIGURE 14. Possible values for stochastic synchronization degrees, σ_1 and σ_2.

Figure 14 indicates the possible values of σ_1 and σ_2, determined from the equation for both degrees of synchronization

$$\sigma_1 + \sigma_2 \leq 1$$

Values for σ_1 and σ_2 are contained within the area of a right triangle. The hypotenuse represents the case for $\sigma_1 + \sigma_2 = 1$, or $F_1 = T_{21}$ and $F_2 = T_{12}$. Generated information is actually exchanged, corresponding to maximum coupling. Both sides of the triangle correspond to the monologue ($\sigma_1 = 0$ or $\sigma_2 = 0$) and to Shannon's unidirectional channel. The two corner points represent a suggestion; the zero point, decoupling. All borderline cases are contained in the diagram, demonstrating that the theory of bidirectional communication is a generalization of Shannon's information theory.

Conclusion

The application of information theory to problems in cybernetics can be rewarding; new insights are gained in the understanding of biological and psychological processes. The concept of bidirectional communication represents the beginning of a quantitative study of information theory applied to man, with the hope that it will be possible to describe quantitatively the generation and transmission of information between individuals. The theory has been described more extensively in previous papers by the author.[8,9] The parameters for communications between three systems under limited conditions have also been described in the literature.[14]

Only stationary conditions have been considered, where the structure of the model is unaltered by the process. As a next step, learning processes should be considered. This will result in changes in structure and in statistical coupling between systems.

REFERENCES

1. Shannon, C. E., "A mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423, July 1948; pp. 623–653, Oct. 1948.

2. Marko, H., "Die Ausnutzbarkeit eines Telegrafiekanals zur Informationsübertragung," *Nachrtech. Z.*, vol. 15, pp. 451–466, 1962.

3. Marko, H., "Die Kanalkapazität des Telegrafiekanals und verwandter Übertragungskanäle," *Nachrtech. Z.*, vol. 16, pp. 480–485, 1963.

4. Marko, H., "Physikalische und biologische Grenzen der Informationsübermittlung," *Kybernetik*, vol. 2, p. 6, 1965.

5. Hauske, G., "Ein elektronisches Funktionsmodell für Verhaltensweisen eines Fisches," *Kybernetik*, vol. 3, p. 1, 1966.

6. Küpfmüller, K., "Informationsverarbeitung durch den Menschen," *Nachrtech. Z.*, vol. 12, pp. 68–74, 1959.

7. Küpfmüller, K., and Jenik, F., "Über die Nachrichtenverarbeitung in der Nervenzelle," *Kybernetik*, vol. 1, pp. 1–6, 1964.

8. Marko, H., "Die Theorie der bidirektionalen Kommunikation und ihre Anwendung auf die Informationsübermittlung zwischen Menschen (subjektive Information)," *Kybernetik*, vol. 3, pp. 128–136, 1966.

9. Marko, H., and Neuburger, E., "Über gerichtete Grössen in der Informationstheorie (Untersuchungen zur Theorie der bidirektionalen Kommunikation)," *Arch. Elek. Übertragung*, vol. 21, no. 2, pp. 61–69, 1967.

10. Frank, H., *Grundlagenprobleme der Informationsästhetik und erste Anwendung auf die mime pure*. Waiblingen, West Germany: Hess, 1959.

11. Spälti, A., "Der Einfluss des thermischen Widerstandsrauschens und des Schroteffektes auf die Störmodulation von Oszillatoren," *Bull. Schweiz. Elek. Ver.*, p. 419, 1948.

12. Marko, H., "Informationstheorie und Kybernetik," *Fortschr. Kybernetik*, pp. 9–28, 1967.

13. Rosenblatt, F., *Principles of Neurodynamics, Perceptrons and the Theory of Brain Mechanisms*. Washington, D.C.: Spartan Books, 1962.

14. Neuburger, E., "Beschreibung der gegenseitigen Abhängigkeit von Signalfolgen durch gerichtete Informationsgrößen," *Nachrichtech. Fach.*, vol. 33, 1967.

The November issue of the Proceedings of the IEEE will be a special issue on computer-aided design, consisting of many papers on specialized aspects of this growing field. To put the new development in proper perspective, this article examines just what it is that justifies all the excitement. It discusses the nature of the engineering design process, man–computer systems, both passive and active computer graphics, nongraphic applications, and problems of computer technology, and points out for engineers the importance of computer-aided design for the future.

Computer-aided

The digital computer has long been utilized by the design engineer to aid him in solving engineering problems. Now, the development of the active computer graphics system will provide him with the capabilities he requires

Design engineers in many disciplines have been using electronic digital computers to solve engineering problems for almost two decades. They have been one of the principal groups of users of this tool—indeed, some of them have been engaged in developing and improving the tool itself. Computer applications in design engineering are far from being a new development.

What is a new development is the use of graphical devices for on-line communication with the computer. There is more excitement among design engineers about this new development than there has been about anything since the advent of the computer itself. It is not surprising that engineers are showing the way in the development of applications for these new devices.

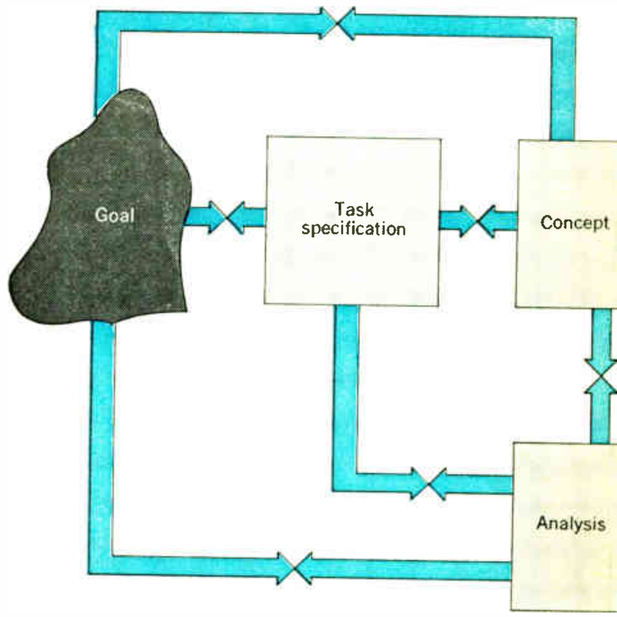


FIGURE 1. The engineering design process.

The design process

The use of active on-line graphical devices as aids to the design process can be understood only when the design process itself is understood. Every useful computer application has the objective of performing some process better or faster or both. The design process has certain key characteristics that must be recognized by a computer system intended to improve the manner in which it is performed.

A particularly useful description of the engineering design process has been articulated by Profs. Robert W. Mann and Steven A. Coons of the Mechanical Engineering Department at the Massachusetts Institute of Technology. A graphical representation of the design process as they described it appears in Fig. 1.

The design process starts when a goal has been perceived and adopted. The goal may be a space vehicle, a bridge, a new model of last year's automobile, or a camera capable of instant photography. The design engineer's first job is to delineate the task in as specific and quantitative a fashion as possible. Such task specifications describe the performance required and serve as guidelines for activity to follow and as criteria for judging the final output of the design process.

To proceed further requires an idea for a solution. This is the creative phase of design. The engineer draws on his experience and knowledge to formulate a concept that may satisfy the specifications. The concept, initially quite vague and ephemeral, firms up as information is gathered and the tentative solution is filled in with detail.

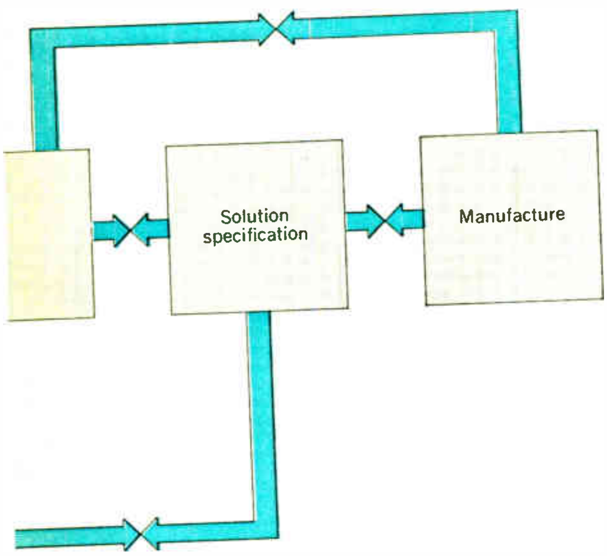
It is then evaluated by being subjected to both qualitative, intuitive judgment and to very precise quantitative analysis. The esthetic value of the proposed solution is evaluated in a purely judgmental fashion, whereas its ability to withstand stress might be evaluated by a mathematical process. The processes of conceptualization and analysis are carried out with heavy reliance on graphical representation.

When these steps are complete, the specification of the solution is largely complete. What remains is the development of manufacturing information and the application of production engineering procedures. The manufacture of prototypes or pilot runs then permits the determination of the extent to which the design fulfills the original goal.

The entire process is characterized by iteration; it does not occur in a simple, sequential fashion. The concept of

design

Ronald A. Siders
Booz, Allen and Hamilton, Inc.



a solution, when analyzed, may exceed or fall short of the task specifications. If so, the goal may be changed by adopting the increased performance, or by relaxing the requirements. As the concept is crystallized and detailed, it may become apparent that manufacturing or cost considerations require a change. Such a change must be appraised for its impact on the rest of the design and on the fulfillment of the task specifications. The process is a sequence of decisions; each decision must be evaluated, and altered if necessary.

The design process is characterized by a few brief moments of creativity and comparatively lengthy periods of mechanical computation and documentation. The conceptualization of a solution requires the engineer to draw on all the experience and knowledge at his disposal. The analysis and documentation of the concept constitute a mechanical process with correspondingly less opportunity for creativity.

Design problems tend to be hierarchical; that is, they tend to consist of a set of problems at several levels. The entire product or system is the highest level. It is composed of subsystems, which, in large design projects, are the responsibility of different individuals. And each subsystem may be made up of elements that may be further delegated. For lesser levels to be specified, the higher level must be specified. At each level there is opportunity for creativity, and a set of appropriate analyses to be conducted. At each level iteration occurs between concept, analysis, and solution specification, and a change at one level can require corresponding changes at the same or other levels. Control and coordination of the design project are of paramount importance.

The individual design engineer must be aware of the specifications for his part of the design. He must be prepared to draw on experience, intuition, and judgment to arrive at a satisfactory solution. He must have standards available for his use and guidance. He must maintain a record of the current form of his design. He must be notified if adjacent elements or his task specifications have changed. He must document his solution.

The salient features of the design process can be summarized as follows:

- Its principal stages are task specification, solution conceptualization, analysis, solution specification.
- The process is iterative.
- Its brief moments of creativity are interspersed with lengthy analyses.
- Intuition and judgment fill the gap when physical principles fail.
- The process proceeds at a number of levels, which must be coordinated.
- Standards are used as aids and controls.
- The design is recorded, primarily in graphical form.

When viewed in an industrial setting, the design process exhibits additional features that provide the key to understanding management interest in on-line graphics systems. These are features that are basic to the industrial activity of developing and introducing new products.

During the period of time between the adoption of the goal of introducing a new product and the introduction of that product, the business fundamentals of risk and gain are at play. As the design process proceeds, the cash exposure of the company to the project increases; funds are expended in anticipation of a future gain. Although the risk of a design failure is reduced as the answers to the design problems are found, the flexibility to change the fundamental approach is lost, and the risk of a shift in consumer tastes increases.

As corporations maneuver for competitive advantage, there is a premium for reducing the time required for design and for increasing the efficiency of the process. In this manner the risk is reduced, and, because it promises to do just that (and by a substantial factor), computer graphics is drawing management attention.

The man and the computer

A computer-aided design system employing on-line graphics devices hypothesizes a close working relationship between the man and the computer, and a well-designed system must consider their respective attributes.

The computer's particular attributes include speed, memory, and reliability. It performs computations and makes logical decisions at a rate that is difficult to comprehend. Its memory is capable of storing and accessing large amounts of data as instructed. And it performs these functions in an extremely reliable fashion. Although the computer cannot execute any calculations or sequences of logical operations unless they can be predetermined in detail, it can go through them rapidly and accurately once they are determined.

The man, on the other hand, is not good at computation. He is slow; he makes mistakes; and he has a poor memory. He does, however, have some attributes that are indispensable to the design process.

The design engineer brings to the design problem his experience, imagination, social and esthetic values, and a tolerance for ambiguity. He has a massive store of information about the world, about design, about relationships. He can be creative, innovative, and inventive.

The designer is able to judge alternative designs against his social and esthetic values. The precise curve of an automobile roof may never have existed before—the designer selects it from among the endless alternatives available because it pleases him. The selection cannot be programmed because it is not made on a basis that is quantifiable.

At some stage in the design process the physical principles and quantitative tools at the disposal of the engineer may not be sufficient to determine the way to proceed. However, the engineer can go ahead without adequate data. He may later reverse his decision, but he can make it. He is capable of intuition and hunch. He can formulate unanticipated questions, and he can decide on a procedure by which he can arrive at an answer.

What the machine can do better, it should do. What the engineer can do better, he should do.

The ideal system

We can now hypothesize the ideal computer-aided design system. We shall then be able to judge what is available today in comparison with this ideal.

The ideal system must facilitate interaction between the engineer and the computer. When the designer needs the computer's assistance, it should be available. The process of brief, seemingly random moments of creativity interspersed with lengthy periods of computation, analysis, and documentation must give way to a process wherein the time devoted to computation, analysis, and documentation approaches zero.

Should the analytical procedure be lengthy, even though done by the computer, the man must be able to monitor its development and intervene if meaningful results are not forthcoming. For this reason, the computer must keep the designer apprised of its progress.

Communication between the man and the computer must be in a language natural to the man; the natural language of design engineers is graphics.

The response time of the computer to instruction must be established by the man's reaction time. The designer must feel that progress toward a solution to the design problem is governed by his own ability to proceed, not that of the machine.

The computer must take over the maintaining and accessing of a file of standards. Standard elements, characteristics of materials, and production standards must all be available instantly when needed.

The computer must be given the job of recording the evolution of the design. A record of the latest version of the design for each element, component, subsystem, and system must be immediately retrievable. Further, the computer should take over the task of integrating the subsystems within the whole with a minimum of guidance by the designer. Interferences between elements must be detected and brought to the attention of the engineer. A change made by one designer should be reflected throughout the system in the affected areas. If changes cannot be made automatically, at least an indication that the engineer's attention is required must be given.

The computer must perform those steps of computation and analysis that are deterministic and programmable. This includes the simulation of performance of the total design by integrating subsystems into an overall model and testing performance as measured by key variables on the basis of test input.

Indeed, the computer should be prepared to provide optimum or standard elements in the design automatically when such are required and are available.

Nongraphic applications

Significant progress has been made toward extensive utilization of the computer in the nongraphic activities of design. It was to the task of computation that the machine was first put. The solving of equations and the manipulation of large amounts of data were the first design tasks assigned to the machine.

Computation occurs principally during the process of analysis, but it is not the whole of analysis. This activity also includes the structuring of a model of the concept, the determination of the pertinent physical principles, and the simulation of performance.

Analytical applications have been put on the machine in growing numbers in the past decade. To do so is a larger task than simple computation. Undertaking analysis on the computer requires that the concept for the design be developed in sufficient detail to permit the construction of a conceptual model. Such a model must reflect the pertinent physical principles, and can be used to predict performance under varying conditions.

Programs have been developed in the aerospace industry to analyze aircraft engine performance. Engineers input the required design parameters in various combinations and receive detailed performance measurements as output. The program permits the testing of various combinations of design parameters in the search for performance fitting the application envisioned.

Programs also have been developed for analysis of the structural dynamics of aircraft and space vehicle design. The objective of this application is to permit the examination of the flutter and vibration characteristics of a given airframe design as early in the design process as possible. These programs assist the designer to answer the fundamental questions of structure—e.g., placement of engine mountings on an aircraft wing or fuselage long before a model can be tested in a wind tunnel.

The analysis of the loading performance of a metal frame building is done today by computer. The engineer enters the loading specifications of the proposed metal frame building and the preliminary design he thinks will meet them. He receives a report showing the forces the proposed design will withstand compared with actual forces resulting from the loading specification—with differences and percentage of variance shown, as well as material and fabrication costs. By varying the input, the engineer approaches the balance of style, strength, and cost that is optimum.

A proposed automobile windshield design must be evaluated for the amount and severity of the distortion it will cause in the driver's view. In the past this evaluation required the construction of a prototype windshield in order to photograph a reference grid with and without the prototype in place. Visual comparisons of the two exposures were used to judge the suitability of the design. Today, a computer program performs the evaluation by calculating the comparative grids based on the space coordinates of the driver's eye and the grid, and the mathematical definitions of the inner and outer windshield surfaces.

A problem in the design process that we previously noted is the integrating and coordinating of individual design efforts to avoid interference and to ensure that components, when assembled, will function together to fulfill the performance requirements established for the system. An example of a computer application serving this purpose is found in the automotive industry in the design of window regulator mechanisms. The location of pivots and the length of arms in the mechanical linkage must be placed so as to meet requirements of mechanical advantage, but interference of the glass, while it is being raised or lowered, with other components of the design must be avoided. The computer program generates a set of values indicating where the key elements can be located and the lengths of the arms that should be used.

An extreme example of engineering drudgery being assigned to the computer is the IBM-developed ADE (automated design engineering) system in use at I-T-E Circuit Breaker Co. This system produces detailed design specifications for transformers. Program input consists of customer specifications for coiling media, kVA, voltages, impedance, cycle, physical characteristics, and other options. The system assumes a standard core and designs low- and high-voltage coils consecutively. Each coil design is iterated until temperature rise, impedance, copper loss, and dimensional clearance are satisfactory. The program prints out detailed instructions for fabrication and assembly of these electrical parts of the transformer. It then proceeds to the mechanical design, ending with the generation of a bill of materials and assembly plans. In this application, each element of the design is formulated, tested for performance and fit, and then kept or rejected. The iterations required to find an acceptable solution are completely specified.

The ADE system represents an automated design application extending far beyond mere computational assistance or analysis. Given the task specifications, the system generates appropriate solution specifications. The highly structured nature of the design process permitted the assignment of the entire design task to the computer.

Despite the extent to which design activities are being transferred to the computer, the design engineer will continue to play the major role in most design processes for the foreseeable future. He performs functions that cannot be displaced by a machine incapable of making judgments, overcoming ambiguity, or expressing esthetic values.

Passive computer graphics

Passive computer graphics devices provide the capability for communicating graphical information to or from the computer in an off-line manner. The hardware devices associated with this kind of application include a variety of image recorders and scanners, cathode-ray-tube displays, and digital plotters.

Some equipment scans drawings on paper and converts them, through a microfilm medium, to digital form for computer storage. The General Motors DAC-I (design augmented by computer) system includes such a device. The scanning is done under program control but with an operator watching over the process to intervene if incorrect interpretations are being made by the scanner. The problem of interpretation arises when there is no clear way for the computer to determine which of several alternative interpretations to make. An example is a figure eight, which could be two juxtaposed closed loops.

Digital plotters have been developed in a variety of sizes and with a variety of capabilities. Large, accurate plotters have been used for some time to produce full-scale drawings and accuracies within 0.0025 cm are common. These plotters are generally controlled by an electronic control unit, which reads magnetic or punched tape and performs the digital-to-analog conversion. Slow plotting speeds make direct, on-line connection to a computer impractical.

The Stromberg Carlson 4020 is an example of equipment that converts and displays on a cathode-ray tube images stored in digital form on magnetic tape. The resulting images can be observed directly or recorded on a film strip by an automatically operated camera.

One current application of this equipment is in the production of wiring and cabling diagrams in the aerospace industry. The file of images on the magnetic tape is prepared on the computer by programs under the control of the design group. The programs are capable of determining optimum lengths and routings for the wiring and harnesses. The SC 4020 converts the magnetic tape file to images on film, which may then be used to produce drawings as needed. This application of passive graphics techniques is in the solution specification phase of the design.

The same device has also been utilized in conjunction with the computer to assist in the analytical phases of design. Specifically, it has been used to produce film strips, which, when projected and viewed, depict the expected performance of aircraft that are still in the initial design phase. This is accomplished by digitizing orthographic views of the aircraft with a Telereader (which transcribes the points to x, y, and z coordinates) and processing these data in the computer to develop digitized perspective views along the simulated flight path. These data on magnetic tape are then processed through the SC 4020 to develop the film strips. Other applications include the evaluation of pilot visibility resulting from alternative cockpit and canopy configurations, of proposed designs for engine mountings to permit rapid removal and replacement, and of aircraft landing characteristics under day and night conditions.

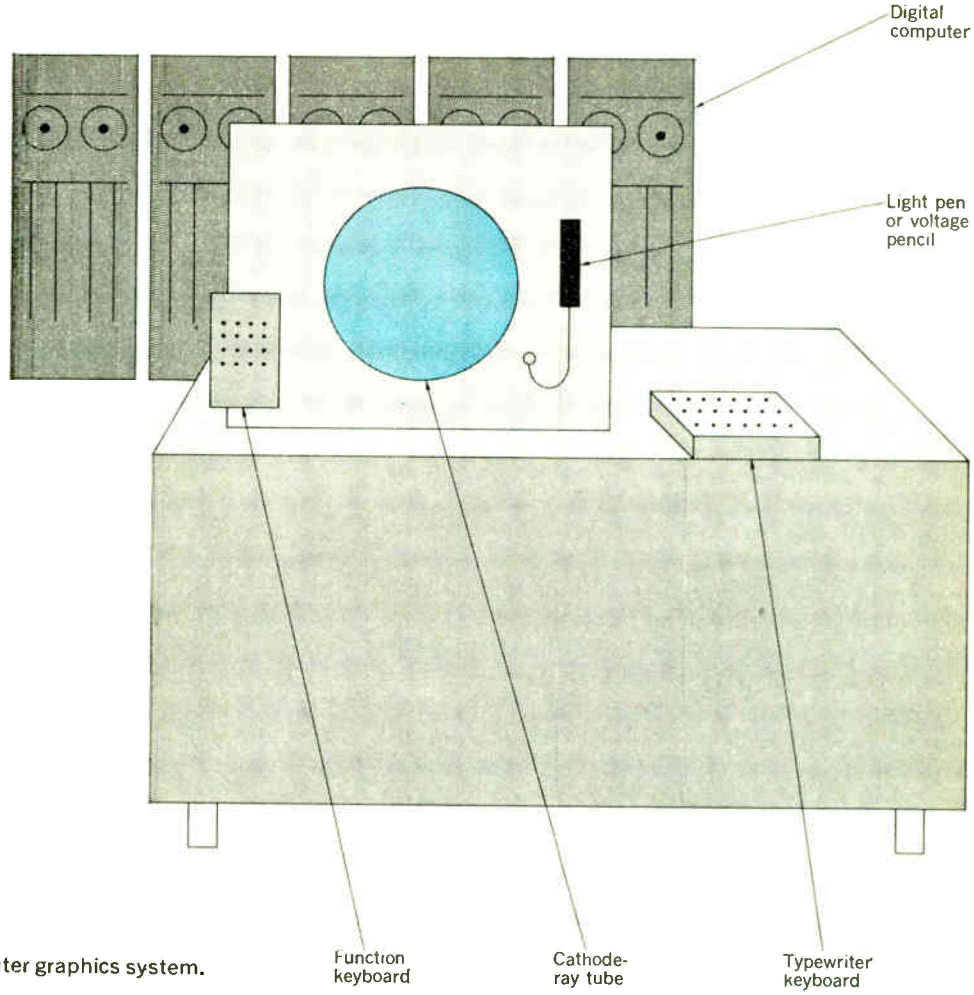


FIGURE 2. A general computer graphics system.

The results of such a system are amazingly realistic and provide real assistance to senior designers in analyzing certain performance characteristics. Such assistance in visualizing a complex process can contribute significantly to avoiding gross errors or oversights, which, when discovered early in the process, are far less costly to correct.

Passive computer graphics has its greatest application value in the solution specification phase of the design process. The preparation of precision drawings in a variety of scales and views and the preparation of templates directly from the computer-stored design provide accurate and reliable manufacturing specifications. The bill of material explosions and the assembly instructions are automatically generated. When combined with automatically prepared control tapes for numerically controlled machine tools, the total package bridges the gap between design and manufacture with a wealth of information to ensure the implementation in production of the design group's intent.

Problems of computer technology

Historically, whenever engineers have sought to use the computer to assist them in the design process, they have encountered two obstacles: the problem of communicating with the computer, and the lack of computer availability in time and space.

The communication problem has existed because of the computer's requirement to have all of its instructions ultimately cast in binary form. Initially, the engineer had to learn machine language in order to express his computations in a form that could be input to the computer. Assemblers were developed to translate machine-language mnemonic codes to binary form.

In addition to the burdensome chore represented by the process of learning the machine language, this restricted means of communication meant that the engineer had to recast his problems in a form that was expressible in the machine's limited vocabulary. This time-consuming chore of recasting the desired procedure into a suitable form could be given to a trained computer programmer; but, if the engineer did not know programming and the programmer was not familiar with engineering, the process was susceptible to gross error.

The introduction of Fortran in 1957 allowed the engineer to express his problems in algebraic terms, which he more readily understood. By and large, the engineer became his own programmer. Computer usage soared as the engineer became capable of visualizing and expressing his computational problems in a language with nearly universal acceptance.

Fortran, however, lacked the facility for expressing problems in the language of particular disciplines and so a number of problem-oriented languages have been developed to overcome this shortcoming. Such languages as

Cogo (Coordinate Geometry for civil engineering) and Stress (STRuctural Engineering System Solver) have helped to overcome the obstacles that prohibited rapid and useful communication with the computer. But they, too, failed to provide the ability to converse in the form most natural to engineers, i.e., graphics.

The second problem faced by the engineer in his quest to make use of the computer was that of availability. The high costs of equipment rentals required the machine to be utilized as fully as possible. The typical scheme for achieving such maximum utilization has been batch processing, in which the problems of a number of users are batched together and run as a single unit.

The frustrations of batch processing for the engineer are many. The turnaround time between the time he submits a run and the time he receives his results is excessive. The rate at which he progresses with his work is paced by the batch-processing schedule and not by his own ability to proceed. He has to rethink his problem every time he gets results in order to prepare input for the next batch. And he wastes large amounts of computer time just trying to localize the computations to the area of interest.

This problem of computer availability has been recognized for a long time, and a great deal of progress has been made in developing methods to solve it. The methods have centered around the use of remote terminals for input and output to the computer, and the development of computer monitoring or operating systems that permit it to respond to remote inquiries on a real-time basis.

Time-shared systems are the most advanced schemes that have been developed to accomplish real-time operation. In these systems, the computer services a large number of remote terminals, scanning the active stations according to a scheduling algorithm, and allocating sufficient equipment and time to each in turn, so that to each user it appears that he has full, continuous use of the computer's capabilities. Such systems have been developed for a variety of equipment.

Multiprogramming schemes have been devised for cases in which it is assumed that a considerable portion of the computer's time will be spent on batch-processing operations. With this arrangement, provision is made for the batch mode to be interrupted by a call from a remote terminal. The call is serviced, usually to some logical stopping point, and the batch mode resumes. This approach is desirable when a computer facility is to run several applications only some of which make use of remote terminals.

Time-sharing and multiprogramming operating systems have come a long way toward solving the problem of computer availability. When combined with problem-oriented languages, as they have been in some cases, the engineer is provided with a very powerful and useful tool. By taking the next step, that of incorporating remote terminals with graphics capabilities, the computer finally approaches its ultimate potential for assisting design.

Active computer graphics and applications

The basic active graphics hardware is typically grouped into a console as shown in Fig. 2. The equipment usually includes a cathode-ray tube, a light pen or voltage pencil, a "function" keyboard, and an alphanumeric keyboard. Graphic input is accomplished by drawing on the face of the scope with the light pen or voltage pencil. Graphic output is displayed on the scope. Nongraphic communication is accomplished by use of the alphanumeric keyboard. The operator typically directs the process by use of the function keyboard and by indicating areas of interest on the display with the light pen or voltage pencil.

The light pen is a fiber optics bundle open at the end directed toward the scope. The other end terminates at a photomultiplier tube, which is actuated whenever the light pen receives a flash of light from an illuminated point on the scope. The point at which the light pen is directed is determined by correlating the time the flash is received with the position of the electron beam that generates the display.

Drawing on the scope face is accomplished by calling onto the screen a "tracking cross." Each leg of the cross is made up of a line of illuminated spots. To begin drawing, the light pen is pointed at the center of the cross. As the light pen is moved, the computer directs the tracking cross to reposition in such a way as always to remain under the center of the pen's field of view. To draw a straight line the operator would move the tracking cross to the starting position and depress the function button "straight line begin." The pen would be moved across the screen to the desired termination point, and the "line end" button would be depressed. If desired, the computer would ensure that the line is geometrically straight.

Additional hardware available at the terminal might include a card reader for inputting limited amounts of data. Some configurations use punched paper tape readers for the same purpose.

The computer can be directed through a set of procedures by the use of a pointing process, in addition to the function keyboard or typewriter. The control program displays a series of alphanumeric lists of alternative procedures, and the operator directs the operation by pointing in turn to those desired. In a long and complicated series of procedures, this "pointing" permits very rapid control as compared with entering typewritten instruction codes.

Consoles of the sort described here are commercially available from a number of hardware manufacturers; some examples are shown in Figs. 3 and 4. They or their experimental forebears have been under development for several years. Today, they are coming into production use in a few large progressive companies.

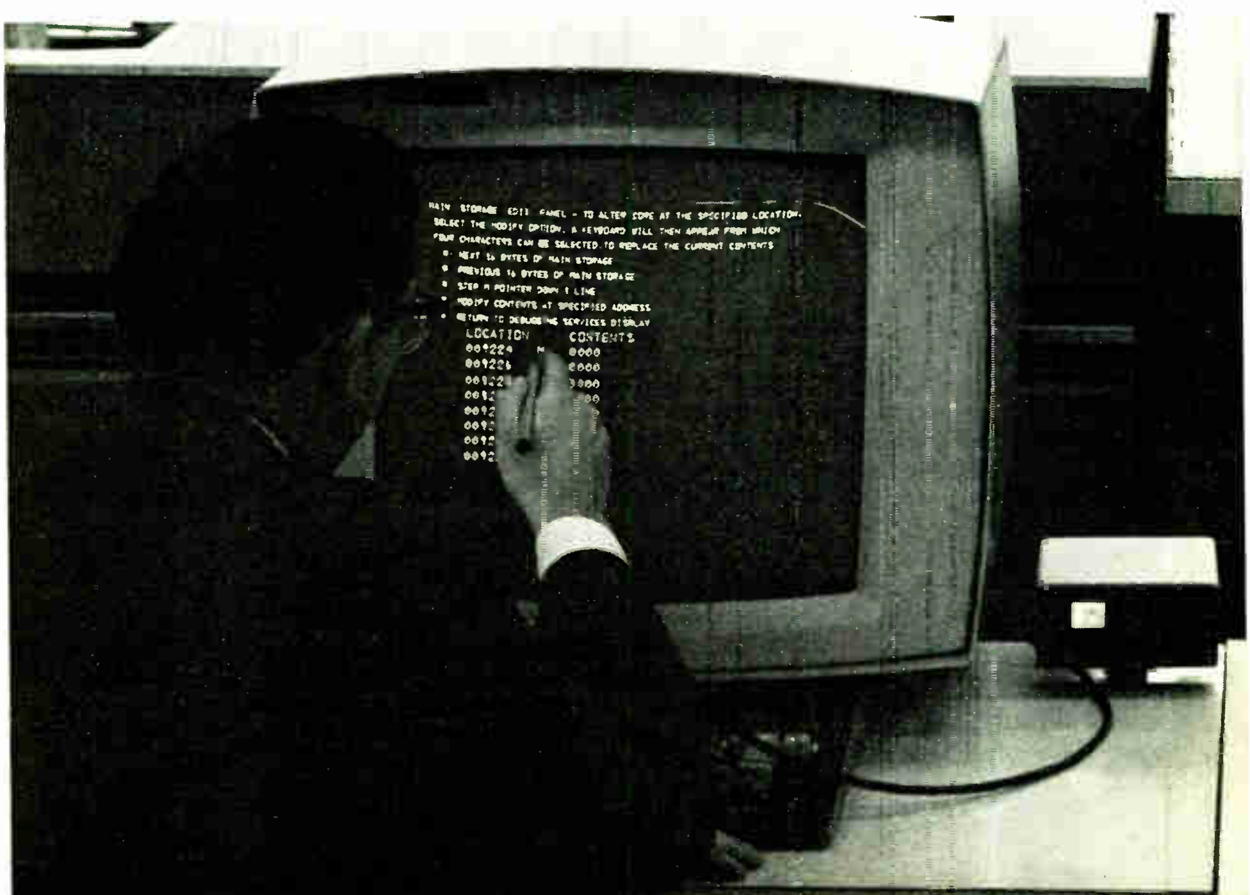
All of the normal operating modes are currently being used to service these consoles. Some have computers available to the console on a full-time basis. In general, this approach is used during the initial stages of the installation while application programs are written and checked out. Multiprogramming operation is used when the full-time use of a computer cannot be provided. Productive batch-processing operations are conducted while servicing the console on an interrupt basis. Multiple consoles operated under a time-sharing system are in existence, and represent the best system for achieving the full benefits of active graphics. Such systems make economic use of the central processor, providing real-time, interactive service to several designers simultaneously.

The amount of central processor time required to service each console is appreciably reduced by incorporating display maintenance circuitry in the console. Most of the time, an operator is studying the display, making some notes, or reaching for the light pen. During these periods the central processor is free to service other consoles.

We noted that passive graphics systems have been ap-

FIGURE 3. The Control Data Corporation's Digigraphic 270 system.

FIGURE 4. The IBM 2250 graphic data-processing system



plied predominantly to the solution specification phases of the design process. In contrast, active graphics systems are being applied to the conceptualization, analysis, and solution specification phases. The ability to change a design configuration rapidly directly on the scope face and to have analytical results quickly displayed enables the designer to make swift progress toward a satisfactory solution to his problem.

The first requirement for an active graphics system is to accept graphical input on the scope face. A variety of capabilities of this type have been developed. Existing systems currently accept points, lines, circles, general conics, and free-form lines directly on the scope. Geometrically perfect forms can be defined by inputting parameters. Scale can be changed by almost any useful factor. Elements of a drawing can be deleted. The drawing can be rotated about any axis, giving the illusion of three dimensions. Dimensions can be automatically calculated and displayed. Components can be duplicated as often as desired and placed in mathematically correct position (such as teeth on a gear). Orthographic projections can be interchanged with perspectives and isometrics.

One of the most powerful capabilities is that of calling up from core a wide range of standard components to add to the design on the scope. A circuit of standard electronic components can be rapidly laid out. At any point, the design being worked on can be filed away in memory for later recall. In this fashion parts of the overall design can be worked on independently and called up for examination singly or as a whole.

With these abilities to construct designs on the scope face and to have the corresponding specifications stored

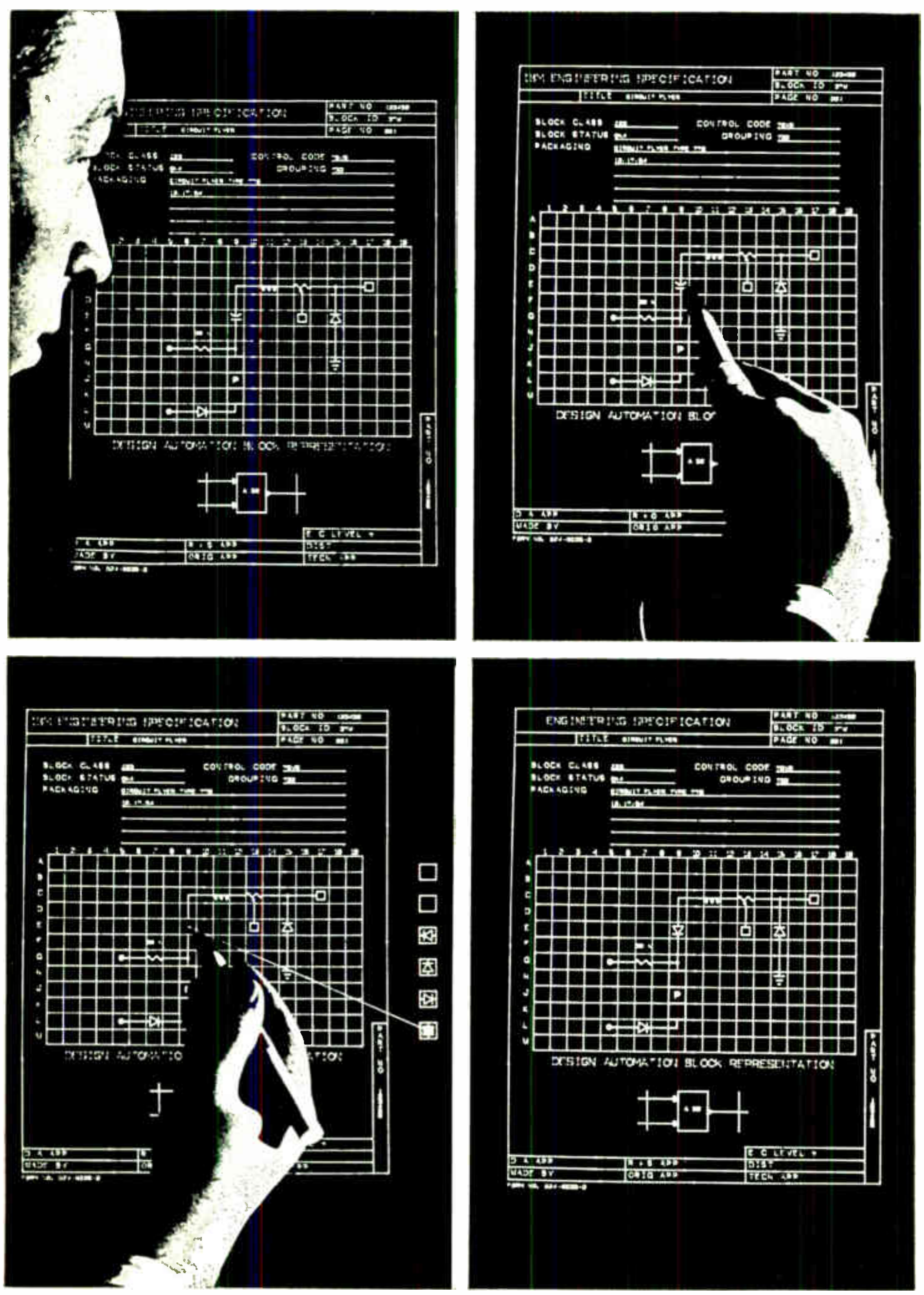


FIGURE 5. An electronic circuit being designed on the IBM 2250 shown in Fig. 4. With this capability, a simple RLC network, for example, can be analyzed by constructing the circuit on the scope, specifying the values of the components and the shape and value of the input waveform.

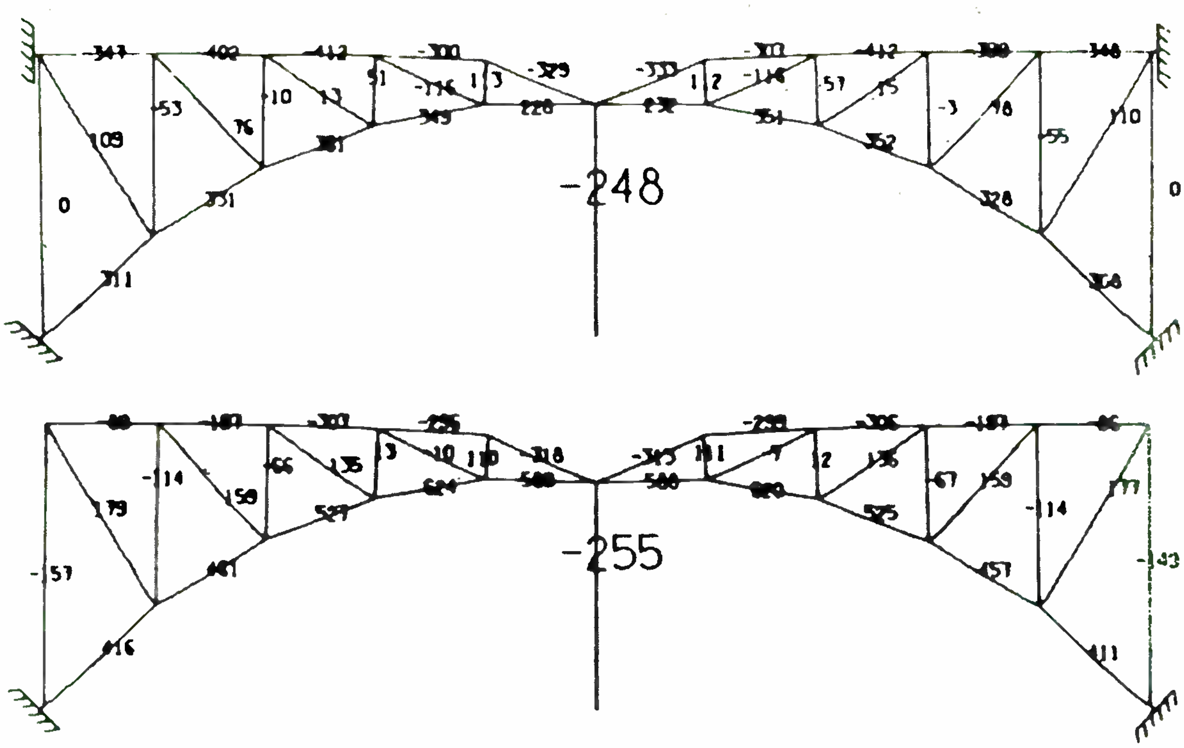


FIGURE 6. Sketchpad truss bridge.

internally in the computer, it is possible to perform certain analyses on the design. For example, a simple *RLC* network can be analyzed by constructing the circuit on the scope, specifying the values of the components and the shape and value of the input waveform, and having the output waveform appear on the scope. The designer can change any component or the input waveform and observe the effect on the output waveform. An example of a circuit being constructed is shown in Fig. 5.

One of the original applications on the prototype graphics hardware developed at the Massachusetts Institute of Technology Electronic Systems Laboratory was the analysis of stresses in a truss bridge. The state of stress or deflection resulting from various loads was indicated directly on the scope face as shown in Fig. 6.

When the engineer completes his analysis and produces a final design, a precise definition of it is available in computer memory from which to generate hard-copy drawings or even control tapes for numerically controlled machine tools. To program the cutting instructions on a drawing on the scope has been shown to take only about one sixth the time formerly required when working on hard-copy drawings.

A wide variety of applications to drafting, drawing, analysis, and parts programming have been developed. Within the past year, these applications have moved from the developmental stages into production use.

Active computer graphics systems are important because they finally provide the engineer with the capabilities he requires. Time-shared operation of graphics consoles provides the close interaction between engineer and computer that is necessary for maximum productivity. Communication is in a form natural to the designer. The designer can proceed at his own pace, relying on the computer to accomplish the otherwise time-consuming analyses. The computer maintains a record of the current version of the design. When given information about the linkages between sections of a design, it can accomplish the integration into a single design and show the result. Files of standard components are kept automatically and can be rapidly accessed. Internally stored designs can be converted to hard-copy documentation at any point.

As these active graphics systems are further developed, the engineer will be freed more and more from the clerical chores associated with design to apply his creativity to the development of better designs. Although increased productivity is hard to measure in an activity such as design, the fact that increased productivity does result is apparent. Current users claim reductions of from 2:1 to 6:1 in the time required to perform certain design functions. There is more than sufficient justification for management interest in this type of increased productivity.

The future

Computer aids to design have come a long way from the original uses of vacuum tube computers for computational purposes, and few would claim the end of this evolution is in sight. For example, exciting developmental efforts are being undertaken in three-dimensional displays. Without looking that far down the road, however, just the application of existing technology to today's design problems promises substantial excitement. Engineers by their very nature have shown the way in the past and can be expected to continue to do so.

Future goals of engineering in biology and medicine

However one labels the recent biomedical engineering conference —interdisciplinary, multidisciplinary, or new major science—its important feature was that its participants illuminated many problems now shared by physicians, life scientists, and engineers, and produced good proposals that are likely to be acted upon

Nilo Lindgren *Staff Writer*

In early September an international conference was held in Washington, D.C., to assess the rapidly changing and vigorous interaction between the engineering sciences and biology and medicine. Leaders from these fields discussed ways and means of broadening the scientific base for the biomedical field, expressed their dissatisfactions with present support mechanisms, explored the difficulties in bringing computers into the clinical and hospital situations, emphasized the need for training the right kind of people for biomedical endeavors, saw the emergence of the "whole systems approach" as a healthy symptom, predicted greater multidisciplinary teamwork, greater use of computers, and a far deeper impact of biomedical engineering in the years ahead. One major hindrance to more rapid advances was said to be not science-centered but society-centered, rooted in our expression of social values. This article highlights this unusual and unusually productive conference.

Looking down the long table, the physician hammered it out to the engineer. "I refuse to wake up one morning and read the newspaper headlines that 50 of my patients have died because a computer error has led to their being given the wrong medicine. I cannot let this happen! I *will not* let it happen! Reliability is the critical issue in the hospital!"

So it went in a vigorous dialogue between men of international standing who not many years ago operated at opposing ends of a spectrum of professions, and who today are concerned with the same problems. They were brought together under the aegis of the U.S. National Institutes of Health in a conference aimed at illuminating the "Future Goals of Engineering in Biology and Medicine." The rather unusual, two-day conference was held in Washington, D.C., in early September, less than two months ago, sponsored by The National Institute of General Medical Sciences, one of eight National Institutes of Health.

The conference was unusual on a number of counts: it brought together "men of substance"—physicians, psychologists, biologists, mathematicians, engineers—concerned with advancing both biomedical science and medical services; it was a relatively small "working" group that was able to maintain sustained serious discussions—participants remarked there had not been a conference quite like this in recent years; and, probably most important, the discussions on the problems that need to be solved in the interplay of engineering with biology and medicine will have a concrete effect on the formation of national programs to meet the goals foreseen. (See Dr. Dickson's remarks in the box on page 95.)

Broad conclusions of discussions

During the impressive and free discussions of two concurrent sessions, one on engineering in biology, the other on engineering in medicine, the numerous questions raised seemed to revolve around two related poles: What were the basic scientific problems and how were they to be solved? What were the most appropriate organizational mechanisms for supporting ongoing research and training new people for the biomedical engineering field?

Again and again, during the two days, there were calls for the establishment of rational national programs to meet and anticipate our health needs, to bring into being the instrumentation, the knowledge, the organizations, the manpower, needed to achieve a higher degree of medical services—in effect, the total systems approach was being advocated from many quarters of the biomedical engineering spectrum.

Biomedical engineering, it was said, has evolved to the stage, and already so proved its effectiveness, where the relatively helter-skelter arrangements that have gone toward its support up till now no longer suffice. A rising population, and rising expectations within that population for better medical services, dictate that appropriate steps be taken now to meet these enlarged future needs. Also, although inspired individuals, generally housed within traditional academic institutions, have brought the basic scientific aspects of biomedical engineering to its present pitch of activity, the scope of the theoretical and experimental activities that now seem obligatory will require new levels of support. The problems in the application of engineering in biology and medicine are of such a scale that neither the universities, largely dedicated to teaching and research, nor industry seems ready or willing to take leadership; so it is falling to the Federal government to take a larger hand through such agencies as the NIH.

Even at that, there is a tight competition for funds. By contrast, the Vietnam war is costing billions, whereas biomedical engineering is nibbling with only millions. Total NIH funding in fiscal year 1966 was $35 million. In view of the looming medical problems, to say nothing of "other" needs, such allocations of our social resources may not be admired as a pinnacle of rationality.

Advancing the science

The technical areas discussed in the session on engineering in biology included biological control systems, neurophysiology, electrophysiology, physical properties and mechanical behavior of biological systems, instrumentation for biophysical research, computers, biomathematics, and biomaterials.

Despite differences, there seemed to be a clear consensus that the future of biomedical engineering depends on the evolution of a sound theoretical base, that it is necessary to determine the specific problems to be solved, so as to establish a system for solving those problems. Considerable discussion centered on the mechanisms for housing basic research, as to whether or not special institutes, perhaps within a quasi-university setting, should be set up. There was broad agreement that there exists a basic problem in training the right kind of people for the biomedical field, people who can think and speak in both engineering and biological languages with equal ease; this raised the question of whether or not biomedical engineering is truly a science in its own right, a problem which in turn reflects on how university departments are presently structured. Said Dr. Warren McCulloch of M.I.T., "What I find everywhere is that it is the engineering schools that are willing to open their doors to biology, whereas the life sciences are hopelessly department-ridden." (See also Dr. Bartholomay's remarks on page 95.) Professor Otto Schmitt of the University of Minnesota quickly agreed. It appeared that the conferees generally agreed that during this "transitional period" most academic institutions are not facing honestly the changing scene in the cross-disciplines; nor, it was said, are the government agencies that have been sponsoring biomedical research. There was talk, too, of the founding of a National Institute of Biomedical Engineering, on a par with the other great national institutes, a concept which, if it were to be realized, would enormously forward the "identity" of the field. There was concern, also, that too much attention was being focused on such structural and organizational questions. Appealing for more of a question or problem orientation in biomedical engineering, Dr. John Moore of Duke University Medical Center said, "The quantitative approach is here . . . let's use it! Let us consider the problems to be solved, and take whatever tools are available to us." Most conferees agreed that during such a transition period, when the biomedical engineering field was pulling itself up to the status of being an independent science, it was difficult for students to get information and training from other disciplines, to extend their knowledge. And so it went, a kind of lively but contained, gentlemanly but impassioned, free-for-all among men who were patently in accord that biomedical engineering had made an important impact in the traditional fields from which it had sprung, and that this impact was but a pale forerunner of what was yet to come.

Advancing the practice of medicine

The engineering-in-medicine sessions were equally wide ranging, covering instrumentation, biomechanics, orthotic/prosthetic systems, automation of clinical laboratories, the use of computers in medicine, patient monitoring and intensive care units, hospital information systems, and the broad development of health systems. If a distinction must be made, it should be said that whereas in the biological realm researchers are most concerned with the ways and means of expanding their theoretical base of understanding of living systems, those on the medical side are most concerned with applications of basic insights. It was said, "Discoveries do not apply themselves—equal money and effort are needed to close the gap between a medical discovery and its application to health services. The great task of research and development to adapt modern technology and automation to the delivery of health services has just begun." Are the problems so immense? Clearly so. As it was almost plaintively noted by one physician, it usually takes almost 25 years, a human generation, for a new scientific instrument to get into medical use. How can this kind of time lag be shortened in a period in which the computer will have an impact as great on medicine as that of the microscope several centuries ago? These questions must be measured against the realities of medical practice. It may shock our complacency, habituated as we are to well-financed and well-organized R&D laboratories, to consider two nodal points of medical practice: at one node is the traditional practice of medicine as a "cottage industry," and at the other node is the hospital, of which one doctor remarked, "Reading about the conditions in hospitals is like reading a history of the Dark Ages!" From the earliest times, hospitals, as Dr. Erling Dessau of Datacentralen, Copenhagen, quoted, "have been highly individualistic institutions, the product of highly individualistic individuals and groups. Dealing as it does with the individual from birth to death, the hospital is like no other product of human society."

Enter the computer

Into that unique institution comes the new hero-demon, the computer. Problems? Yes. With reliability, inadequate software, and so on. Nonetheless, a "total systems approach," like that in defense and space, is

Input to the NAE

NIH has recently contracted with the National Academy of Engineering for the establishment of a committee on the interplay of engineering with biology and medicine. The committee's broad purpose will be to delineate the characteristics of modern engineering and its relevance to biomedical research, to the development of instruments, devices, and systems useful in medical practice and to the delivery of health services. One of the major areas in which this committee will examine the interaction of engineering with biology and medicine will be the utilization of engineering concepts and industrial technology for the development of instrumentation, materials, diagnostic and therapeutic devices, artificial organs, and other constructs that relate to the solution of major problems in the area of biology and medicine. It is likely, then, that the output of this conference will constitute an important input to the deliberations of this new committee that will shortly be formed by NAE.

Dr. James F. Dickson III
Director, Engineering in Biology and Medicine
National Institute of General Medical Sciences, NIH

Homes for biomathematics

I am afraid that even more than literal distinctions may have to be made in a new cross-field such as biomathematics in order to protect the field from the unavoidable onslaughts of those "non-Generalists" in adjacent fields who are really content with the pre-existing barriers and who see the creation of bridges over the barriers as something to be attacked with great ferocity.

We must recognize the fact of nature that new kinds of faculty members and new kinds of courses generated by existing programs pose threats to the security of members of established departments and some administrators because they push beyond those comfortable barriers. Hence, it is up to established biomathematicians and bioengineers in a university setting to do all that they can to advance the field both scientifically and practically. It is encouraging to note that a few departments of biomathematics have already been created. But many more are needed, particularly in the older universities, usually the least interested in new departments.

Dr. Anthony F. Bartholomay
Biomathematics Program
University of North Carolina

seen as the answer to the universal automation of hospital record-keeping, and the proper aim of an "international task force."

At the present time, automation of clinical laboratories (traditionally providing the chemical, hematological, microbiological, tissue analytical, and blood banking services), and of hospital and patient-monitoring services, is not moving as fast as it should. And this is not just because physicians fear automated equipment. This idea was debunked as a myth by Dr. Cesar A. Caceres, chief of the Medical Systems Development Laboratory, U.S. Public Health Service. What the doctor resists, Dr. Caceres said, is an array of unrelated instruments that can be more trouble than they are worth. If physicians are offered a good system, he went on, really useful in improving patient care, you will find them quite willing to try it.

There was the problem, too, of whether or not computerized systems, being developed now for large hospitals, would be suitable for small hospitals. This question led to the problem of standardization of hospital information systems. Dr. Murray Eden of M.I.T. warned that the time is coming soon when arbitrary decisions will need to be made on standardizing hospital information systems, in the same way that arbitrary-but-binding engineering decisions had been made in the past for such things as telephone dialing (in which the zero, the most used number, created the largest time delay in transmitting calls) and right-hand- or left-hand-side auto driving. Such arbitrary, but necessary, decisions could lead to costly mistakes in the formation of the coming hospital systems unless appropriate studies are made early enough in the decision process.

Even without the question of computer automation of medical activities, there are crucial basic questions as to the determination of the "essential variables" that are needed in basic medical science and clinical investigations. The proper selection of essential variables has been going slowly, said Dr. Rushmer of the University of Washington School of Medicine, over the whole range from single physiological items to whole systems. In fact, most present instruments for measuring human variables have not been sensitive enough to distinguish the changes that occur in the individual "normal" person, so sensitive is the human physiological control system. What will be needed, evidently, will be the measurement of multiple parameters and the use of computer integration techniques to single out significant physiological changes.

A focus for action

These are but a few of the problems that were illuminated during the medically oriented discussions. Dr. James F. Dickson III of NIH, chairman of this session, summarized that it would have "a genuine utility in the area of medical engineering development, as it is likely that there will shortly evolve a focus for action in this area that will initially concern itself with planning—with contracting, with preproduction laboratories of one form or another for the development and testing of prototype instrumentations and systems—and with contracting with industry for the manufacture of the prototype instruments and systems developed. Because of the current constraints on funding in this area, the evolution of this activity will likely be quite gradual. However, in time, such a directed research and development effort in medical engineering could become a considerable engineering enterprise in itself."

But even more significant was Dr. Dickson's view that the hindrances to moving ahead in these biomedical engineering areas could be viewed as being far more man- and society-limited than science-centered. The deepest obstacles of all have their roots in the way in which our society expresses its values through social, political, and governmental action.

So much for the "buckshot" summary. What follows are condensations of a few papers that may suggest with a bit more depth the fertility of the Washington discussions.

Special computers for life-science research

The potential power of new computer systems and languages in life-science problems is revealed in Prof. Gilbert D. McCann's description of a system being evolved at the California Institute of Technology. It is worthwhile following McCann's description at some length because this expandable language system is a beautiful new example of how the technology with which many of us are most familiar may be shaped to serve the special problems of the biomedical sciences.

Special systems must be developed because life-science problems differ from those in the physical sciences. Although one can subdivide problems in the physical sciences, it is a mistake to attempt to subdivide too much in the life sciences. McCann contrasted the two approaches. The vast and precisely organized knowledge of the physical sciences, he pointed out, was the result of a long series of relatively simple but accurately correlated experiments. The reason these simple experiments could be correlated and expanded into larger conceptual chunks is because they dealt with pieces of essentially static relationships. Not so for the life-science investigator. Regardless of how simple he tries to make his experiment, says McCann, he is always acutely aware of many more informational relationships and variables that he is not recording accurately or assessing in relation to the factors he is trying to consider. Furthermore, living systems are usually in a process of continual change, frequently as a result of the experimental stimuli. Thus, the process of correlating a chain of relatively simple studies is not fruitful. Many of the important unanswered questions regarding complex living systems require the ability to conduct much more complex experiments that permit the simultaneous correlation of vastly greater amounts of data than has hitherto been possible.

Thus, the CalTech approach aims at surmounting these difficulties through the appropriate application of new concepts from the information sciences and the technology of our steadily evolving high-speed digital computers. We want, McCann says, the ability (1) to acquire simultaneously much larger amounts of information, (2) to examine and characterize more complex informational relationships rapidly, and (3) to formalize by new methods of mathematical analysis the fundamental properties of living systems as interrelated over several levels of system complexity.

To achieve these aims, which require a rich language and sophisticated data memory structures, McCann and his associates have relied on recent developments in mathematical linguistics to provide the foundation for their computer system concept. Their first system, called the Phase II Experimental Data Language System, is now working. It employs an IBM 360/44 and a 360/50 computer, together with an interface computer developed at CalTech. Figure 1 shows how the system provides for a highly interactive combination of the experimental environment, the human investigator, and the computer. The investigator has at his disposal rich, expandable languages for discoursing with the computer. The communications interface consists of an IBM 2250 digital cathode-ray display console augmented by a "functions" keyboard as shown in Fig. 2. The experimenter controls both the experiment and its subsequent analysis with the light pen, function keyboard, and typewriter.

McCann describes the principal characteristics of the system as follows. Its capabilities reside in a family of primitive analysis and display functions, and in the process handler that services them. Each primitive function is accessible through the function keyboard. Sequentially pressing the appropriate keys activates the corresponding series of functions that are thereby applied to the data. Thus, the experimenter is given console programming of his data analysis and display. The input and output structure of any function is displayed on the CRT when the experimenter presses the function key, whereupon he can select data and parameter values to which he wants the function applied, and the names by which he wishes the output identified. This allows him later to prescribe simply by name the cross-linking and composition of various functions sequentially.

The full power of the system is available to the researcher on a real-time basis during the experimental period of data gathering. While the data he has selected and named are flowing automatically into data storage he can apply any of a large variety of analysis and display functions, monitoring now this, now that aspect of the experiment, controlling parameters of the experiment from the console, cross-linking and classifying, during the actual experiment. Upon completion, when the data are all in, he can re-examine, analyze, reduce, and manipulate his data at will, all from the same console.

One good example of the rapid feedback interaction that can be achieved in combined stimulus control and data analysis, McCann says, is illustrated by EYEMAP, a program that has been used in studies of the nervous system of the common housefly. Figure 3 shows the CRT format: in the upper array of dots, rows represent horizontal angles of the spherical visual field of the stimulus screen in Fig. 1; columns represent vertical angles. The system allows rapid mapping of visual properties.

The significant use to which this kind of computer system can be applied has to do with the mathematical formulation of the really sophisticated processes of living nervous systems. The fact that this system involves expandable languages is important (new languages as well as classical axiomatic languages will be available); these languages are richer and more suitable to the thought processes of human beings.

Development of health systems

The broad view of how biomedical engineering could contribute importantly to the evolution of better U.S. national health services, and the obstacles still in the way of achieving those services, was put forward cogently by Dr. James F. Dickson III of NIH.

Applications of modern engineering technology that

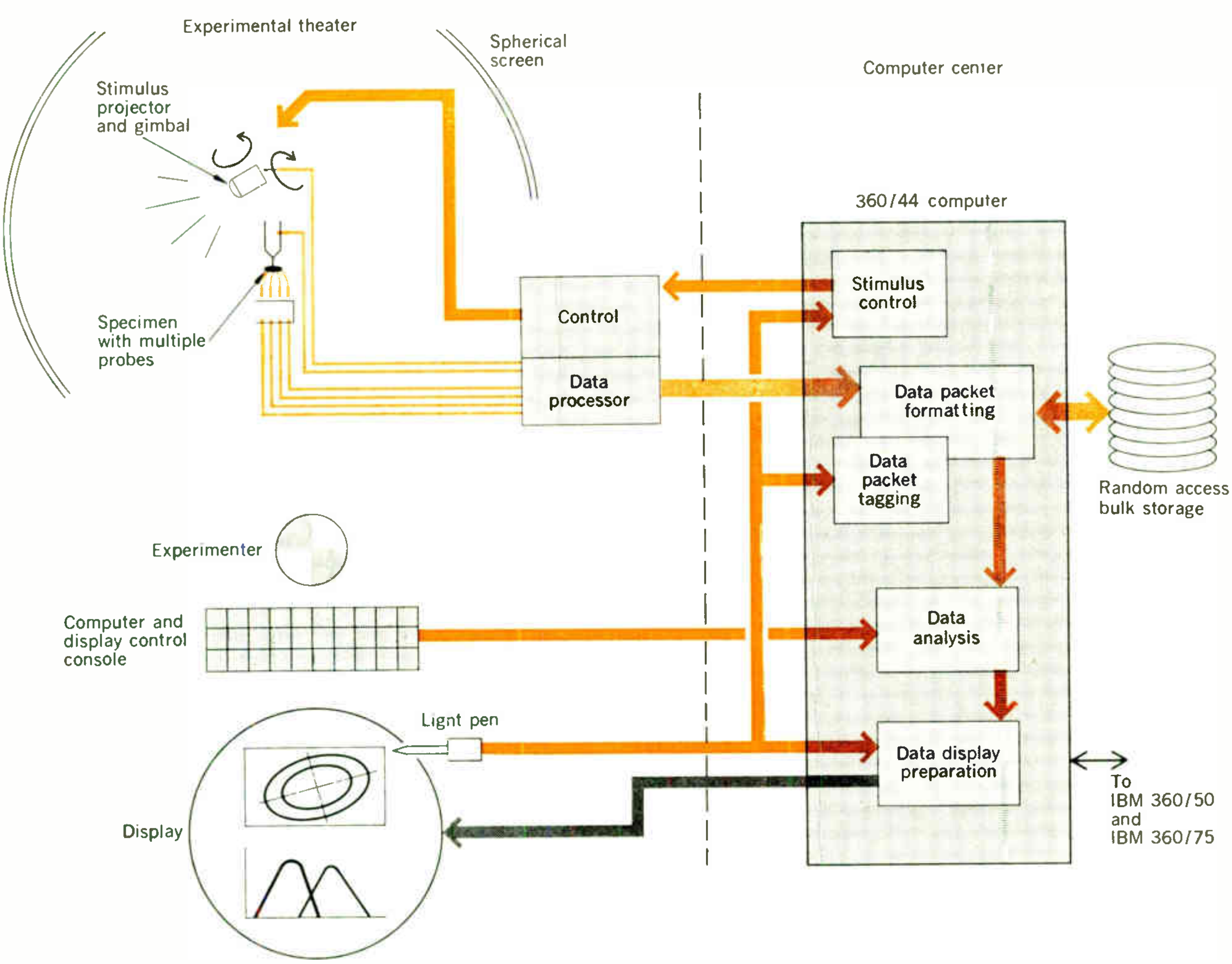


FIGURE 1. Expandable data-language system developed at the California Institute of Technology as applied to visual nervous system research.

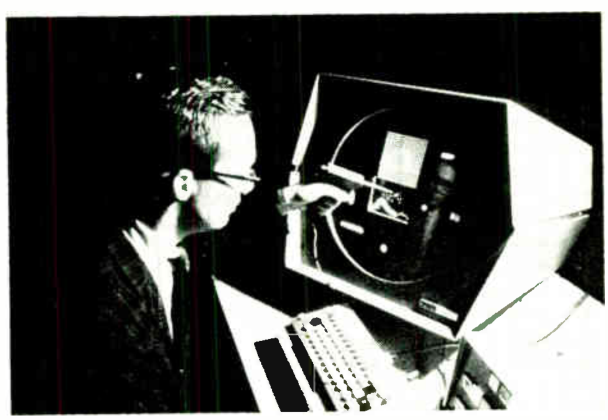


FIGURE 2. The digital cathode-ray display console allows experimental operations on many different types of function formats.

will be of special importance to health services, Dickson said, will be those in instrumentation, automation, computation, communications, and systems engineering that emphasize techniques and systems for diagnosis, treatment, patient care, and the storage and analysis of health data. Pertinent examples relative to these include the automation of clinical laboratories, specific computer-oriented diagnostic testing, computer assistance in differential diagnosis, intensive care units, patient monitoring, hospital information systems, repositories for health data, facilities for screening for disease detection, and local, regional, and network facilities for incorporating some of these techniques and systems.

In the near future, Dickson says, because of the increased participation of Federal programs in providing essential health services, a sharp increase in the demand for clinical laboratories services can be anticipated. A program of directed research and development for automated laboratory technology and electronic data handling can raise the capabilities of clinical laboratories to high levels of efficiency so that (1) the increasing demands can be met, (2) it will be possible and feasible to use these laboratories for the diagnosis and treatment of disease in ways unheard of but a few years ago, and (3) possibilities will be opened up for bringing the full benefits of modern medical laboratory science to the entire population. Accordingly, the automation of clinical laboratories must be considered high on the list of targeted areas for further research and development in the light of the potential likelihood of payoff in the next five to seven years if substantial amounts of money are invested. In the development of this area, demonstration trials must be arranged, not only for the evaluation of the instrument systems per se but also for the purpose of obtaining feedback on performance under prolonged working conditions.

In reviewing industry's activity in the applied biomedical engineering area, Dickson noted that attempts to put off-the-shelf items into service in the health area have often not met with success. It is increasingly apparent, he said, that there is a need to turn industry's capabilities and potentials away from off-the-shelf items to more imaginative developmental work in targeted areas pertinent to the delivery of health systems.

However, it is important to note, Dickson said, that there does not exist at this time an adequate biomedical research base for the development of some of the larger projects that are necessary. For example:

1. It is important, if computer activities are to become more practical, that a more natural language input be developed before a truly across-the-board effectiveness can be realized for health systems.

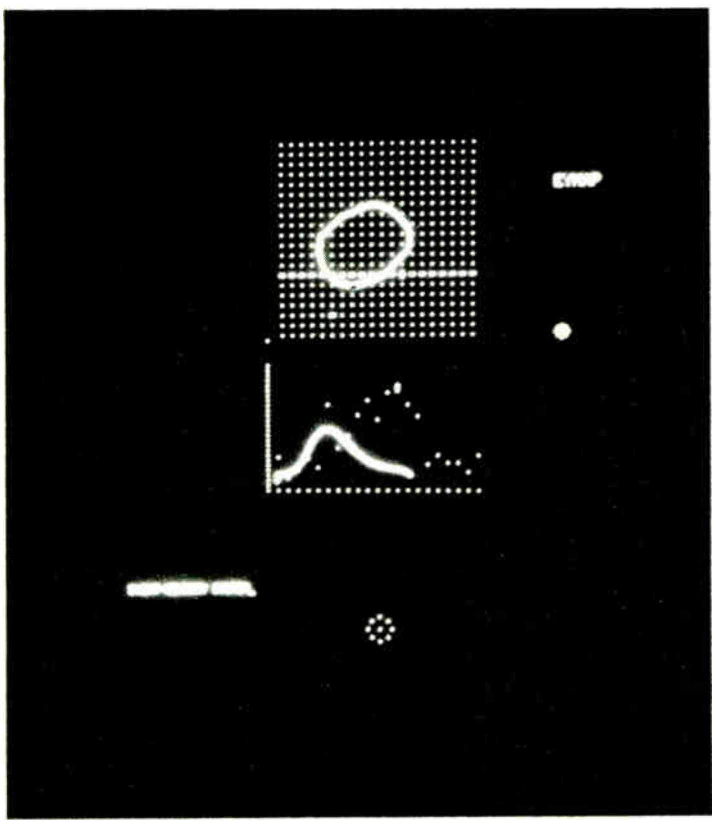
2. There is a need for considerable assistance from mathematicians in developing new techniques, new approaches, and new symbolic representations that are pertinent to problems in the life-sciences area.

3. It is evident that basic knowledge in the area of physiological systems is not deep enough, nor broad enough, to set up patient monitoring systems or intensive care units that will have the desired sophistication.

Nevertheless, the application of the engineering sciences through directed research and development to the applied health sciences and the delivery of health services will be of increasing significance. Given an applied goal in engineering, there may often be nothing but money that stands in the way of reaching the goal, provided basic science has shown the goal to be achievable.

Computers, of course, Dickson goes on, must be counted on to play a key role in the development of adequate health systems. However, an overview of the computer activities that are supported about the United States reveals that there are considerable areas of overlap and that some important areas have not been attended to at all. In short, activity in the area of computer studies must be integrated into a more meaningful whole as regards basic research, applied research, and the delivery of health services.

FIGURE 3. Cathode-ray-tube format used in the EYEMAP operation, which allows rapid scanning of the visual fields of interneurons in a visual nervous system.



Man–machine orthotic–prosthetic systems

One of the major areas in which engineering is being applied in medicine is the development of orthotic and prosthetic systems. Orthoses are used to supplement weakened human limbs and functions, whereas prostheses entirely replace a missing part of the human body with an artificial device that can perform at least part of the function of the missing part. Although many types of orthotic/prosthetic (O/P) systems have been devised over the years, one of the persistent problems is to design systems that will be acceptable to their human users. Too many devices that have already been constructed, it is said, lie abandoned in closets and dark corners.

In their review of the O/P field, Profs. James B. Reswick and Lojze Vodovnik of the Case Institute of Technology noted that the number of people who could be materially helped if highly developed O/P systems were available was not generally appreciated. Considering only the broad areas of disability of human limbs or extremities caused by crippling diseases, neurological disorders and congenital effects, and accidental and war disabilities (i.e., excluding such potentially large fields as artificial organs, sensory aids, and so on), it is estimated that as many as 2½ million persons could conceivably benefit markedly from the use of orthotic devices.

However, one of the major problems in bringing about such assistance is that the need has not been precisely defined. Many persons suffering neurological disorders are not recorded in hospital statistics, and, if they are, the nature of their disability is not. A real need exists to codify the specific types and numbers of disabilities and to do this in a way that would lead to the development of engineering specifications from which decisions on priorities of effort and specific engineering designs could be made.

Another general problem, Reswick and Vodovnik maintain, is *evaluation* of O/P systems, a function that is isolated from ongoing development. The theory that a prototype can be developed by one group, evaluated by another, and manufactured by a third does not work out in practice because O/P problems are more complicated than appears at first sight. Reswick and Vodovnik stress that a constant interplay between patient needs, physician requirements, and the engineer's design work must be maintained. They urge an overall systems management function for all O/P activities in the U.S., vested with Federal agencies and their advising groups, comparable to NASA's space management.

An even broader problem in determining priorities of effort was brought forward quite forcefully by Dr. Bertil Jacobson of Stockholm's Karolinska Institutet. He pointed out that while in the past impressive progress has been made in the application of instrumental technology to diagnostic and therapeutic medicine, preventive medicine has derived far less benefit. An example of the lack of rational planning in the bioengineering field cited by Jacobson is in research on atherosclerosis. It is the most common and important of all destructive diseases, despite the enormous investments of time and effort in cardiovascular research. In Jacobson's view, atherosclerosis is an outcome of the fact that the human species has not yet had time to adapt itself to the rapid rise in the standard of living that we have witnessed over the last few centuries. This period has seen a sudden change in the biological environment, including food habits and a disproportion between mental stress and physical effort—the first cardiovascular sign of which is aortic atheromatosis. This condition may occur early in life—it has been found in an eight-year-old child, and by age 20 probably all persons are affected by it to a greater or lesser extent. And yet, the glamour associated with the development of an aortic prosthesis to replace a hopelessly damaged artery after a further 20 to 40 years' deterioration is, says Jacobson, unfortunately greater than that afforded by attempts to detect, measure, and prevent the first manifestations of the disease. If only we can change our goals and redirect our efforts, he argues, we may justifiably look forward to the time when we shall no longer be concerned with designing replacements or executing repairs.

New O/P directions

Most practical thinking about O/P systems to date has necessarily dealt with various kinds of mechanical–electrical contrivances of varying degrees of sophistication. Such O/P systems have grown out of the tradition of examining only the gross physiological and behavioral elements of functional loss. But, more recently, as Dr. John Lyman of U.C.L.A.'s Department of Engineering points out, attention has been directed to the events of microstructure that may be practically related to the observable gross behavior. One consequence of an understanding of the microbehavior of nerve–muscle as a system would be the possibility of utilizing such behavior for both structural and control purposes. The dream of direct communication, Lyman says, by means of the sensory-motor information-carrying capacities of nerve stumps, may indeed be realizable in man–machine applications as well as with advanced surgical concepts such as transplants from nerve, muscle, bone, and even whole limb banks. A clear goal for biomedical engineering in this area, he goes on, would be an examination in depth of the potentially useful interrelationships between the levels of structure and function as they pertain to human skills.

But even beyond such goals, Lyman speculates, there is the possibility that our accumulating basic knowledge might ultimately permit the regrowth of faulted human subsystems through man-controlled manipulation of events at the subcellular level.

Artificial internal organs

Two noted physicians whose names have become almost synonomous with investigations of artificial internal organs—Dr. Michael E. DeBakey of the Baylor University College of Medicine and Dr. Willem J. Kolff, formerly of the Cleveland Clinic and now at the University of Utah College of Medicine—appraised recent developments in artificial hearts and artificial kidneys.

The single most important impediment in the development of artificial hearts is still, as it has been for some time, the materials problem—the need to find a material that will serve as an interface with body tissue. Until this tissue interface problem is solved, Dr. DeBakey said, the whole approach to artificial organs will be impeded. Power sources and control mechanisms also present formidable problems, but in his view the interface problem must be solved first. In the case of artificial hearts and cardiac assist devices, the need is for materials that will not cause blood clotting. Dr. DeBakey reviewed the materials that his Baylor group has been using—Dacron, Teflon, Silastic—and discussed a new velour technique in which finely spun Dacron fibers are used to encase heart elements in contact with the blood. It has been found that the body forms a natural tissue over the velour so that no blood clotting seems to occur. Figure 4 shows a heart valve coated with velour. Remarking on this new technique, Dr. Willem Kolff pointed out that the results are very encouraging, but there is as yet no proof that it will serve on *flexing surfaces* for longer than a few weeks.

Dr. Kolff's own concerns about artificial kidney developments were of a different order, namely, with the high cost and complexity of such machines. He acidly noted that although the first artificial kidney was invented in 1913, and one was used successfully for the first time 30 years after that, now, 24 years later still, there are in the United States only about 600 patients who use maintenance dialysis. The initial cost for each of these patients is $10 000. Meanwhile, 45 000 people die of renal disease each year. In exasperation at certain industrial "dillydallying" over the development of a new capillary kidney for the past five or six years, Kolff and his colleagues tried out dialysis using an ordinary washing machine whose initial cost was less than $350. Kolff admitted that although such a crude do-it-yourself machine might kill half of the kidney patients, it would at least save half of those who would die otherwise. He cited this washing machine experiment as an example of the kinds of devices that biomedical engineers should aim for. Ingenuity and simplicity in design, he said, are needed much more than complex automatic devices.

On the other hand, it was noted during the subsequent discussions, many engineers were apparently not attracted to biomedical engineering work because the types of devices needed might seem too trivial to engineers. The response to that was a call to change the "pecking order," so as to give biomedical engineers more prestige among their nonmedically oriented peers, and to give them greater remuneration as well. Dr. Reswick opined that it would be hard to change the pecking order in the near future, and that it might be necessary to build biomedical organizations to operate outside the usual academic structure, and thus "run around end."

Better support mechanisms?

The question about appropriate recognition of biomedical engineering, and about whether or not it is an entirely new science, ran like a recurrent thread through the discussions in both concurrent sessions. It is clear that many of these leading biomedical engineers and scientists feel that their ongoing work is seriously impeded by this lack of recognition in the universities, in government, and even among themselves.

Many conferees, most notably Dr. W. R. Adey of U.C.L.A., raised strong questions about present government support mechanisms. It is perhaps a symptom of the present healthy posture of the biomedical field, and a reflection of its present rate of growth, that the funding mechanisms that prevail to date have aroused rather strong dissatisfactions. The kinds of enlarged multidisciplinary teams now envisioned by the substantial leaders of this field would place very large new demands for both more money and a more flexible framework for funneling it to where it will do the most good.

A programmatic challenge

The road to rational biomedical engineering programs is littered with the debris of many arguments and, as Prof. Otto Schmitt of the University of Minnesota called them, "position statements." Schmitt's point was that there are far too many (albeit comforting) position statements. Now it is time, he said, to come to the point of proposing specific individual and scientific-community undertakings from which a set should be selected and actually carried to realization.

It is now time for biomedical engineering, he said, to begin accepting the full responsibility for developing theory and instrumentation specially adapted for the biomedical field, not simply begged or borrowed from other fields. While keeping alert to useful intellectual and technical developments in adjacent fields, it should be, he continued, our responsibility to discover electrophysiological models that fit our observed facts and to develop and see put to enlightened use, instruments and computers evolved for life-science use. Particularly exciting areas of interest are the development of appropriate interpenetrating field theory models, biocomputers, and systems analysis by articulable blocks of functional transfer, both discrete and continuous. There also is challenge in extension of system optimization and information theory to a closed biosystems form in which cost-productivity trade-offs that can be evaluated quantitatively can be made.

I would like, Professor Schmitt went on, to suggest some specific endeavors that we could undertake if we are willing to believe that we, as bioengineers, biophysicists, and biomathematicians, are the initiators of a new major science, not merely hangers-on to a bandwagon.

He said that he would like to challenge each conferee to invent and have on hand for development:

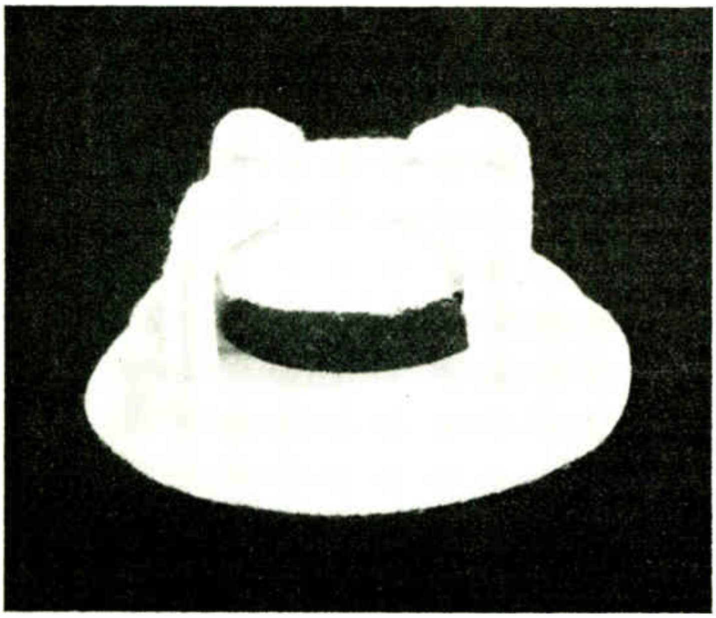
Some one model in biomedical science that seriously implants into the life sciences an algorithm that is usefully manipulable by analytic or machine methods.

A systems problem that suggests a different management of our academic, administrative, industrial, or technical problems (e.g., use of an imagination matrix).

One major and one minor project that is important biomedically to the community.

Dr. Schmitt then offered examples of tasks applicable to electrophysiology that he thought might be generally extended to the other biophysical sciences. These included the development of a core list of tasks pruned from a larger accumulated list, a rational assignment of activities, the building of a modular data base on which a biomedical science could grow adaptively, establishment of a training program for biophysical science technicians, the creation of an umbrella organization to serve the emerging quantitative biomedical sciences, and so on.

FIGURE 4. Artificial heart valve encased in velour. Dr. Michael E. DeBakey says that velour technique sharply reduces blood clotting, which is still a major problem in artificial heart devices.



Talking together

It has already been observed how the limitations of the mathematics available from the physical sciences, and the lack of a formal language to describe psychic activity, life processes, etc., constitute at present a serious roadblock to solving life-science problems. There has been, in addition, another language problem afflicting biomedical engineering endeavors. This has been manifested at a different level—at that of the social interaction and cooperation of investigators from different disciplines. For the past few years, one has heard quite a lot of discussion and complaint about this language barrier, this barrier to understanding.

That such difficulties are being bridged is evidenced by meetings such as the one in Washington. As the noted Pierre Rijlant of the University of Brussels Faculty of Medicine put it, he was heartened by this success in attaining mutual understanding, "where mathematicians, biologists, engineers and doctors live in peace around the same table." Waxing eloquent, while referring to the sometimes impassioned discussions about whether or not old disciplines were now giving way and being supplanted by new ones, Dr. Rijlant went on, "speaking as a physiologist, I am glad to see that the old dying tree is sprouting strong new branches." On more or less that warm note, the conference concluded.

Note: Proceedings of the conference will be published in book form.

IEEE educational activities—an ad hoc committee report

The Board of Directors of the IEEE, at its meeting of August 22, 1967, at WESCON, approved the creation of an Educational Activities Board, to have a status comparable to the three other major Boards of the Institute: Publication Activities, Technical Activities, and Awards. This action represents a reorganization of major importance to the IEEE in improving the effectiveness of its educational activities and furthering its recently launched program of continuing education services for the membership. The action came as a result of a year-long study by the Ad Hoc Committee on IEEE Educational Activities,* culminating in a report to the Board of Directors in August. Because the results and recommendations of the study have important implications for all members of the Institute, we are taking the unusual step of reproducing below the Ad Hoc Committee report in toto.

—The Editor

In August 1966 the IEEE Board of Directors, acting upon the recommendation of its Executive Committee, voted to set aside in the 1967 budget the amount of $45 000 for the purposes of "continuing education." Subsequently, an *ad hoc* committee was appointed by the Executive Committee, with the IEEE Secretary serving as chairman, to study not just the question of how this money might be spent most effectively, but also the broader question of what IEEE's total role should be in carrying out its constitutional objectives as a "scientific and educational organization." Specifically, the charge to the committee was: "to define, in the light of the IEEE Board of Directors' action funding a program of continuing education, the types of educational activity in which the IEEE, as a society, should engage, and to recommend to the Executive Committee the appropriate organizational structure, bylaw changes, and policy statements needed to implement those activities."

The committee has not been able, in the time available, to make a completely detailed set of recommendations, nor does it believe this would, in fact, be desirable at this early stage. It has, however, identified a number of areas in which IEEE could better serve its members, and promote new memberships, by strengthening existing educational activities and by initiating new ones. These fall into the broad categories of:

1. Continuing education for IEEE members
2. Assistance to educational institutions
3. Strengthening the support of Student Branches
4. Career guidance for high school students

Continuing education

In a narrow sense, continuing education may be defined as any process by which a person, having been brought abreast of current knowledge in his chosen field through traditional college and graduate curricula, continues to keep his knowledge up to date as new developments in the field occur. Continuing education, in this narrow sense, is not a new role for IEEE; indeed it is a basic function, as it is of most professional societies.

IEEE technical publications report the most significant research developments and engineering progress of our profession. The PROCEEDINGS OF THE IEEE is (or must be

*Members of the Committee: Bernard M. Oliver; Edward E. David; John B. Little; Roger E. Nolte; M. P. Ristenbatt; George Sinclair; Herbert Trotter, Jr.; Ernest G. Walters; Charles H. Weaver; F. Karl Willenbrock.

made) invaluable to the research-oriented member—the member working at the science–engineering interface where new discoveries are being made and harnessed for use. The TRANSACTIONS and JOURNALS report with fuller coverage the details of technical progress in particular fields. These are the IEEE journals for the creative specialist.

Technical conferences also serve the needs of these members for continuing education in this narrow sense. They are oral, live counterparts of the PROCEEDINGS and TRANSACTIONS, and provide the opportunity for verbal exchange of information—for questions, answers, and discussions among co-workers in a field.

The committee feels that although a reduction in the number of Groups (by combining closely related Groups into a single Group with expanded scope) would provide a wider dissemination of knowledge, with fewer TRANSACTIONS and fewer symposia, the basic structure of both the IEEE publications and meetings is good. The member who is continuously working in the vanguard of progress in his field—the member who is not only a reader of published papers but an author as well, the member who is not only an attendee of conferences but often a speaker as well—this member is well served by the present IEEE format of publications and conferences.

What, then, is the problem? If we provide more continuing education functions, aren't we scratching a place that doesn't itch? The committee believes there is a problem, and that we should add some services, services that are different from those already provided.

In a broader sense, continuing education is simply the acquisition of any knowledge, on a deliberate and sustained basis, in any field whatever, subsequent to one's formal training. It is in this broader sense that IEEE has done very little in continuing education. The committee does not suggest that we should do so in fields far removed from those of our profession, but it does suggest that the archival reporting function of our journals and conferences is not sufficient to meet the continuing education needs of a large fraction, if not a majority, of our membership, even within our profession. In short, we believe *that the tutorial services of IEEE, at the present time, are inadequate.*

Many of our members are managers, administrators, or sales engineers. Their duties do not permit them to keep fully abreast of the technical aspects of their fields, and yet they need to know the basic facts and potentialities of technological developments in order to make correct decisions and in order to be effective at their jobs. Our members change jobs frequently, and more frequently are given new assignments—new responsibilities—that require them to "catch up" in a field that was formerly of little or no interest to them. The engineer, or the manager whose business is impacted by technological advances, must absorb the new technology or go under.

In addition to these categories, many of our members have either never had formal college training, or have discovered that training to be inadequate for their needs. These people need training in what are now considered basic subject areas in order to realize their latent creativity. Inventive technicians who make significant contributions in spite of educational handicaps are legion. We believe IEEE can help, and has an obligation to help, these members as well as the more experienced researcher.

The IEEE is unique in respect to the resources on which it can draw in organizing programs of continuing education. It has access to the services of the most competent engineers in our field, regardless of the affiliations or loyalties of the individuals involved. Other organizations are generally hampered in this respect, due to problems arising from competition between industrial laboratories, or to lack of sufficient prestige to secure the desired cooperation in presenting the latest developments.

The whole thrust of any IEEE tutorial activity should supplement, and not compete with, the activities of our colleges and universities. Indeed, we should utilize the schools' talents, and expect their cooperation, in providing education for people whom they cannot reach but we can: busy people in the midst of their careers who are unable to spend time in graduate work; technicians who need basic courses to enable them to be more effective but who cannot afford the cost of college training, or cannot qualify for entrance, or both. These are the people who are inaccessible to the university in most cases. But we can and should help them.

We recognize that the IEEE does, in fact, at present provide some services that help satisfy the needs outlined above. IEEE SPECTRUM publishes many articles that help acquaint our members with developments outside their own special fields, and that introduce them to new fields. So do the special issues of PROCEEDINGS. Further, many of the symposia and local Section meetings present talks of a tutorial nature. On the other hand, the services provided by these media are uncoordinated. What happens in one Section is not available elsewhere. The articles in SPECTRUM report "news as it happens." An organized tutorial service should provide members, or groups of members, with the material they need, when the need is felt.

We conclude that the tutorial role of IEEE would best be served by providing additional services such as:

1. Published monographs and course material in new technological areas. To some extent such publications could be compendia of selected important papers republished in logical order, but in general they would require the preparation of additional tutorial chapters, problem sets, etc. For this service to be effective, IEEE must monitor the literature for emerging developments of great significance (e.g., the transistor, the laser, the integrated circuit) and anticipate the need for introductory courses and monographs. Such first volumes could be followed, when justified, with second and third volumes on the same subject. The availability of these monographs must be widely advertised; and their cost to members must be reasonable.

2. Lecture series. IEEE should prepare and distribute both live, video-taped, and slide–tape lecture series in new technological areas. These should be designed for both Section and Student Branch use, and for use both in conjunction with and independently of the written material recommended in item 1. To some extent, a "star system" would help the series to succeed, i.e., the use of lecturers who are recognized leaders in their field *and* who are good speakers. Although care should be taken in preparation to assure that the message gets through in spite of the limitations of the medium, money should not be wasted in making the tapes "slick." An occasional goof doesn't detract—it even adds reality. The main objective is to pack a high bit density of information per reel of tape.

3. Information on courses and materials. A great deal of material suitable for continuing education has already been written, and a great many courses already exist, that are available to IEEE members. However, objective appraisals of these materials and offerings are not always available to the member who most needs them; often he is not aware they exist. At the very least, IEEE should, through an appropriate editorial review board, disseminate information to its membership on currently available courses and texts in such basic subject areas as calculus, functions of a complex variable, transform theory, circuit analysis and synthesis, probability and statistics, and computer programming, as well as in new areas of interest such as coherent optics and integrated circuit technology. Reviews should, in good editorial fashion, indicate the level at which the material is treated, the prerequisites, and any special features of the courses and material. Such a guide would be of great help to the member in making an intelligent selection that will best meet his needs.

4. Workshops. The committee recognizes that learning is greatly improved if the audience is involved in an active way. Hence it would be desirable to couple all of the above suggestions to an active workshop program. Lecture series organized at IEEE Headquarters could be followed by or interspersed with workshop sessions covering the same material. These would involve assignment of problems, solving of these by the participants, and question-answer sessions.

The IEEE monographs and the guide to already available materials could form the basis for workshops organized by local Sections.

Eventually, such workshops might mature into regular courses of study, for which minimum standards could be set against which the participants' work would be judged. At this point, the offering of certificates of successful completion might be considered.

The committee believes strongly that the tutorial activities of IEEE can be made largely self-supporting. It feels that appropriate enrollment fees can and should be charged. Paid tutorial series at Section level have been very successful in the past. The payment of a "tuition fee" insures more serious involvement of the individual member, and guarantees his critical analysis of the services offered. The value of the programs can be directly measured by their subscription, whereas the amount of participation in a free service is a poor index. Finally, we feel that the expense of IEEE tutorial activities should not be allowed to diminish substantially the funds available for traditional services.

Assistance to educational institutions

IEEE is dependent upon our colleges and universities for the education of our future members. Our members are on many faculties, and derive some benefits from our archival publications and our conferences. However, as a society, we offer them little in the way of direct help in upgrading their curricula, or even in their efforts to recruit Student members for us. IEEE could have a much greater positive impact on the programs in our colleges and universities, to our mutual advantage, if we undertook to assist them in every way, not merely through ECPD, to evaluate them.

The committee recommends that:

1. The tutorial materials developed by IEEE for continuing education should be made available gratis, or at cost, to educational institutions for use at their discretion.

2. The IEEE should select and review educational films prepared by government and industry. This information, and in some cases the films as well, should be distributed for student use. These films often have continuing education value.

3. The IEEE should sponsor curriculum workshops on new topics that should be included in electrical engineering curricula, with participation both by educators and by outstanding R&D professionals in industry.

Support of Student Branches

A great many of our Student Branches feel like lonely, isolated outposts of IEEE. Students are encouraged to join; but as students they enjoy very little sense of participation in IEEE affairs—not because IEEE is disinterested in them, but because no effective organizational ties exist between the local Student Branches and IEEE Headquarters. It is true that many Sections do sponsor effective student programs. It is true that Sections and Regions do conduct prize paper contests. But it is also true that these often tend to be patronizing, and make the student feel like a freshman (or a gladiator performing for the judges) rather than a true member of the club.

The committee feels that:

1. All Student members should be considered members in good standing of the local Section in which their Branch is located. They should receive all Section publications and meeting notices. Appropriate additional rebates to the Sections should be provided to cover the expense.

2. IEEE Student Branch counselors should be ex officio members of the local Section committees. This is already the case in many Sections, and leads to greater student awareness of Section activities.

3. Student Branches should come under the purview of the Educational Activities Board (EAB); and the chairman of EAB should be the coordinator of Student Branches on the Executive Committee. He, in turn, should designate a member of the EAB as chairman of a permanent Student Branch Committee of the EAB.

4. The Student Branch Committee chairman should make all pertinent output of the EAB available to Student Branches, and should develop further mechanisms for support of Student Branch activity and for selective student participation in Institute events.

Career guidance for high school students

Studies show that most career decisions are made at secondary school level; at least the basic choice of a major field tends to be made there. This committee feels that the IEEE could be helpful to students in making these decisions, and to counselors and science teachers in advising the students. The object of any IEEE career guidance activity should not, of course, be to lure into engineering students whose talents lie elsewhere. Rather, it should be to insure that the student with talent in engineering or science is given an opportunity to develop interest by making him aware of the excitement our field offers.

The committee recommends that:

1. The IEEE should encourage all Sections to assist the secondary schools at a local level by providing speakers, arranging for field trips, securing gifts of equipment

from industry, etc. At present it appears that only about one fifth of our Sections engage in any sort of activity at the high school level.

2. The IEEE should prepare and circulate slide–tape presentations that acquaint students with our field. The cost of duplicating such material is not great, and such presentations can be effective, if well done. We note that the ASME has done this, and that some Groups, such as the Communication Technology Group, are planning such a package.

Secondary school curricula—set theory vs. calculus

Ever since Sputnik, the public schools of the United States have been in a turmoil, particularly with respect to their mathematics and science programs. Concern over inadequacies in our schools' curricula led to the hasty development and adoption of the "new math" (SMSG), "new physics" (PSSC), "new chemistry" (CHEM), and "new biology" (BSCS) courses. Although these courses have many good features, many educators believe that they have done more harm than good. Their high level of abstraction fails to relate science to the young student's world. The courses are often needlessly academic and rigorous, and they are difficult for students and teachers alike. The result is that student enrollment in mathematics and science is declining in the very years when we all hoped it would increase.

It is conceivable that IEEE should appoint a select study commission to look into the problem and develop an Institute position with respect to the desirable content of secondary school math and science curricula.

Organizational structure

If IEEE is to conduct significant educational activities as outlined above, there must be both staff support and continuing committee direction at Institute level. The committee makes the following recommendations:

1. The IEEE should establish an Educational Activities Board, having a status comparable to that of the Publications Board or to TAB, with the chairman of the EAB a member of the Executive Committee.

2. The EAB should be responsible for planning specific activities in continuing education, for monitoring ECPD accreditation activities and developing guidelines for our representatives on the ECPD accreditation committees, for developing IEEE relations with educational institutions, and for making policy recommendations to the Board of Directors.

3. Appropriate staff support and budget should be provided for the EAB.

4. The chairman of the Group on Education should be an ex officio member of the EAB.

5. The chairman of EAB should be also the Coordinator for Student Branches, and be responsible for the support of Student Branches as outlined above.

The EAB would supersede the present IEEE Education Committee. Members of EAB would be chosen for their interest and experience in education, among other qualifications, and so would often be members of the Group on Education as well. There would, however, be little or no functional overlap between EAB and the Group on Education. The former is conceived of as setting Institute policy and directing Institute activities in education, the latter as attending to the needs of its own members as professionals in education.

To be successful, the EAB should work closely with the Sections. In fact, many of the educational activities contemplated would depend for their success on active participation by the Sections. Local Sections are therefore urged to appoint Educational Activities Coordinators to their Section Committees. These coordinators would initiate educational activities and make available to the Section the programs of the EAB.

The EAB should keep abreast of the activities of other educational organizations, such as ASEE, and cooperate with them where useful, but it should not depend on these joint organizations for action that the IEEE needs to initiate now.

Word of caution

Although the committee believes all of its recommendations have merit, no plan for increased tutorial services can succeed without zeal on the part of those who plan and administer it, or motivation to study on the part of those for whom the services are intended. It is very easy to fall into a sort of "welfare syndrome" concerning education, and fallaciously assume that there is a great thirst for knowledge when, in fact, there may be none. It is therefore also recommended that IEEE proceed cautiously in providing these services. Before large-scale, costly programs are attempted, the market demand should be tested with a few smaller-scale, well-planned offerings. Only if these are received with enthusiasm should we escalate the program.

The initial phase

The committee hopes that this report will be considered an adequate fulfillment of its charge, and that the Board of Directors will implement the organizational changes recommended herein as a long-term plan of action. However, there exists some budget money to start certain activities this year; and there is a desire within this committee to make a start in continuing education, even prior to the existence of the EAB. The committee would appreciate authorization for the following immediate short-range programs:

1. Support of the initial costs of the tutorial series on "Integrated Circuits" and "Computer Aided Design" now being developed for the March Convention, and preparation of slide–tape presentations of the series.

2. Preparation of a review guide to existing educational films.

3. Evaluation of the microelectronics material from WESCON in order to make available slides and tapes if suitable.

4. Initiation of steps to have an existing series discussing semiconductor lasers and electroluminescence re-recorded (present tapes are bad) for distribution.

5. A request to TAB to develop suggestions for monographs in the fields of each Group.

6. Solicitation of speakers for a series on coherent optics.

7. Formation of plans to develop the series described in item 1 into a series of workshops.

8. Invitations to Sections to submit to the Ad Hoc Committee on Educational Activities, and later to the EAB, information on courses that have already been developed and that have proved successful.

We feel all the above actions can be accomplished within the existing budget, and request approval to do so.