VOLUME 2 ☐ NUMBER 4 APRIL 1965

IEEE spectrum

features

the cover

Focal point of the American Electric Power System Computer Center in Canton, Ohio, is this 390-foot-high microwave tower which channels coded information to and from the center over AEP's 1700-mile microwave network at the rate of 15 000 data characters per second. This system will be discussed in detail in Part II of the two-part series on computer-controlled power systems that begins on page 60 of this issue.

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

Spectral lines

Publication Objectives. In President Oliver's article entitled "The IEEE Looks Ahead" in this issue of SPECTRUM, he makes reference to the importance of publications to the Institute, and also to the need to improve our present publications. The President and the Editor are in complete agreement on both of these points.

The objectives of our publications are essentially the objectives of the Institute: to advance the profession and to advance the professional capabilities of the members. To accomplish the first objective, it is essential that our publications become the repository for the scientific and technical knowledge that relates to IEEE fields of interest. While some remarks in this article will be directed toward the second objective, it will be primarily concerned with the first; a future "Spectral Lines" presentation will be devoted to the overall effectiveness of the Institute in advancing the capabilities of its members.

Concentrating now on the objective of advancing the profession, it is worthwhile to enumerate some desiderata for IEEE technical publications.

First, each publication should cover a clearly recognizable technical area. Recognizability is important because it enables the reader to know whether or not he needs to monitor the publication to keep abreast of his specialized field, and it enables the author to know whether or not the publication is the appropriate one for his contribution. The author is very properly concerned that his paper be brought to the attention of the right audience. There should not be surprises in the subject content of technical publications.

Second, the technical integrity of our publications must be maintained. At times even the most carefully organized review procedure will fail to eliminate an unsound article. However, it is certainly better to err on the side of occasionally publishing misinformation rather than to refuse to publish material that is sound. But a publication's reputation and, consequently, its ability to attract the best papers in a field, depends on the infrequency of its mistakes. The editor and his reviewing staff are performing a significant service to the profession by maintaining a high standard of integrity. A major difference between the publications of a professional society and some of the no-cost technical publications is—or should be—that the reader can depend on the soundness of the content of the professional society's publications. In maintaining the integrity of his publication, the editor is not only serving his contemporaries but, since these publications are kept in the libraries and thus become a significant part of the literature of the period, he is performing a significant service for posterity.

Third, the IEEE should publish the minimum number of publications that will meet the needs of the profession. The IEEE is presently committed to a decentralized form of organization; as a result, it is committed to a relatively large number of publications covering rather specific technical areas rather than a few publications covering wide areas. Therefore, these publications should be well-focused on their particular areas of interest. The fact that many of our TRANSACTIONS find it difficult to obtain enough good papers to fill a reasonably sized issue and to publish on a regular schedule may indicate that they are too narrowly focused and do not encompass a large enough area of activity. At the other extreme, some TRANSACTIONS may attempt to cover such a wide field that they are not able to attract the right readership. Both extremes should be avoided. The "ideal" TRANSACTIONS should have a wide enough area of interest that it can publish reasonably sized issues—say ten or more papers—on at least a quarterly schedule, and preferably on a bimonthly or monthly schedule, although there well may be exceptions in newly developing areas.

Fourth, a major reason for a reasonably frequent publication schedule is to minimize the time between the submission of a paper and its availability to readers. While the practices of the Groups vary widely with respect to their review procedures and the Institute's policies must be flexible to allow for the different requirements of the different areas, it is certainly desirable to expedite the flow of technical information. Needless duplication of effort can best be avoided by prompt and widespread dissemination of technical knowledge. In this connection, the desirability of publishing a special periodical on perhaps a biweekly schedule, thus permitting rapid dissemination of brief articles of current importance, is at present being studied.

Fifth, the IEEE should publish good periodicals in all its areas of interest. There should not be areas of obvious present or future interest to the Institute in which there is no coverage. At present, there are several such areas and it may well be that the Institute needs to study in a systematic way the coverage of its present publications and develop its activities in those areas of most importance.

Clearly, this list is not complete. There are undoubtedly other significant requirements that our publications should strive to meet. An effort to find and define such requirements may well be the first step toward improvement.

F. Karl Willenbrock

The IEEE looks ahead

B. M. Oliver *President, IEEE*

Two years ago, the IEEE came into existence; a society merged at the top but nowhere else. During the last two years the merger of all the organizational units of AIEE and IRE origin into new IEEE organizational units has taken place. Section mergers have been completed. Technical Committees have merged with Groups or transferred to Group status and Groups have merged with other Groups. Rather than the dual structure we started out with, comprising 29 Groups and some 70 Technical Committees, we now have 35 (soon to become 32) Groups and five general committees all under the guidance of one body: the Technical Activities Board (TAB). We begin 1965 with merger an accomplished fact and it is time to turn our attention from the problems of the past to the problems of the future.

It is tempting to say at this point: "Well, that's that. IEEE is now organized. Let's leave well enough alone." But to do this would be to fail to recognize the continually changing needs of our profession. New technologies appear, grow, flourish, and then either reach a semistatic maturity or gradually wane. Since IEEE must always be concerned with new developments and active fields, the effort and activity we devote to new technologies must grow with them and then subside. As a result, we need a continually changing organizational structure, continual redefinition of Group scopes, continued changes in our pattern of symposia and in our publications. Adaptive change must be the only permanent feature of IEEE.

In the past we have adapted to the rise of new technologies by the formation of new Professional Groups in the IRE, or new Technical Committees in the AIEE. Once created, these Groups and Committees have tended to last indefinitely. We have never yet discontinued a Group because of inadequate activity; however, unless we do so the number of Groups (and the number of publications) will grow without limit. Criteria for the survival of a Group or Committee must be established, and if it fails to measure up it must either be merged with Groups having related interests or be discontinued.

IEEE must be of greater service to the membership than either AIEE or IRE was before the merger. Otherwise all our efforts will have been in vain. Our membership service is principally through publications and meetings. There can be too few or too many of both. I think we have too many. Too many publications of limited circulation are unattractive to members, libraries, and authors alike. We should, I believe, seek the minimum number of publications, consistent with the economics involved, that will adequately cover the field. This can be done by combining those having closely related scopes. The staggering number of symposia can be reduced by making them jointly sponsored, and somewhat broader in scope. At the same time we must not lose sight of the advantages of an occasional small meeting, or workshop, attended principally by workers at the forefront of a particular field, or even an occasional *Transactions* of limited circulation in a newly developing area.

It will be recognized that this kind of publication and meeting coordination means some loss of autonomy for the Groups. Not much, but some. But if it makes sense to have all these Groups under one umbrella—the IEEE—rather than to have 35 separate societies, surely it is because of this very coordination, which the Institute as a whole can provide. Our Groups are like states and in their union lies strength. While we must respect and preserve certain states' rights, we must not carry this principle so far that we imperil or weaken the nation.

Other types of membership service should be explored. As Junior Past President Clarence Linder has remarked, we should try to develop other ways, besides the usual conventions and publications, to provide continuing education for our members. Modern methods of abstracting and information retrieval must be developed so that all the available information on any subject can be obtained almost instantly by any member.

Nationalism has no place in science. The non-national character of IEEE is consonant with this belief and can lead to more rapid technological progress in our profession throughout the world. To increase the world-wide value of IEEE, we must solve the problems Sections and Chapters face in countries outside the United States. We should establish friendly working relations with other professional societies. Our goal should be to make IEEE a *local* reality around the world by virtue of active Sections and ultimately by IEEE-sponsored conventions in centers of electrical technology everywhere.

Some of these goals may be attained this year; others may take several years, but are worth the effort. Now that we're one society it's up to all of us to make it vigorous and active—not just the biggest but also the best professional society, and a world-wide one.

Parametric principles in optics

In this introduction to the newly developing field of nonlinear optics, emphasis is placed upon the similarities and the intrinsic differences between optical processes and the equivalent radio-frequency processes

I. P. Kaminow *Bell Telephone Laboratories, Inc.*

The development of the optical maser has made available for the first time a light source with the same properties as conventional radio-frequency sources; that is, the output may be formed into a plane wave oscillating at a single frequency. These new sources have been used to demonstrate phenomena that were previously unknown at optical frequencies but have long been familiar at radio frequencies—namely, harmonic generation, mixing, heterodyning, modulation, and parametric amplification, all by means of nonlinear reactances. Most of the processes involved in the new field of nonlinear optics can be understood in terms of parametric mixing and amplification. In the optical case, the refractive index is varied by means of the nonlinear response of the medium. (The optical dielectric constant is the square of the index of refraction.)

Some optical masers are capable of emitting 10 MW of power during short pulses. When the light is focused into a volume having dimensions comparable to an optical wavelength, the electric fields may be as great as 10^9 volts/cm and the nonlinear character of the refractive medium is readily manifest. All refractive media exhibit some nonlinearity, however, even at very low electric field strengths. Hence, nonlinear effects may also be observed with a low-power plane-wave optical maser beam if the conditions for traveling-wave parametric interaction are satisfied. Then the interaction products may build up over extended path lengths. The necessary phase-matching conditions will be examined in a later section.

Despite the identity in underlying principles of operation, optical and RF parametric devices differ widely in form because of the disparity in wavelengths of the radiation. At optical wavelengths (10^{-4} cm) it is much simpler to produce the nonlinear interactions in a bulk medium, with dimensions much greater than the wavelength, than to follow the RF practice of trying to construct elements with dimensions comparable to or smaller than the wavelength.

In many respects, the optical device is easier to construct and understand and has greater flexibility than its RF counterpart. Moreover, the optical experiments to be described shortly have a certain elegance and beauty because the results of the parametric interactions can literally be seen.

Before the traveling-wave requirements and nonlinear optical experiments are described, it will be instructive to consider the properties of a nonlinear lumped RF capacitor and, in particular, the restrictions laid down by the symmetry of the element. This treatment will lead to similar symmetry restrictions for bulk optical media.

Symmetry

When a voltage V is applied to a capacitor filled with a nonlinear dielectric, as shown in Fig. 1, a charge Q appears on the plates. The charge and voltage are related by the capacitance $C(V)$:

$$Q = C(V) \cdot V \qquad (1)$$

The voltage may be regarded as the excitation and the induced charge as the response. The capacitance is written as an explicit function of V to emphasize the nonlinearity of (1). Expanding Q in a polynomial series,

$$Q = C_1V + C_2V^2 + C_3V^3 + C_4V^4 + \ldots \qquad (2)$$

The response functions corresponding to the first four terms are illustrated in Fig. 1. The first term (a linear capacitor) produces a response at the exciting frequency; the second term, a response at the second harmonic of

the excitation; the third term, a response at the third harmonic; and so on. If two excitation frequencies are present simultaneously, the response will contain various mixing products. For example, the second term in (2) will give rise to the sum and difference of the exciting frequencies as well as to dc and second-harmonic terms.

Now suppose that the capacitor in Fig. 1 is symmetrical so that, except for the labels, terminals 1 and 2 are indistinguishable. Then, clearly, reversing V will reverse the sign of Q without altering its magnitude. But this behavior, required by symmetry, is consistent only with the response function's odd terms, C_1V, C_3V^3, ..., and not the even ones, C_2V^2, C_4V^4, ... Therefore, C_2, C_4, and all even coefficients must vanish for a truly symmetrical capacitor, and only odd harmonics can be present in the response. In order to generate second and other even harmonics an asymmetrical capacitor is required.

A p-n junction has distinguishable terminals (p and n) and therefore may generate second harmonics. The junction acts as a rectifier and, obviously, does not respond symmetrically on reversal of the applied voltage. If the p-n junction is back-biased, it operates as a nonlinear capacitor for small ac applied voltages.

To extend the discussion to nonlinear optical materials it is necessary to examine the properties of bulk dielectric media rather than lumped circuit elements. It is convenient to form an equation similar to (1) in terms of quantities characteristic of the bulk material rather than the geometry of the element,

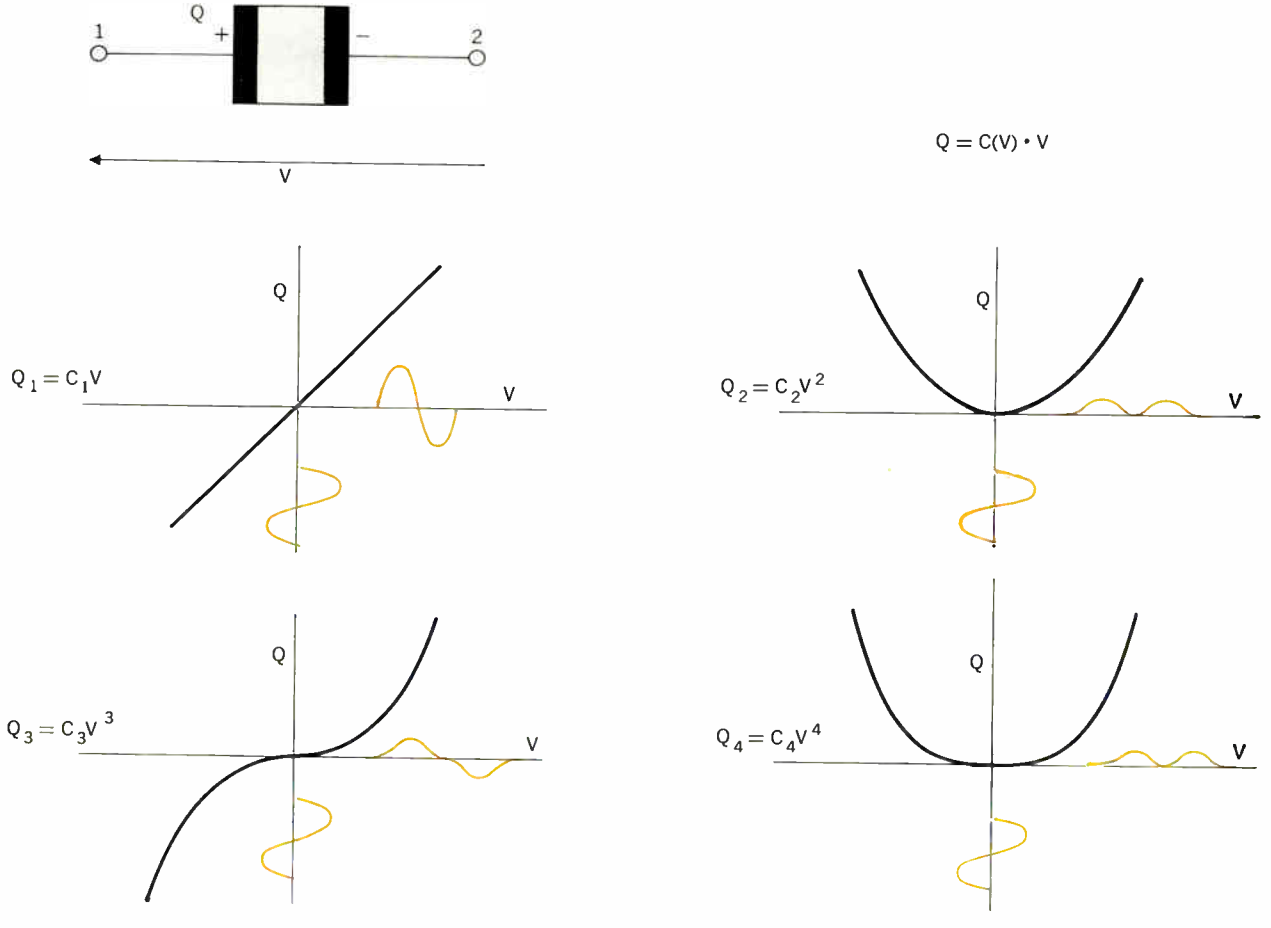
$$P = [K(E) - 1]\epsilon_0 E \qquad (3)$$

where P is the electric polarization (charge per unit area), E the field strength, ϵ_0 the permittivity of free space, and $K(E)$ the (field-dependent) dielectric constant. Here, P may be regarded as the response to the excitation E in analogy with Q and V for the capacitor. Symmetry imposes restrictions on the response function in three-dimensional bulk media just as it does in the case of a lumped capacitor.

A piece of glass is symmetrical in the sense that there are no features of the bulk material (not its bounding surface) that permit a distinction between a plus and minus direction within the medium; hence, reversing the exciting field must reverse the sense of the response without changing its magnitude, as in Fig. 2(A). Materials of this nature, including ceramics, most liquids, and many crystals, are said to possess a center of symmetry. Put differently, any vector relation that holds within the substance in a coordinate system (x, y, z) holds unchanged when the coordinate system is inverted so that (x, y, z) becomes $(-x, -y, -z)$.

Some crystals, such as gallium arsenide (GaAs), lack a center of symmetry. A GaAs crystal consists of unequally spaced Ga and As layers as shown in Fig. 2(B). An imaginary observer inside the crystal would be able to distinguish *up* from *down*. On the other hand, if gallium and arsenic were each replaced by carbon, the resulting crystal, diamond, would have a center of symmetry, and

Fig. 1. Charge Q developed by a nonlinear capacitor C(V) in response to an applied voltage V.



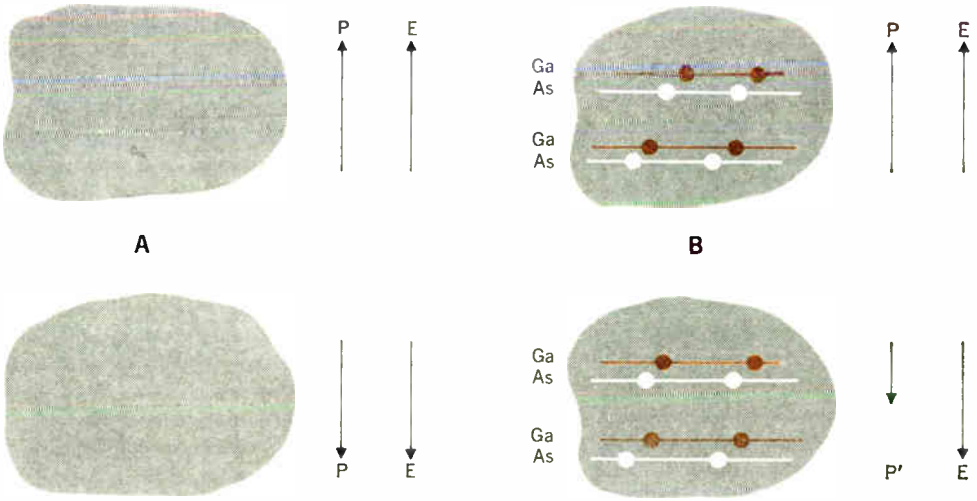


Fig. 2. Excitation E and response P in (A) a centrosymmetric medium, such as glass or diamond, and (B) a noncentrosymmetric crystal, such as gallium arsenide.

up and down would be indistinguishable.* Hence, an electric field E directed either up or down in diamond would be expected to produce a response P with the same magnitude, as shown in Fig. 2(A). But in gallium arsenide the two senses of E are distinguishable and the corresponding responses may differ, as in Fig. 2(B).

The polarization of the nonlinear medium may be expanded in powers of E,

$$P = a_1E + a_2E^2 + a_3E^3 + a_4E^4 + \ldots + a_kE^k + \ldots \quad (4)$$

In general, the coefficients a_k depend upon the frequencies present in E and P. In centrosymmetric materials, P is an odd function of E, $P(E) = -P(-E)$, and all even coefficients must vanish. In noncentrosymmetric materials, P is neither an even nor an odd function of E and all terms in (4) may be present. Thus, in analogy with the nonlinear capacitor, only materials that lack a center of symmetry can generate even harmonics, or sum and difference frequencies, when two signals are present.

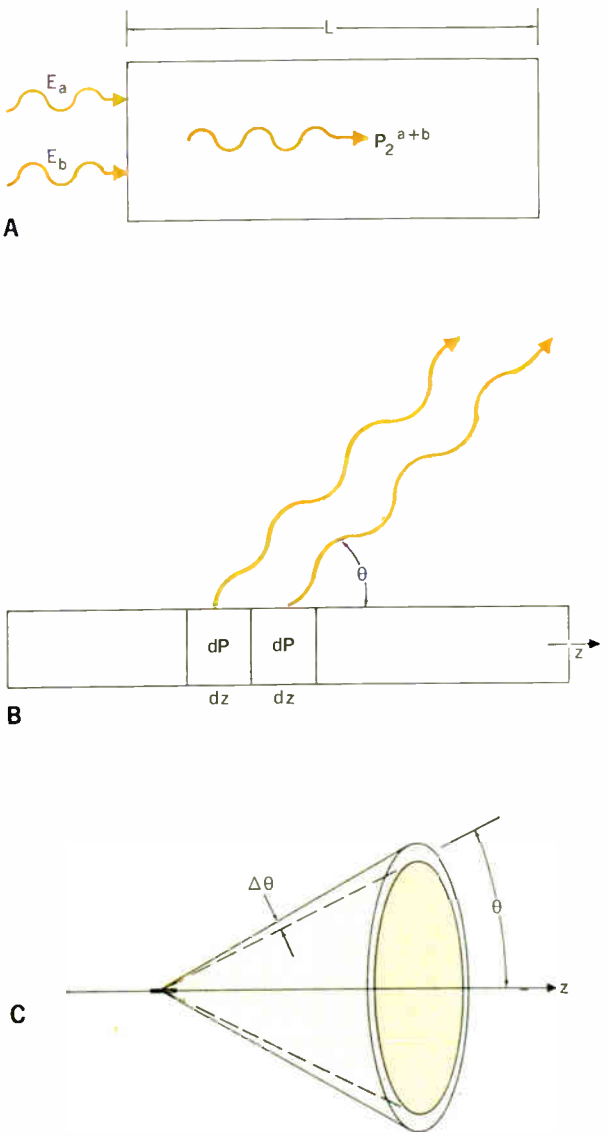
Traveling-wave interaction

So far, only interactions taking place at a point have been considered. If the interaction takes place over an extended region the nonlinear response created at one point will add in phase with that created at another point only if certain conditions are satisfied. If these phase-matching conditions are met, the cumulative response may become large.

Consider the case of two traveling waves passing through a dielectric medium of length L as in Fig. 3(A). The exciting field, in complex form, is

$$E(t, z) = E_a e^{j(\omega_a t - \beta_a z)} + E_b e^{j(\omega_b t - \beta_b z)} \quad (5)$$

Fig. 3. A—Polarization wave P^{a+b} at the sum frequency is produced by two incident waves E_a and E_b. B—Energy radiated by differential elements dz add in phase at angle θ with beam width $\Delta\theta$. C—Cone of radiation produced.



* A homely analogy may be instructive. Consider a vegetable garden with rows running north and south and with the spacing between rows 1 foot, 2 feet, 1 foot, 2 feet, ... If radishes were planted in every row, a rabbit traversing the garden along an east–west line could not tell whether he was coming or going; that is, travel in one direction would be indistinguishable from travel in the other. But if, instead, peas and carrots were planted in alternate rows, movement from peas to carrots between closely spaced rows would define one direction and movement from carrots to peas between closely spaced rows would define the opposite direction. The radish garden is centrosymmetric (in two dimensions) but the peas-and-carrots garden is noncentrosymmetric.

where the phase velocities of the two waves in the medium are given by

$$v_a = \frac{\omega_a}{\beta_a} \quad \text{and} \quad v_b = \frac{\omega_b}{\beta_b} \tag{6}$$

respectively. For the sake of illustration, let us concentrate on only one term in the nonlinear response, a_2E^2, and only one of the modulation products it produces, the sum frequency

$$\omega_{a+b} \equiv \omega_a + \omega_b \tag{7}$$

The corresponding complex polarization obtained by substituting (5) into (4) is

$$P_2^{a+b}(t, z) = a_2E_aE_be^{j[(\omega_a+\omega_b)t-(\beta_a+\beta_b)z]} \tag{8}$$

where the actual polarization is the real part of (8).

The polarization wave acts like a uniformly excited linear antenna oscillating at frequency $\omega_a + \omega_b$ with a progressive phase shift $(\beta_a + \beta_b)z$ along the antenna. The phase velocity in the dielectric medium into which the antenna radiates at the sum frequency is

$$v_{a+b} = \frac{\omega_{a+b}}{\beta_{a+b}} \tag{9}$$

In the far field, the radiation from every differential length dz will add in phase only in a direction θ measured from the antenna axis such that

$$\beta_{a+b}\,dz\cos\theta = (\beta_a + \beta_b)\,dz \tag{10}$$

as illustrated in Fig. 3(B). When the frequencies (ω_a and ω_b) and the phase velocities characterizing the material (v_a, v_b, v_{a+b}) are such that $\cos\theta = 1$, the radiation will be confined to a small solid angle along the z axis (end-fire). It is known from antenna theory that the beam width in the end-fire direction is roughly $2\sqrt{\lambda/L}$, rather than λ/L as in the broadside direction, when $\lambda/L \ll 1$. In the present case, λ is the wavelength at the sum frequency. When the experimental conditions give $\theta > \sqrt{\lambda/L}$ or, equivalently, $\cos\theta < 1 - \lambda/2L$, very little energy at ω_{a+b} will appear along the axis; it will be distributed over a rather large solid angle near the surface of a cone, as in Fig. 3(C). On the other hand when $\cos\theta > 1 + \lambda/2L$, angles within the beam width are imaginary and very little energy at ω_{a+b} is radiated in any direction.

It may be concluded that in order to collimate appreciable energy at ω_{a+b}, it is necessary to satisfy the condition

$$|\cos\theta - 1| < \frac{\lambda}{2L} \tag{11}$$

or, using (10),

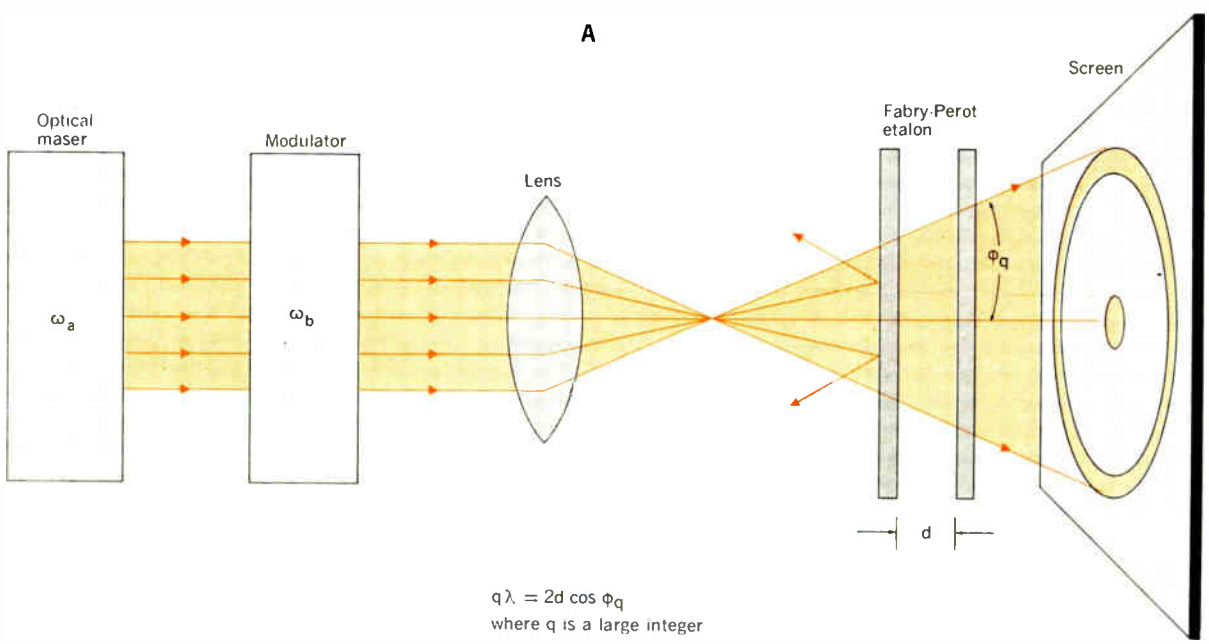
$$|\beta_a + \beta_b - \beta_{a+b}| < \frac{\pi}{L} \tag{12}$$

where it has been assumed that $\lambda/L \ll 1$. The interaction is most efficient, of course, when the left-hand members of (11) and (12) vanish identically. For L less than or comparable in size to λ, the interaction effectively takes place at a point and the traveling-wave restrictions disappear.

A simple case in which phase matching occurs is when all velocities are equal; i.e., $v_a = v_b = v_{a+b}$. Another simple example is when L is small, as when the interaction take place in the focal region of a lens, so that the restriction on the β's is not very stringent.

Equation (12) holds for two interacting waves, E_a and E_b, traveling in the same direction. If the two waves are not parallel, then (12) may be written in vector form with $\mathfrak{z}_a$ and $\mathfrak{z}_b$ directed normal to the wavefronts of E_a and E_b in the direction of propagation. To generalize further, the other modulation products $m\omega_a + n\omega_b$, in which m and n are positive or negative integers, may also be

Fig. 4. A—Schematic arrangement for observing splitting of Fabry-Perot rings by modulated light. (For the sake of simplicity, sideband rings are not indicated.) B—Ring pattern for red light modulated at 9 Gc/s, with d = 5.6 mm.



included. Then the ω, β conditions for traveling-wave interaction of two waves are

$$\omega_{m+n} = m\omega_a + n\omega_b$$
$$\beta_{m+n} = m\beta_a + n\beta_b \qquad (13)$$

Note that the product frequency ω_{m+n} may be generated only when the coefficient a_k in (4) does not vanish for $k = |m| + |n|$.

Mixing experiments

We are now in a position to illustrate some of the foregoing ideas by nonlinear optical experiments performed recently at various laboratories.

Modulation. It has been known for about 90 years that an electric field applied to a transparent substance can vary its index of refraction. A time-varying electric field will modulate the refractive index and hence the phase shift experienced by light passing through the medium. In materials that have a center of symmetry this electrooptic phenomenon is known as the Kerr effect; in noncentrosymmetric materials it is called the Pockels effect. In our terms, light modulation by these electrooptic effects may be viewed as parametric mixing of an optical field and an RF modulating field. Taking ω_a to be the optical carrier frequency and ω_b the modulating frequency, the first sidebands occurring in connection with the Kerr effect are at $\omega_a \pm 2\omega_b$, corresponding to $m = 1$, $n = \pm 2$ in (13), since $a_2 = 0$ and $a_3 \neq 0$. For modulation by the Pockels effect, however, the first sidebands appear at $\omega_a \pm \omega_b$ and are displaced from the carrier by the fundamental modulating frequency rather than the second harmonic. The process corresponds to $m = 1$, $n = \pm 1$ with $a_2 \neq 0$.

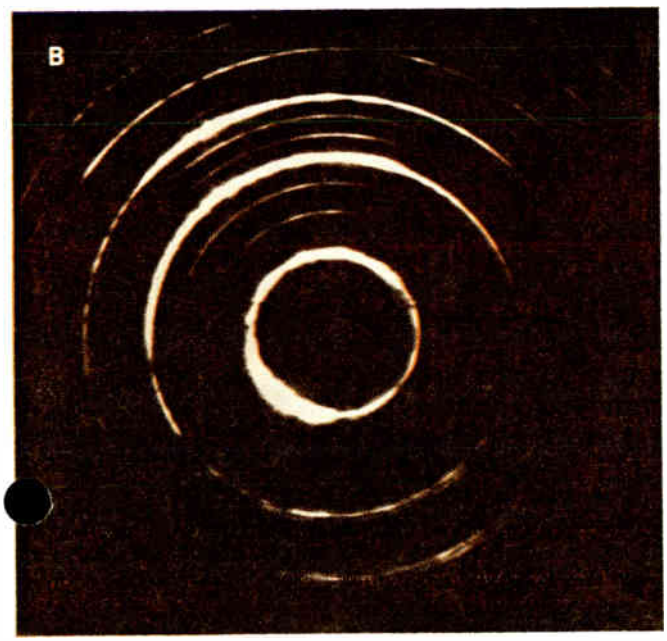
The crystal most widely employed for its Pockels effect is potassium dihydrogen phosphate (KDP), whose electrooptic properties were first reported 20 years ago. In one experiment KDP was used to modulate red light from an optical maser ($\omega_a = 2\pi \times 5 \times 10^{14}$ c/s) at a microwave frequency ($\omega_b = 2\pi \times 9 \times 10^9$ c/s). In order for the carrier and sidebands to be observed directly, the modulated light was passed through a form of spectrum analyzer known as a Fabry–Perot etalon. This device consists of two parallel, partially transparent mirrors separated by a distance d. There are certain discrete angles of incidence for which the structure is resonant when it is excited by monochromatic light. Light rays incident at these angles are transmitted by the resonator, whereas other rays are reflected. When a distribution of angles of incidence is present in the light falling on the device, a series of concentric rings is produced, as shown schematically in Fig. 4(A). When modulated monochromatic light is incident on the Fabry–Perot etalon, two additional series of rings, for the upper and lower sidebands, appear. A Fabry–Perot pattern (with the light directed slightly off the etalon axis) is shown in Fig. 4(B). The bright rings are due to the carrier, and the fainter, partial rings are produced by the sidebands. The modulating frequency is 9 Gc/s and d is one-half wavelength at 27 Gc/s. Under these conditions, the sidebands divide the carrier ring spacing into thirds.

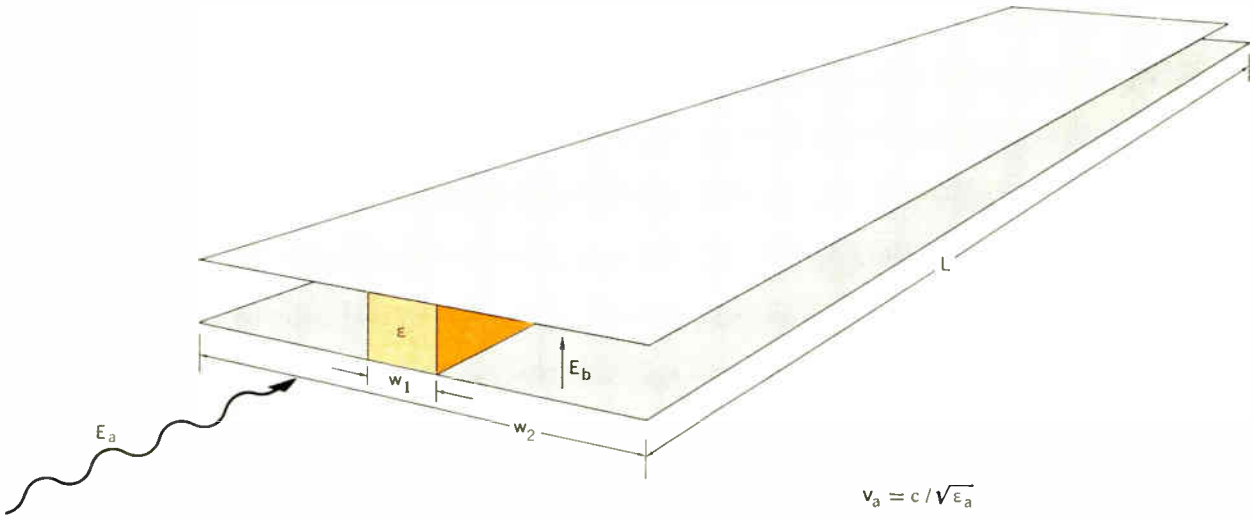
In order to reduce the field strength and power required to produce a given modulation index, it is desirable that the interaction length L be large. Then, for optical and RF waves traveling in the same direction, the RF modulating and optical carrier velocities must be matched. With an RF dielectric constant of 20 and an optical dielectric constant of 2.25 in KDP, one method for achieving this match is to partially fill the cross section of a parallel plate transmission line as shown in Fig. 5. At low enough frequencies, the effective RF dielectric constant of the structure will be determined by the ratio of air to KDP in the cross section of the transmission line as indicated in the figure. Traveling-wave modulators of this sort have been constructed with $L = 1$ meter by joining many individual crystals together with an optical cement.

Harmonic generation, optical mixing, heterodyning, and rectification. Consider now the case in which ω_a and ω_b are both optical frequencies. The sum and difference frequencies may be produced by a noncentrosymmetric crystal ($a_2 \neq 0$). Both a sum frequency (mixing) and a difference frequency (heterodyning) have been observed for $\omega_a \neq \omega_b$. For $\omega_a = \omega_b$, the sum frequency becomes the second harmonic and the difference frequency a dc polarization of the medium. Both effects have been observed. Third-harmonic generation of light has also been observed using a centrosymmetric crystal ($a_2 = 0$, $a_3 \neq 0$).

Much of this work was accomplished using high-power pulsed optical masers focused inside the crystals. With the resultant short interaction length and high field strength, it was not essential that the β conditions be satisfied precisely. However, when these conditions are satisfied, the interaction becomes much more efficient.

The velocity-matching problem for second-harmonic generation in KDP was solved in the following way. The optical velocity at the fundamental frequency (red) differs from that at the second-harmonic frequency (ultraviolet). However, because of the particular symmetry of the crystal, the velocity at any frequency is also a function of the direction of propagation and plane of polarization



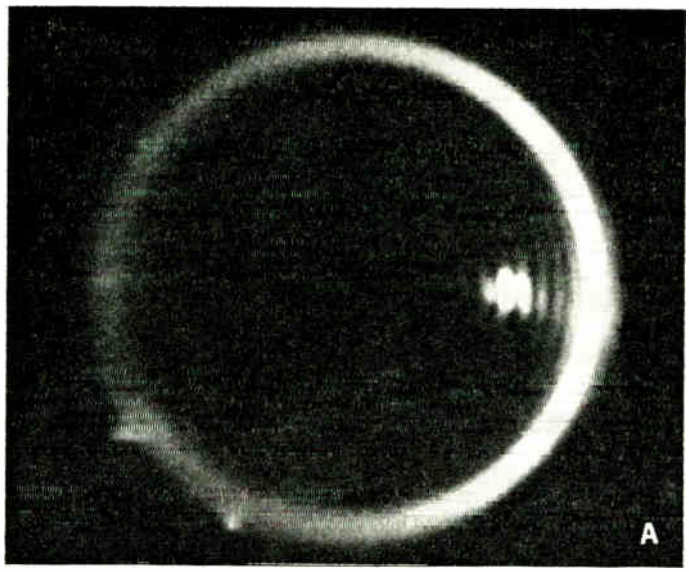


$$v_a = c/\sqrt{\varepsilon_a}$$

$$v_b = c/\sqrt{\varepsilon_{eff}}$$

$$\varepsilon_{eff} = 1 + (\varepsilon_b - 1)\frac{w_1}{w_2}$$

Fig. 5. Parallel-plate structure containing a KDP rod with optical dielectric constant ϵ_a and RF dielectric constant ϵ_b.



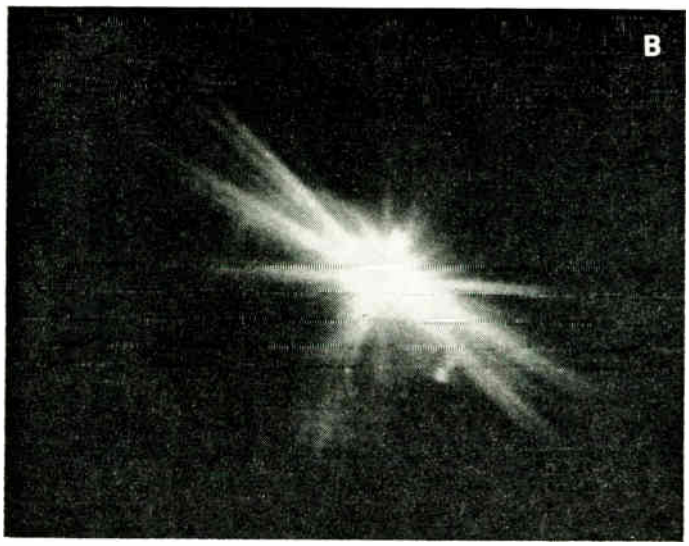


Fig. 6. A—Ring produced by cone of second-harmonic light from a KDP crystal when the incident beam deviated slightly from the matching direction. B—Fundamental and harmonic velocities matched.

of the light. It happens that the properties of the crystal are such that the second-harmonic ultraviolet light is polarized perpendicular to the incident red light and that it is possible to find a particular propagation direction in the crystal for which the red and ultraviolet velocities are equal.

The coalescence of the second-harmonic light cone into a small solid angle, as discussed in connection with Fig. 3, is illustrated in Fig. 6. When the incident maser beam is about 1° from the matching direction in the crystal, the second-harmonic light is distributed on a cone with an apex angle of about 4°, as shown in Fig. 6(A). When the velocity-matching conditions are satisfied exactly ($\cos\theta = 1$), all the ultraviolet light coalesces into a narrow end-fire beam, as in Fig. 6(B). The direction of the maser beam and cone axis in Fig. 6(A) do not coincide because of crystal anisotropy.

Parametric amplification by the Raman effect

Before turning to the optical analog, let us review briefly some aspects of parametric amplification at radio frequencies. Consider the section of the RF circuit in Fig. 7 to the left of the dashed line. It consists of a nonlinear capacitor, a branch tuned to frequency ω_v and another branch with a tuned circuit and generator at the pump frequency ω_p. With this combination, voltages will be developed across the nonlinear capacitor at frequencies $\mu\omega_p + \nu\omega_v$, with μ and ν positive or negative integers. The only currents that may flow, however, are at ω_v and ω_p. If now the series-resonant branches tuned to $\omega_\pm = \omega_p \pm \omega_v$ are included, currents may flow through the nonlinear capacitor at these frequencies. The nonlinear capacitor

$C(V)$ may deliver or absorb power at the four frequencies ω_p, ω_v, ω_+, ω_-, and only at these frequencies. The relationships among the net powers W_p, W_v, W_+, W_-, *delivered* by $C(V)$ to the circuit at the respective frequencies, are defined by the familiar Manley–Rowe equations:

$$\frac{W_+}{\omega_+} + \frac{W_v}{\omega_v} - \frac{W_-}{\omega_-} = 0$$

$$\frac{W_+}{\omega_+} + \frac{W_p}{\omega_p} + \frac{W_-}{\omega_-} = 0 \qquad (14)$$

These relationships are quite general and hold at the terminals or boundaries of any nonlinear (lossless) reactive system, whether lumped or distributed, when it is terminated or bounded as indicated in Fig. 7 so as to permit power to be transferred only at ω_r, ω_p, ω_+, ω_-. The internal details of the nonlinear interaction need not be considered.

In the typical RF parametric amplifier, the ω_+ branch is absent and $W_+ = 0$. If no generator is provided in the ω_r branch, then $W_r \geq 0$ (no external power at ω_r). Thus, the input power provided by the pump at ω_p must be divided between W_v and W_- according to the ratio of the corresponding frequencies. (If an open circuit in the ω_v branch causes W_v to vanish, for example, then $W_- = 0$ also. Moreover, according to (14), $W_p = 0$; that is, the impedance presented to the pump is purely reactive.) When $W_v > 0$, the impedance appearing across the terminals of the generator in the ω_- branch will have a negative real part and power will be delivered *to* the (small-signal) generator rather than *by* it. The phase of the current at ω_-, however, will be determined by the generator e_-. If the incident and output signals at ω_- are separated by a circulator, it becomes clear that the circuit can operate as an amplifier. (The process just described is the counterpart of quantum mechanical *stimulated emission*. The small incident signal e_- stimulates the circuit to emit a greater signal at the same frequency.) When the negative resistance gets to be as large as the generator resistance, oscillation will occur with $e_- \equiv 0$.

If the ω_+ branch is now returned to the circuit, it becomes clear from the Manley–Rowe relations that power may not be delivered by the circuit at ω_+ unless $W_- > 0$—that is, unless power is being extracted simultaneously at ω_-. (It is assumed, as before, that $W_v \geq 0$.)

Let us keep these ideas in mind while examining some properties of refractive media that permit analogous processes to occur at optical frequencies. In place of the *R-L-C* circuit resonant at ω_r, a mechanical vibration of a molecule in a gas or liquid or a lattice vibration in a solid may be employed. When these atomic vibrations occur, they influence the electronic polarizability (or refractive index) of the medium and thereby modulate transmitted light at the vibration frequency ω_v. If the atomic displacement in the normal coordinate corresponding to the ω_v vibrational mode is called x_v, then a term $b_r E x_v$ must be added to the expansion for P in (4) to account for the vibrational–optical interaction. The coefficient b_v is very large only when x_v is driven near the resonant frequency ω_v. If E is an optical field at ω_p and x_v is oscillating at ω_v, then the transmitted light will have components at ω_+ and ω_-. Sidebands produced by atomic vibrations in this fashion are called Raman frequencies, after the man who observed the effect in 1927. Those vibrational modes with symmetry such that $b_r \neq 0$ are said to be Raman active. (A center of symmetry or lack of it in a substance does not enter directly into determining Raman activity of a vibrational mode.) For historical reasons the lower sideband, ω_-, is called the Stokes frequency and the upper one, ω_+, the anti-Stokes frequency.

The response function of a Raman-active medium has the form

$$P = a_1 E + a_2 E^2 + a_3 E^3 + \ldots + b_r E x_v + \ldots \qquad (15)$$

where, as before, a_2 vanishes when a center of symmetry is present. The a_2 term is capable of mixing two electromagnetic waves (optical or RF) to produce a third, the a_3 term mixes three electromagnetic waves to produce a fourth electromagnetic wave, and the b_r term mixes a vibrational wave and an electromagnetic wave to produce another electromagnetic wave. These interactions serve to couple the branches in the equivalent circuit of Fig. 7 through the nonlinear capacitor. The resonant behavior of the coefficient b_r takes the place of the ω_r resonant circuit.

The pump power may be provided by an optical maser. In order to produce gain at the Stokes frequency, with ω_+ suppressed, only the b_r term need be considered:

$$\omega_- = \omega_p - \omega_v \qquad (16)$$

The mixing arises from the polarization product

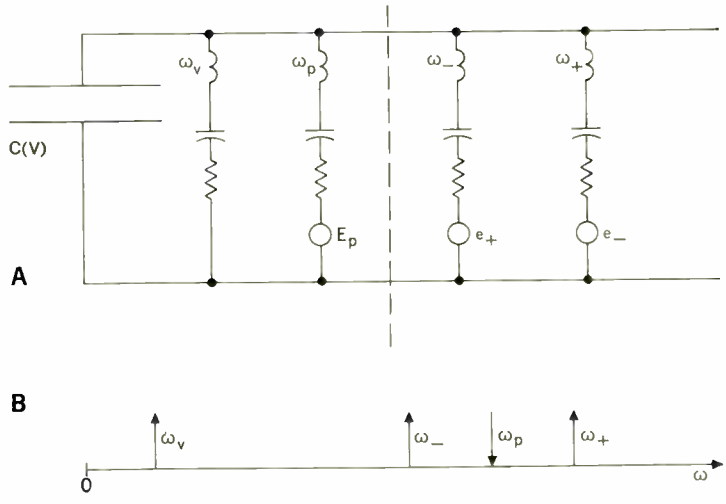
$$P_- = \tfrac{1}{2} b_r E_p e^{j\omega_p t} X_v^* e^{-j\omega_v t} \qquad (17)$$

in which E_p and X_v^* are complex, spatially varying electric-field and vibrational amplitudes. The upper sideband ω_+ might be produced in similar fashion, but the assumption that $W_r \geq 0$, as well as other physical reasoning (to be mentioned), suggests that a process involving the mixture of three optical frequency components

$$\omega_+ = \omega_p + \omega_p - \omega_- \qquad (18)$$

by means of the a_3 term is more efficient in producing anti-Stokes energy when a Stokes component is present in

Fig. 7. A—Nonlinear network containing pump generator E_p and small-signal generators e_+ and e_-. B—Relationships among frequencies showing net power in at ω_p only.



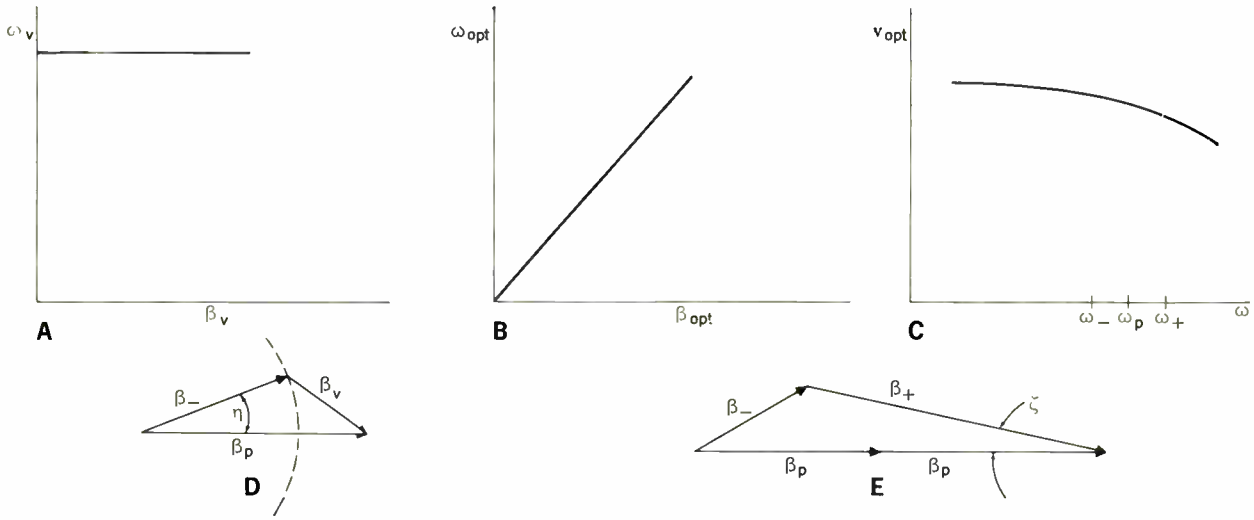


Fig. 8. A—ω vs. β for vibrational waves. B—ω vs. β for optical waves. C—Dispersion in optical velocities. D—Vector diagram for the production of Stokes radiation. E—Vector diagram for the production of anti-Stokes radiation.

the exciting field. The corresponding polarization product is

$$P_+ = \tfrac{3}{4}a_3 E_p e^{j\omega_p t} E_p e^{j\omega_p t} E_-^* e^{-j\omega_- t} \qquad (19)$$

in which E_-^* is the complex conjugate of the Stokes field amplitude. (The a_2 term is not required in any case; if it were present, interactions involving the second-harmonic frequency $2\omega_p$ would also have to be considered.)

In order that the mixing be coherent over an extended region, β conditions analogous to (12) must be satisfied. For the Stokes frequency produced by the b_v term we require that

$$\beta_- = \beta_p - \beta_v \qquad (20)$$

For the anti-Stokes frequency produced by the a_3 term,

$$\beta_+ = \beta_p + \beta_p - \beta_- \qquad (21)$$

where β_p is in the direction of the incident maser beam.

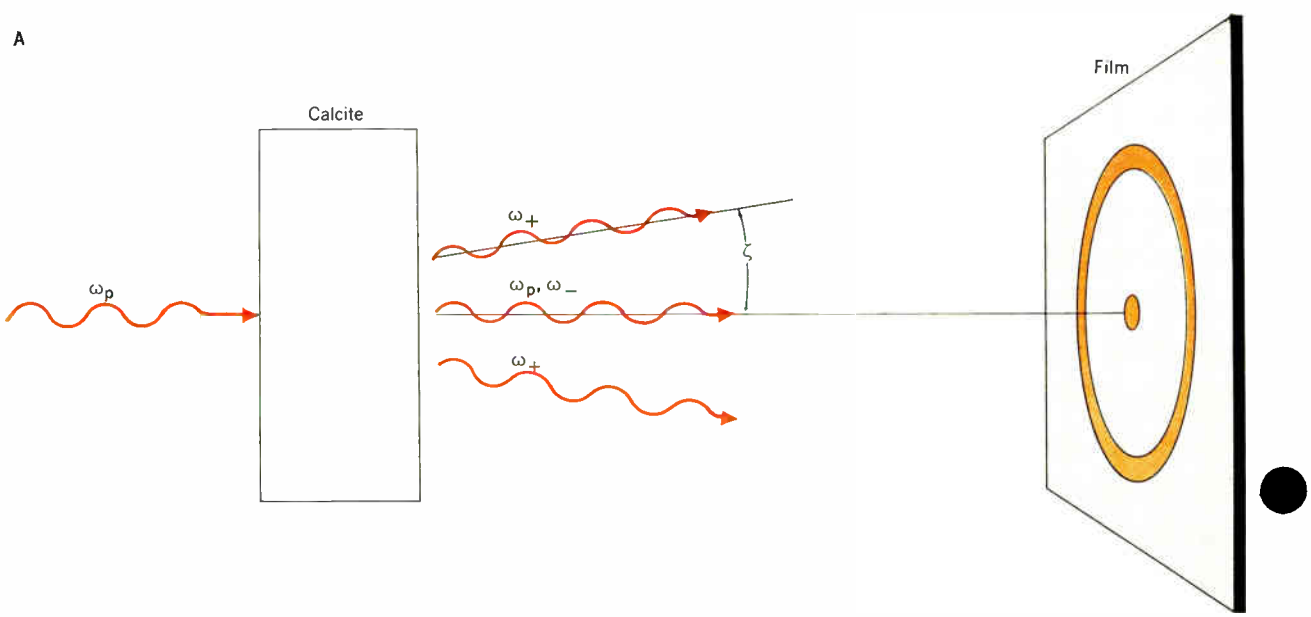
The alternative but equivalent two-step interaction

$$\beta_r = \beta_p - \beta_- \qquad \beta_+ = \beta_p + \beta_v \qquad (22)$$

which involves β_r explicitly, is rejected because the vibrational waves are normally more strongly attenuated than the optical waves, so the interaction remains coherent only over short distances.

A plot of ω vs. β for optical waves has a slope equal to

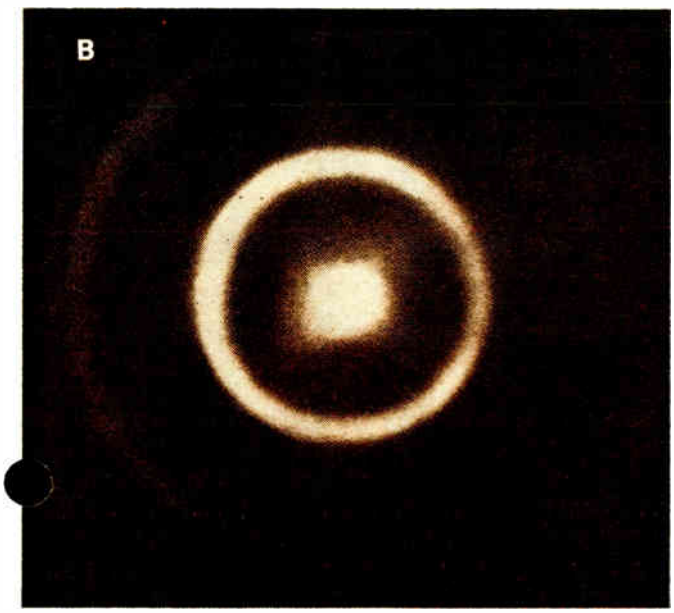
Fig. 9. A—Schematic arrangement for observing anti-Stokes rings. B—First- and second-order rings produced by cones of anti-Stokes radiation.



the velocity of light in the medium; see Fig. 8(B). However, the ω vs. β curve near $\beta = 0$ for the vibrational wave is very nearly a horizontal line, $\omega_v \neq f(\beta_v)$; see Fig. 8(A). The phase velocity at ω_v is arbitrary and the group velocity approaches zero. The reason is that the vibrating molecules in a liquid, for example, are not tightly coupled to one another and cannot propagate a signal effectively. A wave at ω_v can be set up with any propagation vector β_v (or wavelength) and (20) can be satisfied for a wide range of angles η as in Fig. 8(D). However, since (21) involves only optical waves with well-defined propagation vectors at the respective frequencies, it will be satisfied for a discrete angle ζ, as in Fig. 8(E), which depends upon the dispersion in optical phase velocity $v(\omega)$ shown in Fig. 8(C). The maximum output at ω_+ occurs near this matching angle and the resulting radiation is emitted on a cone with apex angle 2ζ. Energy generated at ω_v is dissipated as heat.

The conversion of the red light from a pulsed optical maser ($\omega_p = 2\pi \times 5 \times 10^{14}$ c/s) to anti-Stokes emission has been strikingly demonstrated by focusing a high-power pulsed maser into a crystal of calcite, which has a Raman-active vibration at $\omega_v = 2\pi \times 3 \times 10^{13}$ c/s; see Fig. 9(A). The ring produced by an anti-Stokes cone is shown in Fig. 9(B). The angle ζ is about 0.2°. The Stokes radiation, from the process in (20), is strongest in the forward direction and appears at the center of the anti-Stokes ring along with transmitted maser light. One reason the Stokes radiation may prefer the forward direction is that the smallest value of β_v (longest vibrational wavelength) occurs for this direction. The smallest attenuation of the vibrational wave is expected under this circumstance, permitting maximum net gain for the Stokes line. Other reasons, involving the character of the exciting maser light, have also been suggested.

The second ring in Fig. 9(B) corresponds to the second-order anti-Stokes line $\omega_p + 2\omega_v$. This and higher-order rings are permitted because no means for suppressing them (other than observation at a particular angle) is provided.



Conclusion

An attempt has been made to avoid the superficial distinctions between parametric processes at radio and optical frequencies and to elucidate the intrinsic differences. Some recent optical experiments have been described in an intuitive fashion. These topics are described in greater detail in the sources listed in the Bibliography, which is meant to be representative rather than exhaustive.

This article is based on a talk presented at the National Electronics Conference, Chicago, Ill., October 19–21, 1964.

The author gratefully acknowledges the permission of J. A. Giordmaine for the use of Fig. 6, and of R. Chiao and B. P. Stoicheff for the use of Fig. 9(B).

CHRONOLOGICAL BIBLIOGRAPHY

Wilson, E. B., Decius, J. C., and Cross, P. C., *Molecular Vibrations, Theory of Infrared and Raman Vibration Spectra*. New York: McGraw-Hill Book Co., Inc., 1955 (for a discussion of Raman effect).

Manley, J. M., and Rowe, H. E., "Some General Properties of Nonlinear Elements—I. General Energy Relations," *Proc. IRE*, vol. 44, July 1956, pp. 904–913.

Jenkins, F. A., and White, H. E., *Fundamentals of Optics*. New York: McGraw-Hill Book Co., Inc., 1957 (for a discussion of Fabry–Perot etalon).

Weiss, M. T., "Quantum Derivation of Energy Relations Analogous to Those for Nonlinear Reactances," *Proc. IRE*, vol. 45, July 1957, pp. 1012–1013.

Rowe, H. E., "Some General Properties of Nonlinear Elements—II. Small Signal Theory," *Ibid.*, vol. 46, May 1958, pp. 850–860.

Tien, P. K., "Parametric Amplification and Frequency Mixing in Propagating Circuits," *J. Appl. Phys.*, vol. 29, Sept. 1958, pp. 1347–1357.

Franken, P. A., Hill, A. E., Peters, C. W., and Weinreich, G., "Generation of Optical Harmonics," *Phys. Rev. Letters*, vol. 7, Aug. 15, 1961, pp. 118–119.

Bass, M., Franken, P. A., Hill, A. E., Peters, C. W., and Weinreich, G., "Optical Mixing," *Ibid.*, vol. 8, Jan. 1, 1962, p. 18.

Giordmaine, J. A., "Mixing of Light Beams in Crystals," *Ibid.*, pp. 19–20.

Kaminow, I. P., and Liu, Julia, "Propagation Characteristics of Partially Loaded Two-Conductor Transmission Line for Broadband Light Modulators," *Proc. IEEE*, vol. 51, Jan. 1963, pp. 132–136.

Peters, C. J., "Gigacycle Bandwidth Coherent Light Traveling-Wave Phase Modulator," *Ibid.*, pp. 147–153.

Kaminow, I. P., "Splitting of Fabry–Perot Rings by Microwave Modulation of Light," *Appl. Phys. Letters*, vol. 2, Jan. 15, 1963, pp. 41–42.

Terhune, R. W., Maker, P. D., and Savage, C. M., "Observation of Saturation Effects in Optical Harmonic Generation," *Ibid.*, Feb. 1, 1963, pp. 54–55 (for a discussion of third-harmonic experiments).

Smith, A. W., and Braslau, N., "Observation of an Optical Difference Frequency," *J. Appl. Phys.*, vol. 34, July 1963, pp. 2105–2106.

Garmire, E., Pandarese, F., and Townes, C. H., "Coherently Driven Molecular Vibrations and Light Modulation," *Phys. Rev. Letters*, vol. 11, Aug. 15, 1963, pp. 160–163.

Zeiger, H. J., Tannenwald, P. E., Kern, S., and Herendeen, R., "Two-Step Raman Scattering in Nitrobenzene," *Ibid.*, Nov. 1, 1963, pp. 419–422.

Chiao, R., and Stoicheff, B. P., "Angular Dependence of Maser-Stimulated Raman Radiation in Calcite," *Ibid.*, vol. 12, Mar. 16, 1964, p. 290.

Davis, L. W., McCall, S. L., and Rodgers, A. P., "Raman Maser Study of Optical Difference Frequency Production," *J. Appl. Phys.*, vol. 35, Aug. 1964, pp. 2289–2290.

Bloembergen, N., and Shen, Y. R., "Coupling Between Vibrations and Light Waves in Raman Laser Media," *Phys. Rev. Letters*, vol. 12, May 4, 1964, p. 504.

Bloembergen, N., and Shen, Y. R., "Multimode Effects in Stimulated Raman Emission," *Ibid.*, vol. 13, Dec. 4, 1964, p. 720.

A Philoſophicall DISCOURSE Concerning SPEECH, Conformable to the *CARTESIAN PRINCIPLES.* Dedicated to The Moſt Chriſtian King.

Engliſhed out of *French.*

In the *SAVOY*,
Printed for *John Martin*, Printer to the *Royal ſociety*, and are to be ſold at the *Bell*, a little without *Temple-Bar*, 1668.

Machine recognition of human language

Part II—Theoretical models of speech perception and language

Men of science and philosophy have always shaped out theoretical models purporting to explain natural events. Today, the principle of invariance, fathered by the physical sciences, looms large in the new models of human speech and human language

Nilo Lindgren *Staff Writer*

Lastly, I am to take notice, that there is so great a communication and correspondency between the Nerves of the Ear, and those of the *Larynx*, that whensoever any sound agitates the Brain, there flow immediately spirits towards the Muscles of the *Larynx*, which duely dispose them to form a sound altogether like that, which was just now striking the Brain. And although I well conceive, that there needs some *time* to facilitate those motions of the Muscles of the Throat, so that the Sounds, which excite the Brain the first time, cannot be easily expressed by the Throat, yet not withstanding I doe as well conceive, that by virtue of repeating them it will come to pass, that the Brain, which thereby is often shaken in the same places, sends such a plenty of spirits through the nerves, that are inserted in the Muscles of the Throat, that at length they easily move all the cartilages, which serve for that action, as tis requisite they should be moved to form Sounds like those, that have shaken the Brain.

Fig. 1. This antique view of speech events, "Englished" in 1668, bears a resemblance to recent speech models.

By the beginning of this decade, the foundations of an acoustic theory of speech production had been firmly established, notably in the work of Dr. C. Gunnar M. Fant, Director of the Speech Transmission Laboratory at the Royal Institute of Technology in Stockholm. His book,[1] published in 1960, contains a detailed mathematical account of the acoustic theory of speech production, correlating vocal tract configurations, speech sounds, formant frequency patterns, and speech wave and resonator analyses treated on the basis of equivalent circuit theory. This theory, however, deals mainly with *static* aspects of speech production.

In the past few years, then, there has been a discernible trend toward a deeper examination of the *dynamic* properties of the articulatory system. Although a dynamic theory of speech production has yet to be thoroughly formulated and quantified, the work has begun. Such a theory, it is said, would need to "describe and predict articulatory movement in quantitative terms and elucidate the various modifications that speech sounds suffer in context."[2] This would mean a "return to some of the problems that were once of interest to phoneticians prior to the advent of the sound spectrograph and a revival of a genetical and physiological orientation in experimental work."[2] Reports of many such experimental ventures have appeared in the past few years.[3–5] These include acoustical studies of modifications in certain speech sounds in connected speech uttered at different speeds,[2] and the development of much more sophisticated dynamic models of the vocal tract,[6] which utilize, among other things, information gained from X-ray films (cineradiography) of vocal tract activity during speech and high-speed films of vocal cord action.[7] Electromyographic studies (measures of voltage patterns in the tongue and other muscles through means of surface recording electrodes) have also begun.[8,9]

Another aspect of recent research that must also be taken into account concerns the recent neurophysiological studies on the sensory systems of lower animals.[10–12] Studies on the visual and auditory systems of the frog

suggest that much more information processing or feature abstraction goes on at the peripheral levels of the sensory system than had previously been supposed.[13] Comparative studies between visual and auditory recognition systems have become valuable aids in the construction of models of perception. For instance, the striking model of the frog's visual perception system, put forward in 1959 by Lettvin[14] and his colleagues at M.I.T., stimulated other investigators to search for similar possibilities in the auditory perceptual system.

Inasmuch as recent models relate the speech-production and speech-perception systems quite closely, it is conceivable that the experimental findings on either system could contribute importantly toward "cracking" the nucleus of the human speech recognition process.

Still another area of experiment that has grown more germane to the problem of automatic speech recognition are the psycholinguistic studies, which have begun to give new insights into how humans process speech.

Having examined the processes whereby humans produce and audit speech, we must at last confront the problem of language. As we have been stating in many different ways, if engineers are to take seriously the idea of building automatic speech recognition machines, they can no longer avoid the questions related to language itself. Studies of the statistics of language have been made, and are being made—but beyond these are the findings and the formal models of language being postulated in recent linguistic theory.[15] Since some observers place the resolution of speech recognition problems squarely on the determination of the laws of language, the direction of modern linguistic studies (which constitute essentially a new, and some say revolutionary, approach to language) must at least be entertained, although no full description should be expected here.

In all of the studies discussed, we should note, we shall have moved well beyond the level of "mere" acoustics. In these recent studies, we have become involved in not only linguistic (or language) context, but in *situational* context as well. In fact, some recent studies have made particularly significant progress through a strategy of neurophysiological experiments correlated with behavioral experiments,[16] and it may well be that workers involved in other frontiers of research will have something to learn from such a strategy.

Taken together, the varied endeavors now going on confront us with a rich and complex field that we could not hope to cover definitively in one minor survey. It may prove most rewarding, then, to turn our major attention to the theoretical models that illuminate the specific works. We shall begin with a consideration of some of the theoretical models of speech perception and speech production.

Models of speech perception

Because there are so many curious, and really quite amazing, speech-perception phenomena that are unexplained, and in part because machine recognition performance thus far is so dismally low in contrast to human performance, many new theories purporting to clarify aspects of the human perceptual system have been proposed.

Theories of speech perception and speech production are by no means new; read, for instance, G. de Cordemoy's views (1668), reproduced in Fig. 1. Nor are the new theories in many cases such radical departures from the old. What *is* new is the fact that modern technology has given speech researchers such powerful new tools and capabilities for experimentally checking out their theories.

In fact, the incentive for building many speech recognition machines or models often comes from the desire to understand human verbal behavior and to test theories, rather than to imitate such human processes in machines intended for practical purposes. Nonetheless, we can be pretty sure, based on the experience of our period of research and development, that whatever the motives, these machines will eventuate in some form of useful devices.

Apart from such considerations, these hypothetical models of speech perception and speech production should be of interest to both designers of automata and students of perception because they are the springboards for many recent specific research projects and set off their significance.

Although no one can doubt the power of theoretical models, such models can also, of course, be seductive and misleading. It might be well, then, to go forewarned by an observation of Gunnar Fant: "I feel a little strange and out of place," he said (while discussing theoretical models), "I am used to discussing experiments, measurements, results; I am not so used to discussing *fiction*."

On the motor theory. One of the most influential, and most debated, of recent theories has been the motor theory of speech perception, put forward in perhaps its purest form by A. M. Liberman and his colleagues at the Haskins Laboratories.[17,18] Quite briefly, this theory supposes, on the evidence of perceptual studies, that a human perceives speech sounds (at one stage) by reference to the articulatory movements he knows are necessary to produce those sounds. Thus, for instance, it is assumed that part of a child's learning to recognize speech comes from his own training, his mimicry, in producing speech. Such referral to the articulatory system, and to the motor commands that actuate the articulators, might consist only of some form of neurophysiological sampling, or feedback, but the idea that some such stage of information processing forms part of the human speech recognition system is evidently quite compelling to some thinkers on the subject.

In part, researchers have been led to the motor theory through experiments that demonstrate that the human perception of speech is highly distinctive, and that such distinctiveness is not inherent or wholly attributable to the acoustic signal alone. Rather, it seems that when sounds are heard *as* speech sounds they then engage a speech perceptual system that heightens or sharpens the distinctiveness of the incoming sounds. Thus, it is speculated, the incoming signal stimulates some kind of *active* perceptual system that in itself gives the incoming signal this added distinctive quality. It is further speculated that this "capability" on the part of the human listener comes from his long linguistic experience, that it is a result of learning, "namely, that the perception of speech is tightly linked to the feedback from the speaker's own articulatory movements."[17] The theory clearly derives its name from the view that the task of perception engages the motor centers.

Part of the evidence for this view comes from experiments that indicate that some of the consonants are perceived *categorically*, that is, in absolute terms, rather than relativistically to other stimuli. "The impressionistic

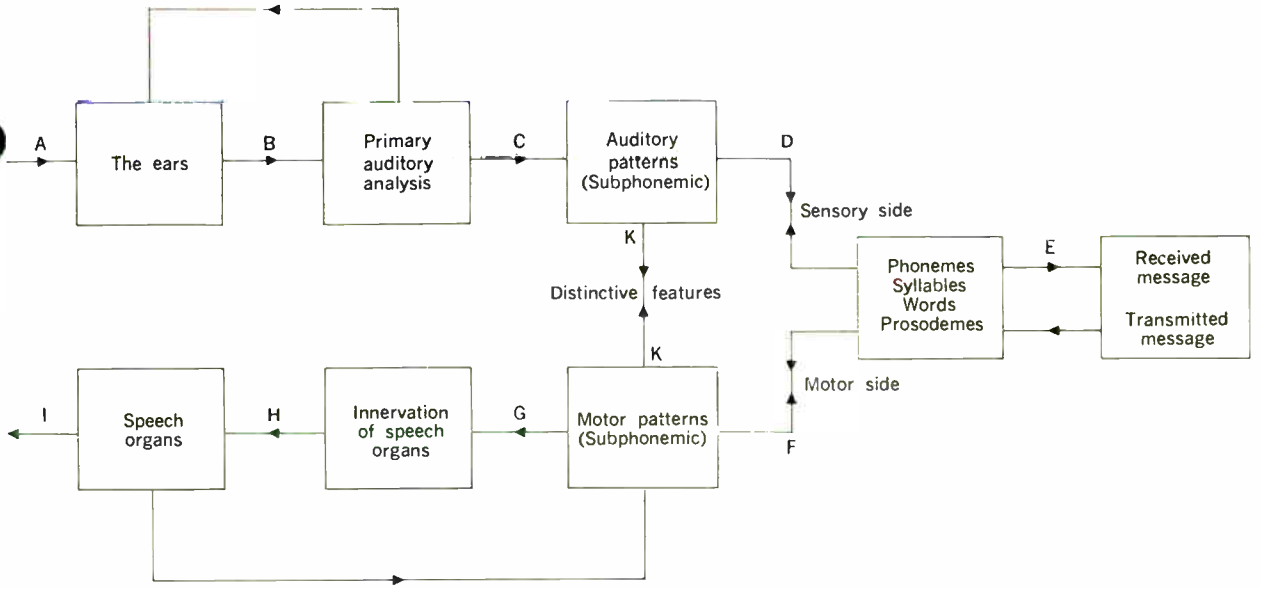


Fig. 2. Hypothetical model of brain functions in speech perception and production.

data, in a typical case, are to this effect: when we listen to a series of synthetic speech sounds in which the second-formant transition is varied progressively in such a way as to produce in succession /b/, /d/, and /g/, we do not hear a gradual change corresponding to the gradually changing stimulus; rather, we hear the first three or four stimuli as identical /b/'s, then, very abruptly with the next stimulus, the perception is of /d/, where it remains essentially unchanged until, again abruptly, it shifts to /g/."[17]

Thus, these stop consonants lie on an acoustic continuum, but they are perceived with sharp discontinuities. (This kind of sharp contrast in perception is reminiscent of the action of certain "crab's eye" and other electronic devices, containing many parallel input channels built up with artificial neurons, whose cross-influence is such as to sharpen the contrast between adjacent channels.[19])

However, it was found that in contrast to the categorical perception of stops, the perception of synthetic vowels was continuous. For the automata designers, this contrast is most interesting. As we have seen in Part I, it is the vowels that thus far have proved to be most easily discriminated by machines. But as Liberman points out:

"Although speech is highly distinctive in human perception, no machine has yet been designed which finds it so. Indeed, it may well be true—and, if so, ironic—that machines will have their greatest difficulty with those very phonemes (e.g., the stop and nasal consonants) which are for human beings most highly distinctive and which probably carry the heaviest load of information."[17]

Because perception seems to follow articulation rather than sound, the speculation arose that the relation between phoneme and articulation might be more nearly one-to-one than the relation between phoneme and acoustic signal. This speculation led rather naturally to experiments on the voltage patterns of the tongue and other muscles, using specially designed surface electrodes with a fairly high degree of resolution. These electromyographic studies, carried out chiefly by Katherine S. Harris, Peter MacNeilage[8,9] and others at Haskins, although not yet definitive, appear to bear out the motor hypothesis to some extent. For instance, recordings made for /f/ sounds in different contexts showed that although the acoustic signal changed from one context to the next, the electromyographic signals were virtually identical.

These studies, however, have a long way to go, and it may yet develop that the relation between phonemic structuring and neurophysiological command messages may not be as "simple" as is hoped.

Other types of experiments have also recently been performed in an effort to bear out the motor theory. For instance, Peter Denes, of the Bell Telephone Laboratories, recently carried out an experiment based on a corollary of the theory, namely, that part of the process of learning to recognize speech is training in producing speech ourselves. He tried to get a measure of how the ability of hearing one's own voice affected the learning process, the supposition being that when one hears one's voice one can associate articulatory movements with the sounds produced, so that learning should go faster. Although the results of this experiment were not conclusive, they did not tend to support the motor theory.[20]

In any event, the motor theory school of thinking has dominated current trends in speech research, and has proved fruitful in that it has provided a new organizational principle for correlating acoustic and articulatory information. However, as we have indicated, there has been considerable controversy about the motor theory. With respect to it, Gunnar Fant said recently:

"My own opinion would be that the speaking ability is not a necessary requirement for the perception of speech but that it enters as a conditioning factor. I would guess, though, that the importance of the speaking capacity for the development of speech perception is less than the importance of the hearing capacity for the development of normal speech. Children learn to understand speech before they talk and people born with complete lack of hearing have great difficulties in learning to speak."[21]

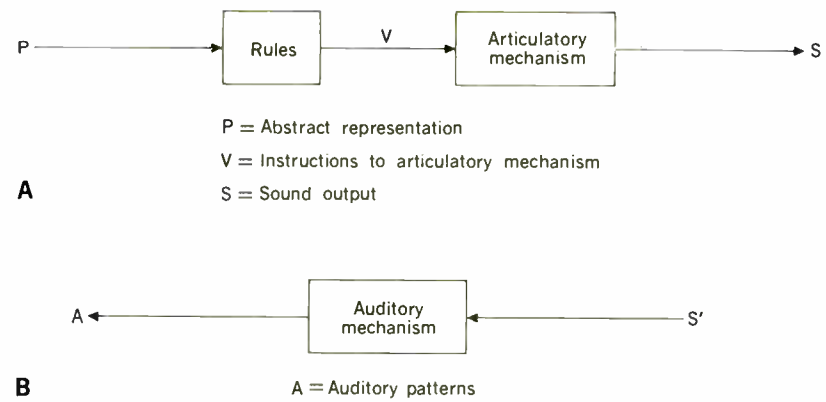
Fant then proposed a speculative model of the speech perceptive and productive functions; see Fig. 2. This model is useful for a comparative discussion of the most recent models that have been put forward. A common principle of the various models, Fant said, is that motor and sensory centers become more and more involved as one moves inward, from the level of the ears and speech organs, and approaches the central stages.

Apparently, it is the extent and the functional role of this postulated interconnection between the sensor and motor sides that has been the focus of most discussion and debate about the motor theory. In the Haskins picture of the inner events, according to Fant, the criteria of recognition are established by reference to the set of articulatory (motor) commands the person would associate with perceived sound patterns. "Once the correct association is set up, the decoding could proceed along the branch KFE (in Fig. 2) or directly along CDE, if box CD contains an image of the neural motor patterns stored in box GF."

Fant goes on: "From my point of view, the capacity of perceiving distinctive auditory patterns on the subphonemic level would be developed in the early learning process, prior to and not critically dependent on corresponding motor patterns. However, because of the cause-to-effect relation between articulation, speech wave, and perception, the inventory of auditory patterns would be structured rather similar to the patterns of motor commands. Each gross category of articulatory or phonatory events on the distinctive feature level in box GF would be paralleled by a corresponding auditory pattern in box CD."

Another question Fant brings up reflects somewhat on how the motor theory is interpreted. As recent psycholinguistic experiments have shown, and as Fant points out, a listener makes running predictions (based on his experience with language, with the person he is auditing, with the subject, and so on) about what he is about to hear. The probable function of such running predictions is to limit the range of search before the incoming sounds are identified. (In a machine, some such running prediction would presumably be included as part of the recognition strategy.) The extent to which such a predicting function engages the speech motor centers—whether along loop EFKDE or EFDE—and how far down toward the peripheral end of the system the motor activity mediates the sensory discriminations, says Fant, is (today) a pertinent problem. Such motor activity extending into the peripheral auditory regions "need not imply an active engagement of motor centers in auditory discriminations but merely the back scatter of neural activity along the path KGHI or FGHI, or both."[21]

By way of pursuing these speculations further, we may find it useful, for our purposes, to resurface and take a fresh tack with another model, which is also a kind of motor theory.



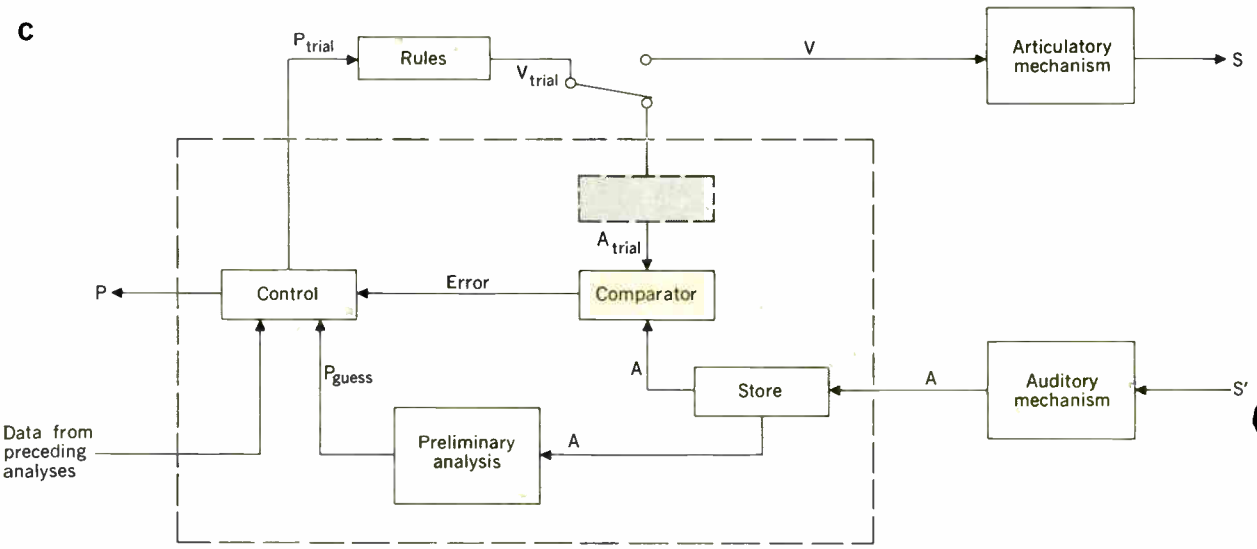


Fig. 3. Analysis-by-synthesis model. A—In the speech-production system, the abstract representation of an utterance is in matrix form P, which is operated on by a set of rules to yield the instructions V, which then signal the articulatory mechanism to voice the sounds S. B—In the perceptual system, sounds S′ are transformed into A patterns which are said to be similar to V patterns. C—The complete model showing the ties between the speech perception and production systems.

Analysis-by-synthesis model. A model of speech perception called analysis-by-synthesis was proposed, in 1958, by K. N. Stevens and M. Halle of M.I.T. (It is said to be a revival of a suggestion made by Wilhelm von Humboldt over a century ago.) They postulated that in perceiving speech, a listener's brain synthesizes patterns by following certain rules and then matches these internally generated patterns against the incoming patterns. These internal rules (generative rules), Stevens and Halle suggested, are largely identical to rules used in speech production. Fundamental to both the speech-perception and speech-production processes, they claimed, was an abstract representation of the speech events.

This model of speech perception has undergone modifications since it was first proposed.[22] The evolution of the model to an even more sophisticated form was described by Stevens and Halle only a few months ago.[23] Because this model (and the program for automatic speech recognition that flows from it) is regarded by most specialists in this field as one of the most promising and comprehensive of all such efforts, we should consider it in some detail.

The Stevens–Halle view of the speech-production process is summarized. "The speaker of a language has stored in his memory an abstract representation of lexical items of the language. These representations consist of segments which themselves are complexes of features. The representations can, therefore, be pictured as two-dimensional matrices in which the columns represent the string of segments and the rows different features.* The syntactic component of the language establishes the order in which different lexical items appear in the sentence as well as the relationship between the items in the string. The string of matrices is then operated on by a set of rules, which tranforms the matrices into sets of instructions for vocal-tract behavior. Ultimately the execution of these instructions produces the acoustical signal. While these instructions bear a close relationship to the articulatory behavior they elicit and to the acoustical properties of the utterances produced by it, there is no implication that the acoustic output must be necessarily decomposable into sequences of discrete segments or that instructions or features are directly recoverable from the signal. We *are* asserting that the acoustical output is a joint function of the abstract representation, the rules of the language, and the dynamics of the vocal tract, but we do not mean to imply that this is a linear function of the segments in the abstract representation; nor is there any reason to suppose that it must be so. As a result it cannot be expected that the underlying representation should in all cases be recoverable from the signal by simple techniques of signal analysis."[23]

How then do they view the process of speech perception? The input acoustic signal, they say, is decoded into an abstract representation identical to that employed in speech production—a representation in terms of segments and features. Moreover, they argue that when a listener decodes a signal into such a representation, he employs the same phonological rules that are used in generating the speech signal from the abstract representation.

They argue further: "That such a representation plays a role in speech perception is strongly suggested by the fact that normal speakers of a language understand without apparent difficulty utterances in a fairly wide range of dialects. This obvious fact would be quite inexplicable if we assumed that utterances are identified solely by means of direct classification of successive segments of the acoustical signal or by comparison of the input signal against an inventory of stored patterns, for the dialectical differences that do not inhibit comprehension are precisely differences in the inventory of the sounds. If, on the other hand, we assume that dialectical differences in the sounds are due to the fact that a given abstract representation of a speech event is actualized in accordance with *different* phonological rules, then the performance of the normal speakers becomes at once understandable. Having listened to a relatively small sample of utterances in a dialect different from his own, the speaker of a language is evidently able to determine the modifications of a few phonological rules of the dialect as compared with those in his own dialect. He is then able to utilize these rules to identify correctly combinations of elements or words he has never heard before in that dialect."[23]

The reader may be interested in comparing and correlating this explicit view with the "vowel-loop" theory of how a listener tunes in on a new speaker, discussed in Part I, which is a less formalized way of looking at the question.

What do these speech-production and -perception processes, as envisioned by Halle and Stevens, look like schematically? Figure 3(A) represents the speech-production system, in which P is the abstract representation of an utterance in matrix form. A set of rules operates on P to yield the instructions V. These instructions are visualized as "patterns that can be actualized in the form of appropriate sequences of motor commands only after certain motor skills have been acquired." That is, as an individual learns the language of his community, he makes observations on how the others, the fluent members, talk, and he attempts to turn these observations into motor commands; but he needs practice before he gets the right assemblage of command signals.

Figure 3(B) shows the auditory mechanism in the same terms, where an input sound S' is processed in some fairly complex fashion to yield auditory patterns A. These auditory patterns are said to have close ties with articulatory activity.

The transformation of S' into A was at one time viewed as being akin to the kind of simple frequency analysis performed by a sound spectrograph, but neurophysiological studies in the past few years on the visual and auditory systems of lower animals suggest that the feature-abstracting capabilities of the nerve systems at the periphery are powerful indeed, and evidently very selective.[12] In any event, Stevens and Halle suggest in their model that the V patterns and the A patterns have very similar properties, and that speech is "anchored equally in the motor and in the auditory system of man."

What, then, goes on beyond the level of the A patterns, that is, how are the A patterns transformed into the abstract representation P? The complete and most recent analysis-by-synthesis model proposed by Stevens and Halle, incorporating the elements already discussed, appears in Fig. 3(C). The box marked "Store" within the dotted section is a short-term memory where the auditory patterns A may be stored.

The authors of this model emphasize that it is a model

* Distinctive features, Table II, are discussed on p. 54.

for the perception of speech, not a model for auditory perception in general. Their point is that the speech-perception system is a special system that a listener switches in when he knows he is hearing speech sounds. (On this point, one should also see the work of A. S. House and others.[24])

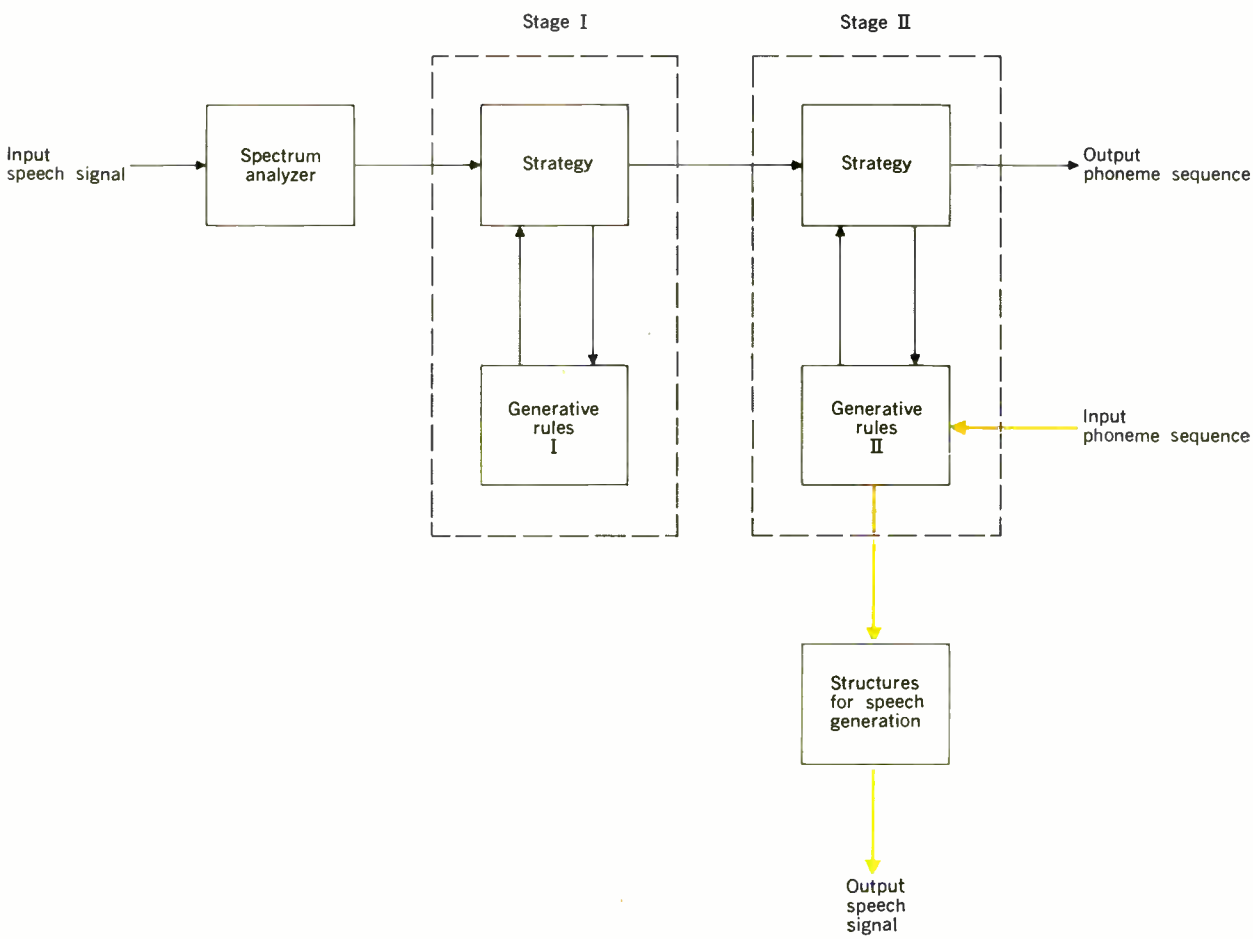
The subjective impression of this hypothetical event is certainly reflected in the cliché, "he pricks up his ears." A person suddenly realizes he is hearing speech, not just noise, and he pays attention. In this connection, it is interesting to draw into the same orbit of speculation some remarks of H. L. Cramer[25] at Harvard, who has been experimenting with the intelligibility of compressed speech. One method of compressing speech is to take tapes of running speech, and cut out varying percentages of the content of these tapes through the simple device of cutting out short alternate segments and piecing the remainder together. Up to the point where the result is still intelligible speech, Cramer reports, subjectively the listener discovers that his attention is uncannily "riveted" to the incoming message, as if the speech-information processing system were being galvanized to its utmost limits.

In any event, Stevens and Halle suggest that when a listener is operating in the speech-perception mode, he is always carrying out a categorization of the input; and, in this sense, their views and those of Liberman, discussed earlier, are in accord.

Stevens and Halle describe the operation of their model, Fig. 3(C), as follows: The auditory patterns resulting from the acoustic signal at the ears undergo a preliminary analysis. The output of this analysis, together with contextual information derived from adjacent elements of the signal, allows a guess to be made in the control section about what the P pattern is likely to be. (Most automatic speech recognition systems, discussed in Part I, dealing with the acoustic level in association with some local context, can be viewed as being at the level of making this first guess.) This P_{trial} is subjected to the phonological rules (the same ones used in speech generation) to produce the pattern V_{trial}. This is the pattern that would normally give rise to motor commands to produce speech, but during speech perception, it is hypothesized, the V path is inhibited, and instead V_{trial} gives rise to an equivalent auditory pattern in the shaded box marked A_{trial}. This synthesized pattern is then compared with the pattern under analysis. When the patterns agree, the pattern or sequence of units goes on for processing at deeper levels; when they disagree, the control section notes the error, makes a new guess, and so on, until the correct P is established.

The authors of the perceptual model suggest that this matching process always be employed as a check, even

Fig. 4. Two-stage scheme for processing speech on the nonsense-syllable level. The input speech signal is reduced first to a set of quasi-continuous phonetic parameters, which are further reduced to yield a sequence of discrete phonemes. Heavy tinted arrows indicate how a speech signal would be generated from phonemes.



when the initial guess is correct. Their point is that normal speech perception always involves active participation of the listener, as opposed to the notion that listening is somehow a passive process. They also point out the relationship of this analysis-by-synthesis model to the models developed by D. M. MacKay,[26] and others,[27] in which it is suggested that all types of perception involve an internal replication and matching process. MacKay has also developed the very interesting idea that perception is a process of "updating," that the living system engages in this "synthetic activity" of constantly keeping its internal map up to date vis-à-vis the environment.[28]

Those who have studied previous models offered by Stevens and Halle will note some differences in the details of this newest model. But, it seems, the basic principle, of how speech perception and speech production function, is in no way abandoned. Even apart from this, whether or not this model truly describes what happens in perception, whether or not new experimental evidence should strengthen or weaken it, the application of this kind of model in a speech recognition machine should not be vitally affected. Let us, then, take a look at the physical "issue" of this theoretical model.

Automatic speech recognition scheme

Up to this point, we have been following a description of a hypothetical model of the *human* speech-perception and speech-production systems. An automatic speech recognition scheme capable of processing any but the most trivial classes of utterances, Halle and Stevens assert, must incorporate all of the features of that model. In such a scheme, the speech inputs are transformed into a sequence of phonemes through an active or feedback process.

The basic skeleton of their proposed automatic recognition scheme includes, in various forms, the following: "the input signal must be matched against a comparison signal; a set of generative rules must be stored within the machine; preliminary analysis must be performed; and a strategy included to control the order in which internal comparison signals are to be generated."[22]

As has been pointed out, there is no one-to-one relation between the acoustic segments of an utterance and the phonemes. For those who analyze speech, and who have attempted to design speech recognition machines, this has raised the so-called segmentation problem. As K. N. Stevens[29] succinctly puts it, "the segmentation problem is the fact that you can't segment." The proposed recognition scheme attempts to transform continuously changing speech signals into discrete phonemic outputs "without depending crucially on segmentation."[22]

In its 1962 elaboration, the scheme consists of two major stages of speech processing, that is, the incoming speech undergoes reduction in two analysis-by-synthesis loops, each equipped with different sets of generative rules.

Figure 4 shows the two-stage scheme in its barest form. Each box marked "Strategy" would perform the functions incorporated within the large dotted box of Fig. 3(C). "In the first stage the spectral representation is reduced to a set of parameters which describe the pertinent motions and excitations of the vocal tract, that is, the phonetic parameters."[22] Elsewhere, it has been hypothesized that for a given speech sound the neurophysiological instructions or nerve excitations assume a set of steady values, jumping discontinuously to a new set of values to produce a subsequent speech sound. This is to say that at this level there is more of a one-to-one correspondence between the discrete phonemes and the electrical instructions to the vocal tract that gives the phonemes voice.

In the second stage, the phonetic parameters are transformed (again through a separate strategy, and embodying generative rules made up of several distinct parts) to a sequence of discrete phonemes.

"These steps," say the authors, "provide a natural division of the analysis procedure into one part concerned primarily with the physical and physiological processes of speech, and the other concerned with those aspects of speech primarily dependent on linguistic and social factors. In the first stage, variance in the signal due to differences in the speech mechanism of different talkers (or of a given talker in different situations) would be largely eliminated. The second stage would account for influences such as rate of talking, linguistic background or dialect of the talker, and contextual variants of phonemes."[22]

However, even this two-stage scheme can handle only speech sounds in the category of nonsense syllables. If words, phrases, or continuous natural speech are to be processed, then still further stages must be added that will take into account higher-order linguistic constraints (that is, generative rules on the level of syntax and beyond).

The importance of the generative rules in the scheme should perhaps be stressed even further. In earlier techniques of speech processing, as discussed in Part I, the recognition devices were equipped with what were essentially dictionaries of spectral characteristics for very limited vocabularies. For larger vocabularies, the dictionary would grow rapidly, and the physical implementation of the memory system thus would rapidly grow impractical. By replacing the dictionary by the sets of generative rules (that is, the principles of construction of the dictionary entries), the size of the memory is more restricted. (Note that the rules for speech synthesis assembled by Haskins, discussed in Part I, are an important contribution to one level of such a dictionary.) The time needed for matching internally generated spectra for comparison with incoming spectra would still be long in this scheme, but, it is argued, the use of the preliminary analysis, or first guess, should restrict the range to a small number of possibilities. Psycholinguistic experiments on how humans use language, and studies of language statistics, made in the past few years (discussed later), certainly lend strong support to this argument.

The two-stage scheme of Fig. 4 also shows peripheral speech generating structures. With this addition, the model is capable of both speech recognition and speech production.

Although we have been obliged to omit extensive details of this scheme, it should be clear, hopefully, that we are in the presence here of a most comprehensive, ambitious, and long-term program of research. Halle and Stevens say that certain components of both stages could be designed from present knowledge, but that considerable research remains before the system can be designed to function as a whole.

Very broadly, perhaps it will not be too wrong to say that quite a bit is now known about the "static" modes of

speech production, and that recent research has been looking into dynamic physiological factors.

Articulatory analog

One of the major preoccupations of the M.I.T. speech analysis and synthesis research program has been with the generative rules for speech, the process whereby, it is postulated, a human talker encodes a sequence or string of discrete linguistic symbols (phonemes) into an acoustic signal, that is, into speech that can be understood or decoded by another member of the same linguistic community. Another general objective of the M.I.T. research program is to further the understanding of the decoding process, whereby a human listener can decode an acoustic speech signal into a sequence of discrete linguistic symbols. These two general, related questions have been studied on a number of levels.

A general experimental approach to the study of the human speech generation processes, and to the determination of the generative rules, has been through the use of articulatory analogs.[6]

The synthesis of speech as a method of study is hardly new. It goes back at least two centuries to 1791, when Wolfgang de Kempelen constructed a mechanical device that synthesized many vowels and consonants. In this century, Voders and vocoders and the like have been used for speech synthesis. And in Part I, we have already seen how the use of the pattern playback for synthesizing speech has contributed so importantly to the determination of the acoustic cues for speech.

The use of vocal tract analogs allows for deeper studies of anatomical and physiological aspects of speech. In 1950, H. K. Dunn demonstrated an electric vocal tract. It had three controls, one for the position of the tongue hump, another for the magnitude of the tongue hump, and a third for the magnitude of the lip constriction. However, this electrical analog was limited in various respects.

The first M.I.T. (Research Laboratory of Electronics) electrical analog (1953) was an effort to develop a much more flexible representation. The human vocal tract may be viewed as an acoustic tube with vocal cords at the lower end (the glottis) and lips at the other end, with everything in between being possessed of varying degrees of variability. Different speech sounds are produced by this "tube" by controlled variations of its walls, by the nature and position of one or more sources of excitation, by the glottis, or by turbulence along points of the tube. The differing shapes and sizes of the cavities of this tube produce different resonances. Theoretically, there are an infinite number of these resonances, but in actual speech, the energies of the formants higher than the second and third are rapidly damped out, and for practical research purposes are not considered.

An electrical analog of the tube was built, based on an idealization of the vocal tract as a cascade of 35 acoustic cylinders. Ideally, the dimensions and activity of each element of the vocal tract, from the larynx to the lips, ought to be known quantitatively for any human sounds so that the synthesizer can be adjusted to make corresponding sounds. Research of the past decade, along with the increasing complexity of the synthesizer, has been part of the effort to fill in these unknowns. Much has been learned about acoustic outputs and their relation to articulatory configurations.

The earliest M.I.T. articulatory analog was a static vocal tract that could produce "acceptable" vowels and some consonants. This analog was developed even further in 1960 into a lumped-parameter electric transmission-line synthesizer, a dynamic model that could be varied during an utterance to produce two-component syllables. A year later, in 1961, a dynamic nasal analog was added to the vocal-tract analog, so that it became possible to investigate the production of nasal consonants and nasalized vowels. The overall arrangement of the articulatory analog is shown in Fig. 5.

The addition of a pulse generator, followed by a filter, enabled the simulation of glottal excitation, and the addition of a noise generator led to the synthesis of consonants characterized by frication noise. By modulating the noise signal with glottal pulses, certain voiced consonants were made more natural. However, as experiments progressed, it became clear that even this rather elaborate analog speech synthesizer was too limited, chiefly because only simple speech utterances could be dealt with, and at the additional cost of manual setting of many articulatory parameters.

To overcome the manual control problem, the next step was to control the synthesizer with the M.I.T. TX-O digital computer, using as an interface a 32-channel digital-to-analog converter that could supply individual control voltages for each transmission line section, as well as control signals for glottal-pulse amplitude, noise amplitude, and nasal coupling. This setup has also exhibited certain shortcomings, problems of calibration and maintenance, so that a new dynamic speech synthesizer has been under design that features direct digital computer control of the individual synthesizer sections.

With this new facility, equipped with a programming system allowing flexible control of the articulatory analog, an investigator will be able to give his specifications for an utterance (e.g., a list of vocal events and the time at which they occur—start of voicing, duration of frication, timing of transitions, and so on), perform the synthesis, listen to the result, evaluate it, readjust his specifications, and make further trials as he wishes. Through this kind of flexible program of control of the parameters (as many as 40) of a synthetic utterance, the objective is to determine quantitatively the significant parameters of articulatory commands needed to make certain utterances.[6]

Most recently, the Stevens research group has been working to give its articulatory model even more human characteristics through direct observations of the human articulatory processes using photographic and cineradiographic techniques. Part of this work, still going on, is making X-ray movies of the human speaker. By plotting the movements of various sections of the human vocal tract, through assiduous frame-by-frame analyses, a series of "programs" are being prepared for the computer-controlled sections of the articulatory analog.

Figure 6 shows one of these X-ray frames taken from a movie. This lateral cineradiograph represents a mid-line section of a talker at rest, that is, the vocal tract is in the rest or breathing position. The overlay tracing calls out the parts more distinctly. Present studies using these techniques include analyses of the movements of the labeled structures and correlation of these movements with speech sounds.[30]

The M.I.T. researchers are also considering, at the

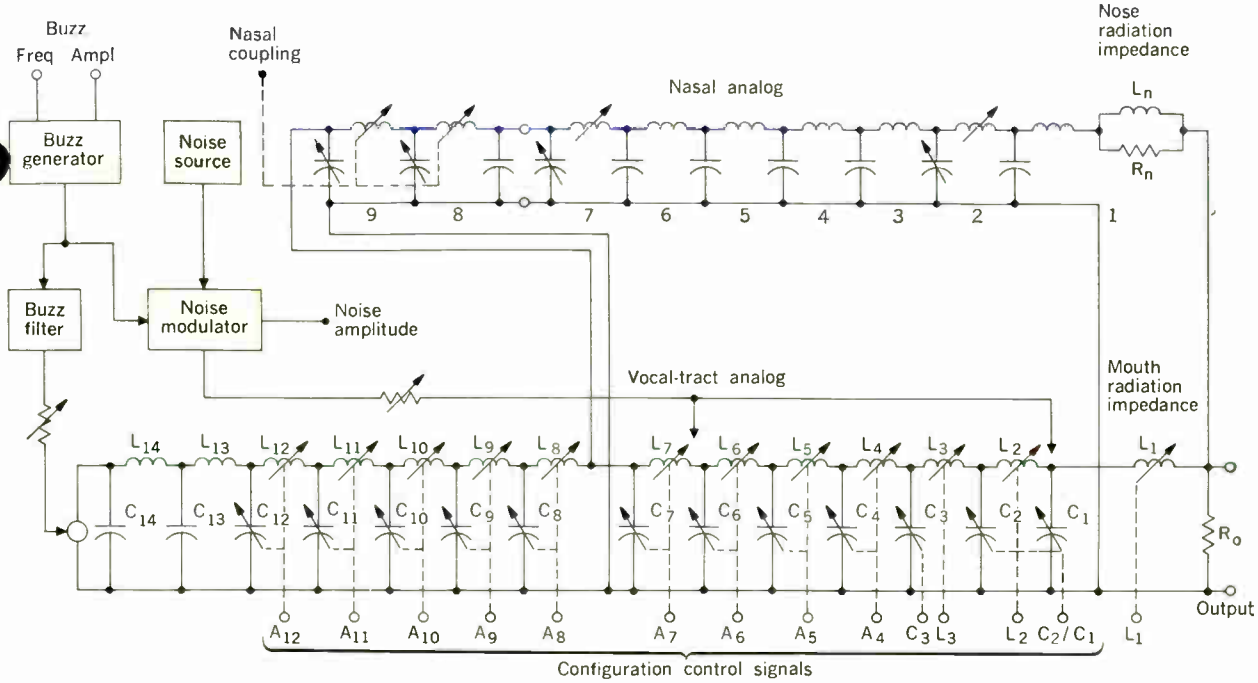
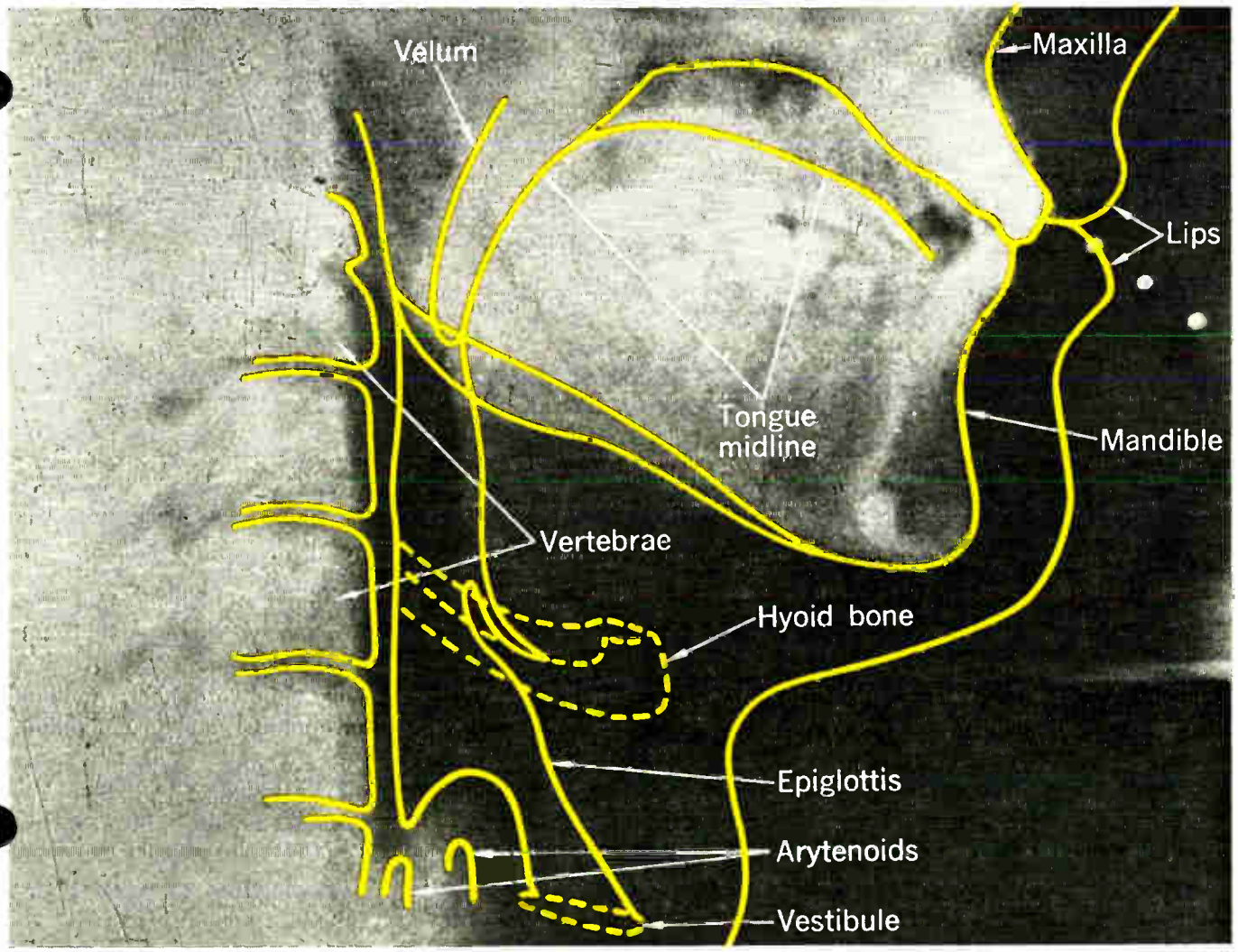


Fig. 5. Complete articulatory analog, consisting of 35 electrical sections.

Fig. 6. Simultaneous sound recordings and cineradiographic films obtained for a variety of utterances. Frame-by-frame tracings are made from the film, as shown here, to produce quantitative data on the movements of various speech generating structures. Lead pellets, shaped pieces of metal, and adhesive barium compound on the tongue are used to improve contrast and to indicate structural positions.



present time, cooperative research with Haskins Laboratories, incorporating some of the findings of their electromyographic studies in the dynamic articulatory analog.[29]

Results of speech-synthesis research

Over the past four years, numerous studies with speech synthesis have been carried out at M.I.T., with the aim of formulating and clarifying the rules that should govern the activities of the articulatory analog when sequences of input phonetic symbols are given to it.

The speech studies include experiments on the production of stop consonants in the initial position (by Fujimura in 1960, using two different synthesizers, a resonance synthesizer as well as the dynamic analog of the vocal tract), in which versions of /b/, /d/, and /g/ were produced in combination with the vowel /ɛ/. Detailed studies were made on how the rate of formant transition affected the output of syllable combinations made from these same consonants and vowel. During the same period, studies were made of the cues for the tense–lax opposition of stop consonants (roughly speaking, the lax consonant is voiced, the tense is voiceless). In 1961 and '62, following the development of the dynamic nasal analog, experiments were carried out by Hecker on nasal consonants and nasalized vowels. Certain timing experiments were also carried out, in which elements of the timing patterns were changed (rates of transition from one articulatory configuration to another, and duration of nasal coupling). These timing studies are viewed rather optimistically by the experimenters, for they gave results supporting their view that studies of the performance of an articulatory synthesizer can reveal important features of the production and perception of natural speech.

One upshot of this synthesizer work is that the gross features of the rules for most consonant and vowel phonemes of American English as they occur in monosyllabic CV (consonant–vowel) and VC syllables are largely known in a quantitative form, at least to a kind of first approximation. The refinement of these rules awaits the development of the more flexible digital control system already mentioned, as well as the results of the X-ray movies of the human vocal tract in action.

Some work was also done by the Stevens group on connected speech. The prosodic features, that is, those features which might be said to characterize natural speech and which are its most distinguishing marks—intonational patterns, stress, vowel durations, and so on—were not programmed by *rule*, however; rather, their specification was guided by a less formal study of sound spectrograms of utterances of a human talker. The Stevens group also programmed their synthesizer to produce a short song and two sentences, which together included all the consonantal phonemes of American English. Here again, naturalness was avoided by superimposing on the articulatory rules the information of a simple musical score, which gave the stress, duration, fundamental voicing frequency, and so on. The output was simplified even further by separating syllables, etc., to produce an infantile singing style.

Most of this research belongs to what Stevens and Halle have described as the first stage of analysis: devising a procedure for specifying in quantitative terms the phonetic parameters. The second task, according to them, is "to establish the generative rules describing the conversion of phonetic parameters to time-varying speech spectra."

Stevens and Halle view this conversion as consisting of several distinct parts. The generative rules must embody, in part, "the relation between what linguists have called a 'narrow phonetic transcription of an utterance' and its 'phonemic or morphophonemic transcription.' " They must also describe how to deal with phonetic parameters not governed by the language. And third, the rules must "specify the transformation from discrete to continuous signals that results from the inertia of the neural and muscular structures involved in speech production."[22]

The generative rules of speech, on which work began in earnest just in the past few years, must be described "precisely and exhaustively." Only then, Stevens and Halle say, will it become possible to evolve (using the powerful tools of mathematics) the optimal strategies that will be required by an automatic speech recognition machine based on the analysis-by-synthesis approach.[22]

The distinctive features

The distinctive features approach to language has already been mentioned in Part I. These features are the ultimate discrete oppositions by which linguistic signals (at any stage) are distinguished from one another.

In the distinctive features approach, one seems to take essentially a nuclearizing point of view on language—the phonemic molecules are separated, isolated, purified, and analyzed down to their particle constituents. And even these, as the analysis proceeds, seem to lose almost all physical or substantive feeling. At the distinctive features level, one seems to be looking at the binding forces: diffuse/compact, acute/grave, tense/lax, and so forth. At this level, one is examining the qualities that make one phoneme differ from another, one is seeing the unique oppositions that distinguish one sound from its successor, one is hearing the "otherness" of the next sound, and the next. And out of these elementary or fundamental linguistic "atoms" is built up, level by level, a world of language, bound by these forces, and operating evidently under precise laws. These "laws" are what the linguists seek to isolate, identify, and relate in a formal system.

In their 1952 exposition, "Preliminaries to Speech Analysis,"[31] Jakobson, Fant, and Halle illustrate the evolution of their concept of distinctive features. They start with some simple, one-syllable words—bill, pull, pill, bull—and they ask: how many significant units, i.e., units relevant for the discrimination of the samples, do the sound shapes of the samples contain? The words *bill* and *pull* are distinguishable by /bi/ and /pu/, which may be decomposed further; but *bill* and *pill* are differentiated by /b/ and /p/, which cannot be decomposed further, and the same is true for *bull* and *pull*.

Through sets of such systematic comparisons, whose aim is to isolate the *minimal* contrasts between word elements, these authors evolve the general rule that a listener must make a series of two-choice (binary) selections to differentiate words. Without going into the subtle detailed deductions and evidence on which this system of distinctive features is explicated, we shall present only their definition, and hope that at least the principle is apparent:

"Any minimal distinction carried by the message confronts the listener with a two-choice situation. Within a given language each of these oppositions has a specific

property which differentiates it from all the others. The listener is obliged to choose either between two polar qualities of the same category, such as grave vs. acute, compact vs. diffuse, or between the presence and absence of a certain quality, such as voiced vs. unvoiced, nasalized vs. non-nasalized, sharpened vs. non-sharpened (plain). The choice between the two opposites may be termed *distinctive feature*. The distinctive features are the ultimate distinctive entities of language since no one of them can be broken down into smaller linguistic units. The distinctive features combined into one simultaneous bundle form a *phoneme*."[31]

And further, they say: "Any one language code has a finite set of distinctive features and a finite set of rules for grouping them into phonemes and also for grouping the latter into sequences; this multiple set is termed *phonemic pattern*."

However, the "distinctive features and the phonemes possess no meaning of their own. Their only semantic load is to signalize that a morpheme (the smallest meaningful unit in language—e.g., a root, a prefix, and a suffix are morphemes) which, all other things being equal, exhibits an opposite feature is a different morpheme."

The question arises as to how many distinctive features there are. Comparative phonological studies (phonology being the study of how language utilizes sound matter) have steadily reduced the number of distinctive features in the languages of the world, i.e., "universals," to about 12 oppositions. That is, physiologically or biologically, or however you want to say it, the human animal is capable of producing only a dozen or so distinctive oppositions with which to build his speech. And out of this number (of universal oppositions), each language has made its selections; out of this number each language has developed its own sound pattern. Table I lists the 12 distinctive features classified by Jakobson and Halle,[32] along with their brief description of the *acoustic* characteristics of each opposition.

Table II, taken from the "Preliminaries,"[31] shows the English phonemes broken down into their distinctive features on a yes–no basis. The table shows, for example, that the vowel /o/ is vocalic, is nonconsonantal, is compact, is grave, and is flat. The table shows further that the bundle of features that characterizes the phoneme

I. Some characteristics of the distinctive features on the acoustic level

1. Vocalic/nonvocalic
Presence vs. absence of a sharply defined formant structure

2. Consonantal/nonconsonantal
Low vs. high total energy

3. Compact/diffuse
Higher vs. lower concentration of energy (intensity) in a relatively narrow, central region of the spectrum, accompanied by an increase (vs. decrease) of the total energy

4. Tense/lax
Higher vs. lower total energy in conjunction with a greater vs. smaller spread of the energy in the spectrum and in time

5. Voiced/voiceless
Presence vs. absence of periodic low-frequency excitation

6. Nasal/oral
Spreading the available energy over wider (vs. narrower) frequency regions by a reduction in the intensity of certain (primarily the first) formants and introduction of additional (nasal) formants

7. Discontinuous/continuant
Silence followed and/or preceded by spread of energy over a wide frequency region (either as burst or a rapid transition of vowel formants) vs. absence of abrupt transition between sound and such a silence

8. Strident/mellow
Higher intensity noise vs. lower intensity noise

9. Checked/unchecked
Higher rate of discharge of energy within a reduced interval of time vs. lower rate of discharge within a longer interval

10. Grave/acute
Concentration of energy in the lower (vs. upper) frequencies of the spectrum

11. Flat/plain
Flat phonemes in contradistinction to the corresponding plain ones are characterized by a downward shift or weakening of some of their upper-frequency components

12. Sharp/plain
Sharp phonemes in contradistinction to the corresponding plain ones are characterized by an upward shift of some of their upper frequency components

II. Distinctive-features pattern of English phonemes

	o	a	e	u	ɔ	i	l	ŋ	ʃ	ʧ	k	ʒ	ʤ	g	m	f	p	v	b	n	s	θ	t	z	ð	d	h	#
1. Vocalic/nonvocalic	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
2. Consonantal/non-consonantal	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−
3. Compact/diffuse	+	+	+	−	−	−		+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−		
4. Grave/acute	+	+	−	+	+	−									+	+	+	+	+	−	−	−	−	−	−	−		
5. Flat/plain	+	−		+	−																							
6. Nasal/oral								+	−	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−	−		
7. Tense/lax									+	+	+	−	−	−		+	+	−	−		+	+	+	−	−	−	+	−
8. Continuant/interrupted									+	−	−	+	−	−		+	−	+	−		+	+	−	+	+	−		
9. Strident/mellow									+	+	−	+	+	−							+	−	−	+	−	−		

Key to phonemic transcription: /o/—p<u>o</u>t, /a/—p<u>a</u>t; /e/—p<u>e</u>t, /u/—p<u>u</u>t, /ə/—p<u>u</u>tt, /i/—p<u>i</u>t, /l/—<u>l</u>ull, /ŋ/—lu<u>ng</u>, /ʃ/—<u>sh</u>ip, /ʧ/—<u>ch</u>ip, /k/—<u>k</u>ip, /ʒ/—a<u>z</u>ure, /ʤ/—<u>j</u>uice, /g/—<u>g</u>oose, /m/—<u>m</u>ill, /f/—<u>f</u>ill, /p/—<u>p</u>ill, /v/—<u>v</u>im, /b/—<u>b</u>ill, /n/—<u>n</u>il, /s/—<u>s</u>ill, /θ/—<u>th</u>ill, /t/—<u>t</u>ill, /z/—<u>z</u>ip, /ð/—<u>th</u>is, /d/—<u>d</u>ill, /h/—<u>h</u>ill, /#/—_ill.

/o/ differs from /a/ only in the flat/plain feature. Many phonemes are distinct from one another in only one feature. This fact makes it possible, from a constructive point of view, to lump certain phonemes into broader classes, since they share so many common features. But from the point of view of the automata builder, these single-feature distinctions, in league with the variability of human speakers, raise difficult problems of identification for machines; which is another way of looking at why a machine may be easily confused about which phoneme it is hearing.

In the Jakobson and Halle work,[32] published in 1956, the authors describe their view of how language units are constructed from the distinctive features:

"The distinctive features are aligned into simultaneous bundles called phonemes; phonemes are concatenated into sequences; the elementary pattern underlying any grouping of phonemes is the syllable. The phonemic structure of the syllable is determined by a set of rules and any sequence is based on the regular recurrence of this constructive model. A free form (a sequence, separable by means of pauses) must contain an integral number of syllables. Obviously, the number of different syllables in a language is a small submultiple of the number of free forms, just as the number of phonemes is a small submultiple of the number of syllables, and the number of distinctive features, a submultiple of the number of phonemes.

"The pivotal principle of syllable structure is the contrast of successive features within the syllable. One part of the syllable stands out from the others. It is mainly the contrast vowel vs. consonant which is used to render one part of the syllable more prominent. There are languages where every syllable consists of a consonant and succeeding vowel (CV): in such a case it is possible from any point of sequence to predict the class of phonemes that is to follow. In a language with a greater variety of syllable types, the recurrence of a phonemic class presents different degrees of probability. In addition to CV, other schemes may be used: CVC, V, VC. In contradistinction to C, the part V can neither be omitted, nor figure twice in the syllable."

In those two paragraphs lies a powerful rationale of the basis of language. The importance of this kind of formulation for a program of automatic speech recognition should be plain.

Besides the distinctive features, Jakobson, Fant, and Halle define other types of coded information-bearing features:[31] *configurative features*, which signal the division of an utterance into grammatical units of different degrees of complexity, particularly into sentences and words, either by singling out these units and indicating their hierarchy (*culminative features*) or by delimiting and integrating them (*demarcative features*). *Expressive features* (or emphatics) put the relative emphasis on different parts of the utterance or on different utterances and suggest the emotional attitudes of the utterer.

There are also redundant features that serve an important role in different situations. For instance, in some cases of shouting or whispering, the normally distinctive features no longer serve, and the redundant feature takes over the distinctive function.

These very brief excisions, taken from the works cited, hardly begin to exhaust the richness and interest of this conceptual structuring, but perhaps they will at least *signify* the direction such work is taking.

Factors of speech intelligibility

Thus far in this survey, we have been looking fairly exclusively at experiments and theoretical models that relate to linguistic events on the level of language "particles." The suggestion throughout, although not explicitly stated, has been to the effect that humans make decisions about what they are hearing at the lowest levels first (at the phonemic level), and then progressively make the decisions at higher levels (e.g., at the levels of syntax and semantics), or possibly on many levels simultaneously (see box at left). We should, therefore, turn our attention to some experimental findings on the nature of speech perception that come from somewhat another direction, and whose results are of a different character. These experiments do not deal so exclusively with the smallest linguistic particles, but rather with larger linguistic segments. We are therefore coming here to a more "naturalistic" scale, by which we mean a province of psycholinguistics where it is easier to exercise a layman's intuitive appreciation.

Some of the conclusions of these psycholinguistic experiments run like this:

Early experiments by George A. Miller, and others, at Harvard, showed that content words, such as nouns, verbs, and adjectives, were more intelligible when heard in the context of a grammatical sentence than when they were scrambled and spoken as items on a list.[33] Later, fuller experiments showed that, in general, intelligibility is highest for a human listener when he is given meaningful, grammatical sentences; intelligibility is lower for

On models of speech perception

Speech research, in general, has been influenced by this school of linguistics (called Taxonemic phonemics), which until recently was representative of virtually all American linguistics, to a far greater extent than is commonly supposed. There is a central underlying model of perception which is the basis of virtually all research on automatic speech recognition that comes from this school of linguistics. Perception is viewed as the concatenation of little chunks of information that are recognized in sequence from the acoustic signal. These "phonemes" form "morphemes" which in turn form words, clauses, and sentences. Perception goes in one way, so to speak. We recognize phonemes from phonetic, i.e., acoustic information. We then group the phonemes into morphemes and then into words, clauses, etc., into bigger and bigger units. However, all perception ultimately rests on the initial recognition of the acoustically defined phonemic elements. Much recent evidence suggests that this is not the case.—*Philip Lieberman*
Speech Research Branch
Air Force Cambridge Research Laboratories
December 24, 1964

semantically anomalous sentences (sentences that are grammatically correct but meaningless, such as "Gadgets kill passengers from the eyes"); and intelligibility is lowest of all for ungrammatical strings of words.[34] Other experiments showed that semantic and syntactic constraints also facilitate learning.[35] Repeatedly, and in various ways, it has been shown that expected words (high-probability words) are more intelligible (accuracy scores are higher) than unexpected words (low-probability words).[36]

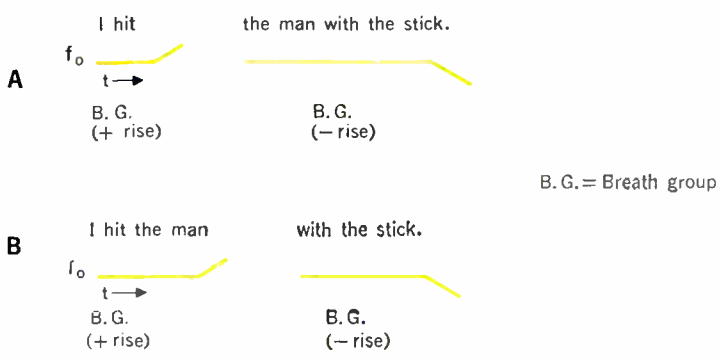
All these experiments indicate that a human listener somehow uses both syntactic and semantic rules (linguistic constraints) to reduce the number of alternatives he might expect to hear in a string of successive words. That is, when the strings of words follow such rules, the listener's decision-making faculties (about what he is hearing) are enhanced. These and other experiments have led researchers to hypothesize that human listeners employ a delayed-decision strategy. That is, they may make only tentative decisions, at lower levels, on recognition of words or parts of words (based on phonological rules), and reserve their final decision until they perceive a whole phrase or sentence (based on syntactic rules). Very informally, we might say that a listener waits for "semantic markers," or whole ideational structures. In this view of matters, the human stores relatively large amounts of information and processes such information in parallel, in hierarchical structures, but the decision mechanism is itself a "sluggish serial device."[37]

Although recognition machines are not bound to mimic the human perceptual system in this regard, it is reasoned that because human speech is intended for the highly selective perceptual system of human ears, there is the possibility that the ongoing speech signal produced by a human will not always contain the kind of information that will support successive "quick" decisions, i.e., at the rate they presumably would be made at the level of word particles. In this view, then, successful recognizers of human speech might also require delayed-decision strategies of a hierarchically related complexity: they would need to know the two types of grammatical rules (phonological and syntactical), and they would need to know semantic rules as well. This means, of course, that the engineer building the recognizer would first need to know such rules, and his "knowing" would need to be of the most rigorous kind.

Some specific experiments related to the hypothesis stated above have been conducted by a number of experimenters.[38,39] Philip Lieberman, for instance, at the Air Force Cambridge Research Laboratories, set out to demonstrate that the semantic and grammatical context of a sentence affected both the production and the perception of speech in a consistent complementary fashion.[38] In speech production, for instance, the final acoustic signal will reflect decisions taken at the highest semantic level. For instance, Lieberman points out, a speaker uttering a well-known phrase like "A stitch in time saves nine," will not articulate the word nine as clearly as he will in a sentence like "The number you will hear is nine," because he knows that in the first case his listener will be able to fill in the missing parts, whereas in the second case he knows he will not. (Another way of saying this is to describe the speaker as following a principle of least work.) Tape recordings were made of many types of sentences containing certain test words that were later excised and used in listening tests. These experiments bore out the contention that speakers did modify the way they pronounced words in sentences, according to how well they knew their listeners would understand them. That is, in cases where they knew the listeners could rely on contextual semantic and grammatical information from the entire sentence, they pronounced certain words with less care and less stress. Or, to put it another way, the speaker knows he can rely on the listener to reserve his final recognition decision until he has heard the whole sentence. In this model of events, the listener is using results of higher-level processing to resolve uncertainties and confusions at lower levels. Lieberman cites other experiments and other evidence to support these views (notably, results from the work of Hemdal and Hughes, mentioned in Part I, who found that their computer-recognition program did less well with utterances that were actual words than did human listeners; thus, through an argument that cannot be reproduced here, Lieberman concludes "even for the simplest of all recognition schemes, isolated words, the computer should have available to it stored linguistic information").[40] On the basis of findings such as those above, Lieberman concludes, "The current search for invariant articulatory gestures is likely to be as futile as the previous hunt for invariant acoustic correlates."[40]

Lieberman has recently turned his attention to the role of intonation in the syntactic processing of speech.[41] The primary linguistic function of intonation, he says, is to furnish acoustic cues that allow the listener to segment speech into blocks for syntactic processing. That is, intonation can furnish different *meanings* to words by grouping the words into different blocks. Thus, a listener is cued in to direct his recognition routines toward one underlying phrase marker rather than another. In this interesting research, too, intonation is related to normal "breath groups," which are defined as the acoustic output that results from the synchronized activity of the chest, abdominal, and laryngeal muscles during the course of a single expiration. For most languages, normal breath groups end with falling fundamental frequency and falling amplitude, used to mark the limits of complete thoughts. Figure 7 shows how intonation, or breath groups, can change the meaning of the words of a sentence. Un-

Fig. 7. How breath groups can change the meaning of the words in a sentence. The sentence can mean either: A—I hit the man who had the stick; or B—The man was hit with a stick by me. Intonation can indicate either A or B by manifesting those aspects of the surface structure that differentiate the two underlying phrase markers.



fortunately, we cannot go into details, or even into the significance, of this work, which is said to reflect on the "folly of certain research on intonation during the past 30 years," but for those who are interested it will be published in due course.[42,43]

Even if it turns out that there need be no one-to-one relation between the way machines do it and the way men do it, we now can be fairly sure that machines for recognizing human speech will require a complexity of levels and strategies that approaches the complexity of the human brain.

Such a competent machine might require the use of language statistics at all levels—not just frequency distributions of spoken phonemes, phoneme diagrams, phoneme trigrams, minimal pairs (some of which have already been obtained).[44] It may be necessary to obtain a large variety of such statistics, by computer techniques, about morphemes, syllables, words, and their interrelations, and perhaps even about sentences, but such a task, "theoretically a trivial problem," may prove in practice to be quite a difficult one.[44]

There is some hope that help may come from the work on the use of learning or adaptive devices. Some work has already been done on speech recognition using such elements, but most such research can be said to be still in its earliest stages.[45]

In sum, it can be said that men have already learned very much about how men perceive speech, but as George A. Miller pointed out some time ago, "an engineer hoping to build devices that will recognize speech has a right to be discouraged."[37]

On the invariants of language

Most of the problems we have been considering have dealt with language on the level of the phoneme, the phonemic level. We have also considered the distinctive features, which go in the other direction, to the subphonemic level. Higher than phonemes, we encounter the level of morphology, the building of morphemes from the phonemic bundles of distinctive features. Higher still, we come to the construction of words, then phrases. At the level of syntax, we encounter the study of the principles and processes by which sentences are constructed in particular languages. Higher yet, there are formal studies of Grammar, the level of structures from which particular grammars for particular languages stem. These levels of studies, these related levels, which successively grow more complex, all relate to the general problem of the nature of Language.

Linguistic studies, as we might guess, are very old. Jakobson and Halle state that the search for the ultimate discrete differential constituents of language can be traced back to the Sanskrit grammarians, although, they say, the actual linguistic study of these invariants started only in the 1870s, developing intensively after World War I, side by side with the gradual expansion of the principle of invariance in the sciences.[32]

This search for the invariants of language has brought us to the most important philosophical concept we have encountered thus far in this survey—the concept of *rule* in language, the concept that language users rely on sets of rules, the concept that sets of ordered rules govern both the production and perception of speech at all levels.

This concept of grammatical rules carries us well beyond familiar limits. Noam Chomsky, the logician and linguist, says, "it seems that a really insightful formulation of linguistic theory will have to begin by a determination of the kinds of permitted grammatical rules and an exact specification of their form and the manner in which they impose structural descriptions on each of an infinite set of grammatical sentences."[46] And in defense of his formal methods, he says: "I think that some of those linguists who have questioned the value of precise and technical development of linguistic theory may have failed to recognize the productive potential in the method of rigorously stating a proposed theory and applying it strictly to linguistic material with no attempt to avoid unacceptable conclusions by *ad hoc* adjustments or loose formulation."[47]

Also, "by pushing a precise but inadequate formulation to an unacceptable conclusion, we can often expose the exact source of this inadequacy and, consequently, gain a deeper understanding of the linguistic data."[47]

For those who wish to go further into these revolutionary and productive studies of language, there already lies waiting a vigorous literature (Fodor and Katz's anthology seems particularly useful).[15] For the studies of phonological rules, one should seek out especially the works of Morris Halle,[48] and for the studies of syntactic structures, the works of Noam Chomsky[19–51] (five of his papers appear in Fodor and Katz[15]).

Allied with the concept of rules is the idea that the construction of language is not arbitrary; that despite the variety in languages and despite the variety possible in a single language, the structure on which such variety is based grows organically out of the biological nature of man. Perhaps a crude way of putting this is to say that men grow languages the way they grow hair. The style may change, but the thing stylized operates on the same functional laws. This concept is thus opposed to the concept that human language is a free creation, based on the collective and arbitrary fancy of a linguistic community.

People everywhere and always, so far as is known, have had spoken languages. Modern studies are revealing that such languages have pervasive similarities.

George A. Miller, a psycholinguist at Harvard, says of the languages of the world:

"The language always has a lexicon and a grammar. The lexicon is not a haphazard collection of vocalizations, but is highly organized; it always has pronouns, means for dealing with time, space, and number, words to represent true and false, the basic concepts necessary for propositional logic. The grammar has distinguishable levels of structure, some phonological, some syntactic. The phonology always contains both vowels and consonants, and the phonemes can always be described in terms of distinctive features drawn from a limited set of possibilities. The syntax always specifies rules for grouping elements sequentially into phrases and sentences, rules governing normal intonation, rules for transforming some types of sentences into other types.

"The nature and importance of these common properties, called 'linguistic universals,' are only beginning to emerge as our knowledge of the world's languages grows more systematic."[52]

The concept, then, that there are these invariants in human language makes the present search for such invariants a *meaningful* search. Further than that, it brings

the notion of building speech-recognizing automata within the range of being a practicable goal. It would be one thing to design a machine that must cope with an infinite, and essentially lawless, number of sentences, and quite another to design a machine with a finite set of rules. This latter (hypothetical) automaton would know in advance that it must search for certain kinds of regularities and correspondences unique to the human speaker so that its search time, instead of being inhumanly enormous, might at least be brought down to feasible limits, down to the human scale, where all these notions start in the first place.

REFERENCES

1. Fant, G., *Acoustic Theory of Speech Production*. 's Gravenhage, Netherlands: Mouton & Co., 1960.

2. Lindblom, B., "Spectrographic Study of Vowel Reduction," *J. Acoust. Soc. Am.* vol. 35, Nov. 1963, pp. 1773–1781.

3. Miller, J. E., *et al.*, "Study of Articulator Dynamics," *Ibid.*, vol. 34, Dec. 1962, p. 1978.

4. House, A. S., *et al.*, "Acoustical Description of Syllabic Nuclei: II. Interpretation in Terms of a Dynamical Model of Articulation," *Ibid.*, vol. 35, July 1963, p. 1112.

5. Stevens, K. N., and House, A. S., "Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study," *J. Speech & Hearing Res.*, vol. 6, 1963, pp. 111–128.

6. "Speech Analysis and Synthesis," Final Rept., AFCRL-64-300, Research Laboratory of Electronics, M.I.T., Cambridge, Mass., Dec. 1963; available from U.S. Dept. of Commerce, Office of Technical Services, Washington, D.C.

7. Soron, H. I., and Lieberman, P., "Some Measurements of the Glottal-Area Waveform," *J. Acoust. Soc. Am.*, vol. 35, Nov. 1963, p. 1876.

8. MacNeilage, P. F., "Electromyographic and Acoustic Study of the Production of Certain Final Clusters," *Ibid.*, Apr. 1963, pp. 461–463.

9. MacNeilage, P. F., and Sholes, G. N., "An Electromyographic Study of the Tongue During Vowel Production," *J. Speech & Hearing Res.*, vol. 7, no. 3, Sept. 1964, pp. 209–232.

10. Rosenblith, W. A., *Sensory Communication*. New York: John Wiley & Sons, Inc., 1961.

11. Hubel, D. H., and Wiesel, T. N., *J. Physiol.*, vol. 160, 1962, pp. 106–154.

12. Muntz, W. R. A., "Mechanisms of Visual Form Discrimination in Animals," *Proc. of Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Mass., Nov. 1964, to be published.

13. Frishkopf, L., and Goldstein, Jr., M. H., "Responses to Acoustic Stimuli from Single Units in the Eighth Nerve of the Bullfrog," *J. Acoust. Soc. Am.*, vol. 35, Aug. 1963, pp. 1219–1228.

14. Lettvin, J. Y., *et al.*, "What the Frog's Eye Tells the Frog's Brain," *Proc. IRE*, vol. 47, Nov. 1959, pp. 1940–1951.

15. Fodor, J. A., and Katz, J. J., *The Structure of Language—Readings in the Philosophy of Language*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1964.

16. Capranica, R., Frishkopf, L., and Goldstein, Jr., M. H., "Voice and Hearing in the Bullfrog," *Proc. of Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Nov. 1964, to be published.

17. Liberman, A. M., *et al.*, "A Motor Theory of Speech Perception," *Proc. of Speech Communication Seminar*, Stockholm, Sweden, 1962.

18. Liberman, A. M., *et al.*, "Some Observations on a Model for Speech Perception," *Proc. of Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Nov. 1964, to be published.

19. Hildebrand, D., "An Electronic Model of the Limulus Eye," in *Biological Prototypes and Synthetic Systems*, vol. 1, E. E. Bernard and M. R. Kare, eds. New York: Plenum Press, 1962, pp. 104–109.

20. Denes, P. B., "On the Motor Theory of Speech Perception," *Proc. of Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Nov. 1964 to be published.

21. Fant, G., "Auditory Patterns of Speech," *Ibid.*

22. Halle, M., and Stevens, K., "Speech Recognition: A Model and a Program for Research," *IRE Trans. on Information Theory*, vol. IT-8, no. 2, Feb. 1962, pp. 155–159.

23. Stevens, K. N., and Halle, M., "Remarks on Analysis-by-Synthesis and Distinctive Features," *Proc. of Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Nov. 1964, to be published.

24. House, A. S., *et al.*, "On the Learning of Speechlike Vocabularies," *J. Verbal Learning and Verbal Behavior*, vol. 1, no. 2, Sept. 1962, pp. 133–143.

25. Cramer, H. L., Private communication, Laboratory for Research in Instruction, Harvard University, Cambridge, Mass.

26. MacKay, D. M., "Mindlike Behavior in Artefacts," *Brit. J. for Philosophy of Science*, vol. 2, 1951.

27. Miller, G. A., *et al.*, *Plans and the Structure of Behavior*. New York: Henry Holt and Co., 1960.

28. MacKay, D. M., "Ways of Looking at Perception," *Proc. of Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Nov. 1964, to be published.

29. Stevens, K. N., Private communication, Research Laboratory of Electronics, M.I.T.

30. Öhman, S., and Stevens, K. N., "Cineradiographic Studies of Speech: Procedures and Objectives," Rept., Research Laboratory of Electronics, M.I.T (unpublished); also *J. Acoust. Soc. Am.*, vol. 35, Nov. 1963, p. 1889.

31. Jakobson, R., Fant, C. G. M., and Halle, M., "Preliminaries to Speech Analysis," Tech. Rept. No. 13, Acoustics Lab., M.I.T., 1952.

32. Jakobson, R., and Halle, M., *Fundamentals of Language*. 's Gravenhage, Netherlands: Mouton & Co., 1956.

33. Miller, G. A., Heise, G. A., and Lichten, W., "The Intelligibility of Speech as a Function of the Context of the Test Materials," *J. Exp. Psychol.*, vol. 41, 1951, pp. 329–335.

34. Miller, G. A., and Isard, S., "Some Perceptual Consequences of Linguistic Rules," *J. Verbal Learning and Verbal Behavior*, vol. 2, no. 3, Sept. 1963, pp. 217–228.

35. Marks, L. E., and Miller, G. A., "The Role of Semantic and Syntactic Constraints in the Memorization of English Sentences," *Ibid.*, vol. 3, no. 1, Feb. 1964, pp. 1–5.

36. Pollack, I., "Message Probability and Message Reception," *J. Acoust. Soc. Am.*, vol. 36, May 1964, pp. 937–945.

37. Miller, G. A., "Decision Units in the Perception of Speech," *IRE Trans. on Information Theory*, vol. IT-8, no. 2, Feb. 1962, pp. 81–83.

38. Lieberman, P., "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech," *Lang. & Speech*, vol. 6, pt. 3, July–Sept. 1963, pp. 172–187.

39. Pickett, J. M., and Pollack, I., "Intelligibility of Excerpts from Fluent Speech: Effects of Rate of Utterance and Duration of Excerpt," *Ibid.*, 1963, p. 151.

40. Lieberman, P., Private communication, AFCRL, Speech Research Branch, Data Sciences Laboratory.

41. Lieberman, P., "Intonation and the Syntactic Processing of Speech," *Proc. of Symp. on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, Nov. 1964, to be published.

42. Lieberman, P., "On the Acoustic Basis of the Perception of Intonation by Linguists," *Word*, scheduled for publication in Apr. 1965.

43. Lieberman, P., "Intonation and the Perception of Speech," in preparation.

44. Denes, P. B., "On the Statistics of Spoken English," *J. Acoust. Soc. Am.*, vol. 35, June 1963, pp. 892–904.

45. Talbert, L. R., *et al.*, "A Real-Time Adaptive Speech-Recognition System," Tech. Report No. 6760-1, ASD-TDR-63-660, Systems Theory Laboratory, Stanford Electronics Laboratories, Stanford, Calif., May 1963

46. Chomsky, N., "On the Notion 'Rule of Grammar,'" in *The Structure of Language*, J. A. Fodor and J. J. Katz, eds. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1964, pp. 119–136.

47. Chomsky, N., *Syntactic Structures*. 's Gravenhage, Netherlands: Mouton & Co., 1957.

48. Halle, M., "On the Bases of Phonology," and "Phonology in Generative Grammar," in *The Structure of Language*, J. A. Fodor and J. J. Katz, eds. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1964, pp. 324–352.

49. Chomsky, N., "Three Models for the Description of Language," *IRE Trans. on Information Theory*, vol. IT-2, 1956, pp. 113–124.

50. Chomsky, N., "On Certain Formal Properties of Grammars," *Information & Control*, vol. 2, 1959, pp. 137–167.

51. Chomsky, N., and Halle, M., *The Sound Pattern of English*, to be published.

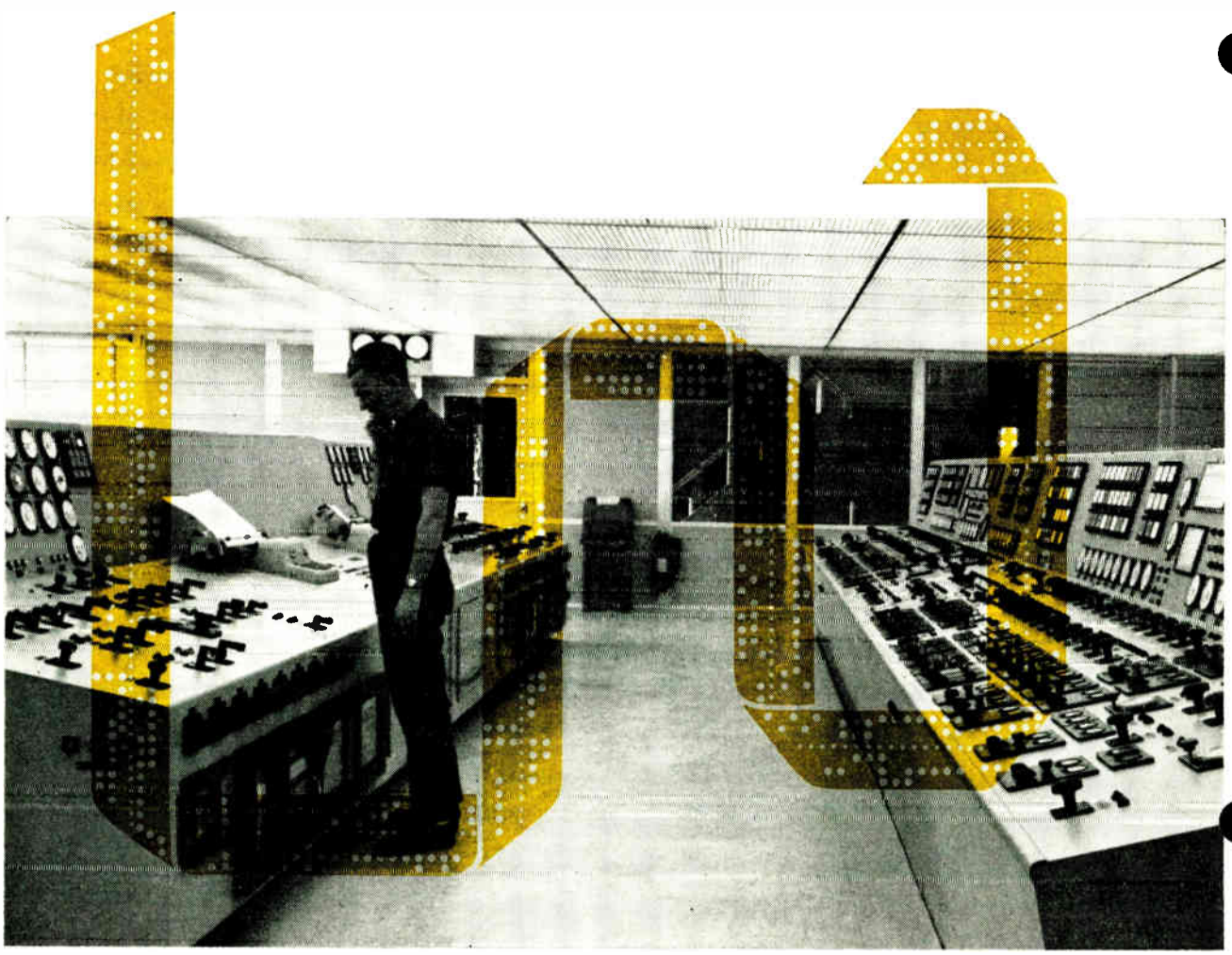
52. Miller, G. A., "The Psycholinguists—On the New Scientists of Language," *Encounter*, vol. 23, no. 1, 1964.

Computer-controlled power systems

Part I—Boiler–turbine unit controls

Over the past decade, automatic power plant monitoring and supervisory instrumentation, area load–frequency, and transmission control of interconnected power systems have developed into sophisticated on-line analog–digital computer and telemetering centers that may be the key factors in the establishment of a future firm national grid

Gordon D. Friedlander *Staff Writer*



There is the apochryphal story popular several years ago that a middle-echelon employee of a major utility company, who had served faithfully in his administrative position for more than 30 years, arrived at his place of employment one Monday morning and, to his amazement, was informed that he had been promoted to executive vice president. He occupied his new and plush office in a state of shock and bewilderment for the better part of a week, and tried—without success—to rationalize his most unexpected promotion.

Finally, he reached the point of perplexity where he just had to get some answers, so he asked his secretary if she could offer any possible explanation for his good fortune. The girl pondered for several minutes with a furrowed brow—then she brightened: "I think I know what may have happened. Last Friday night I dropped your punch card on the floor, and the chairman of the board walked over it with his golf shoes."

It was perfectly natural, therefore, on the basis of this remarkable breakthrough, for the private utilities to proceed with the automation of not only plant monitoring and supervisory controls, but also area load dispatch controls.

Some historical footnotes

In the early days of power generating stations,[1] their design was conceived as an assembly of individual components such as the boiler plant, turbine units, fuel conveyor system, and water-cooling system.

This division also existed in operation. For example, the boiler fireman's duty was to produce steam at a predetermined pressure by using bunker coal that was provided by the mine operators. The control room was concerned with the generation of electric energy, voltage control, and the distribution of the energy into the local networks. Other operators independently attended to the coal delivery and transportation system, the ash removal from the boilers, the provision of cooling water, etc. Each of these systems had its independent and unintegrated control center, and, if any instrumentation was provided, it was quite rudimentary and primitive.

As the state of the art developed, however, the individual control centers became more complex, and the operator was provided with further instrumentation and plant controls to enable him to regulate the plant from the control panel of the control center. This concept was further refined by the inclusion of alarm and alert indicators and more sophisticated instrumentation and controls. Thus the greater part of the operator's tasks was concentrated on the monitoring and supervision of the control panel. By placing two similar panels adjacent to each other, one operator could satisfactorily supervise two major plant functions.

Trend of the past decade. The trend over the past ten years has been to integrate further control of the station by the provision of a plant central control room, from which all functions and operations of some or all of the boilers and turbines are controlled. This concept has been accelerated by the recent advances in boiler and turbine design. It has also been aided by the parallel development of instrumentation telemetry, which permits the detector portion of the measurement system to be separated from the indication by greater distances than could be achieved by Bourdon tube gauges, mercury-in-steel thermometers, etc.

Unit controls—the present position

In the logical evolution, automated controls have been applied first to the unit controls within the power generating station.

With the present increasing demand for electric power, the trend is toward larger capacity and more complex turbine-generators.[2] Steam pressure and temperature conditions of these units are being geared to the demand for higher efficiency. To realize the full benefits of these larger capacity machines, utilities must have full plant availability and performance. Yet the success of such systems still depends largely on human operators. But can operators continue indefinitely to assimilate an increasing amount of data and make the correct decisions without improved information and control systems?

Actually, the increased unit capacity is such that the instrumentation, recording, alarm systems, and the remote controls to enable the plant to be operated from the control room under very adverse conditions will produce a mass of data too great for one operator either to act upon or comprehend.

Boiler warm-up, turbine acceleration, and turbine-generator loading at South Carolina Electric and Gas Company's Canadys Plant are shown being compared to desired sequences and programmed into a computer.

For example, a generator prime mover normally has governing equipment which controls the flow of energy in response to speed and load changes. As system load increases, the result is a slight drop in system speed or frequency, and the governors operate to admit more energy to the prime movers to match the new load. The reverse occurs when the load decreases. Thus manual or automatic control is required to adjust the governor mechanism to restore speed to normal after a load change.

The elementary form of automatic combustion control has evolved into a comprehensive boiler–turbine control system that incorporates numerous subloops. Given a set point, the closed loop control maintains the controlled variable at the desired value within fixed limits. In some cases, provision has been made for complete automatic control during start-up and shutdown. And in line with the continuous developments in power generation, a parallel development has been triggered in new concepts of instrumentation technology, electronics, and data processing.

There are several schemes that employ these new techniques. They are either in operation already or in the design stages. Figure 1 shows how the boiler controls and turbine supervisory equipment of one major manufacturer are integrated in an automated power plant. Other similar schemes, each designed to meet a particular station requirement, are based upon the data accumulation of past operational experience.

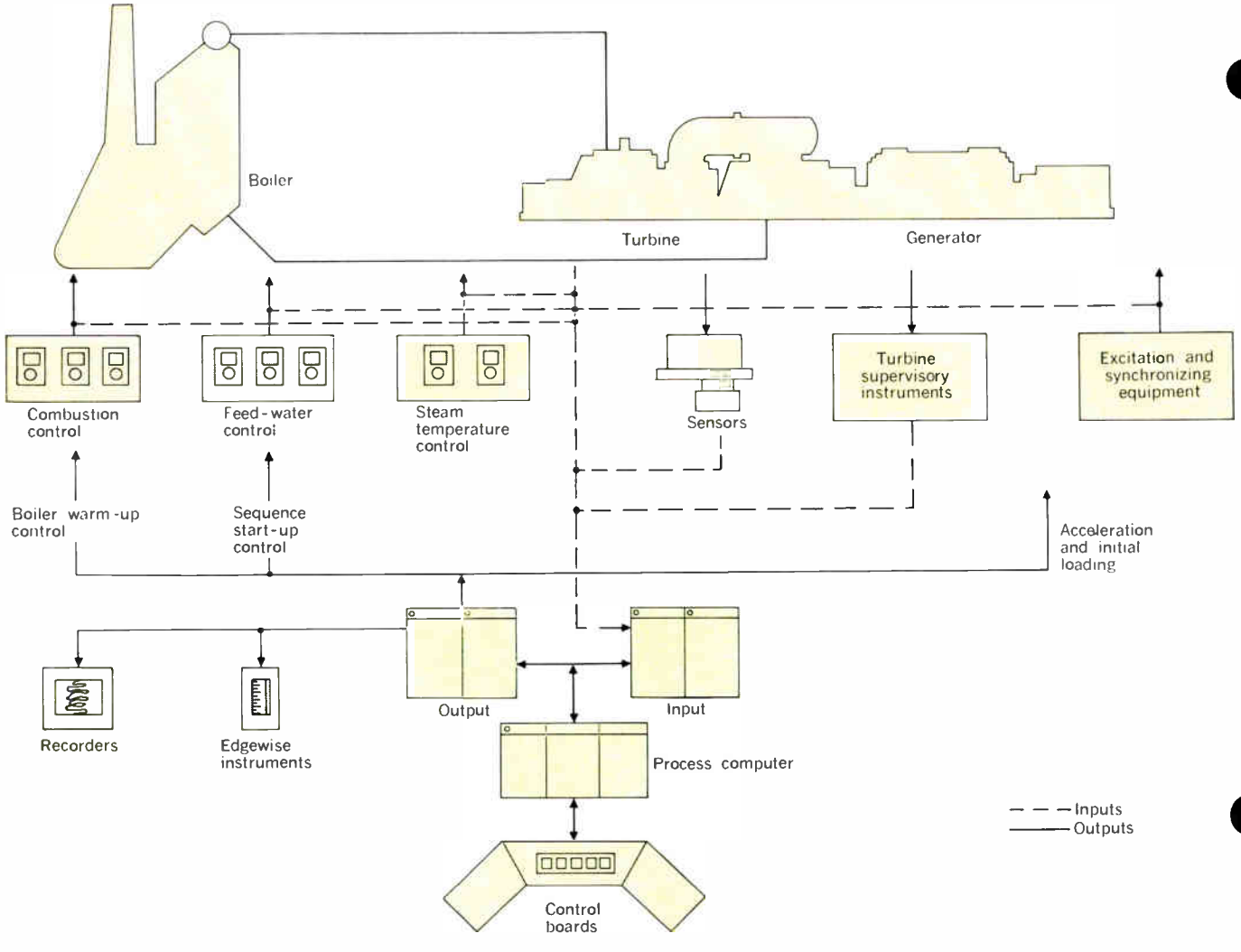
For coal-burning plants, additional development in unit controls will be required before full automation can be achieved. But in the case of gas- and oil-fired plants, the problems are less difficult, and therefore automation has been applied first to such power stations.

Advantages of automatic control and supervisory systems

The several advantages in favor of control and supervisory systems are

1. The continuous supervision of a plant by an on-line system which is always alert and capable of instantaneous action.
2. The elimination of the routine tasks performed by the operator, thereby leaving him free to concentrate on the optimum operation of the plant, and to correct plant abnormalities and faults.

Fig. 1. Diagram showing how typical boiler controls and supervisory equipment are integrated into an automated steam-electric power plant.



5. The use of a consistent and repeatable procedure for start-up that will produce minimal thermal stresses and shocks to vulnerable parts of the plant. This can attain more reliable machine performance and a reduction in outage and maintenance costs.

Unit control hardware—and software

Control schemes. The logical development toward a completely automatic station requires, as the first step, some form of data-acquisition system to obtain time-correlated plant data during normal and abnormal operation.

Initially, this operational data logging is used to establish the parameters of operation through the safe and unsafe ranges by the correlation of the data with the specified alarm limits that require immediate operator attention. Sophistication of this system would allow performance data to be calculated on the basis of short-time average values. A by-product of this procedure is a reduction of the work load for the clerical staff who normally maintains operational records.

The second step utilizes the results of these data to obtain the relationships between the operational parameters and to optimize the operation of the unit.

In parallel with the programming just mentioned, consideration is given to start-up, shutdown, and fault correction by automatic means to eliminate operator error during these critical phases of operation.

This system prepares the boiler and turbine units for operation by starting the auxiliary machinery, warming up the boiler, and performing the pressure-raising procedure. At the correct steam conditions, the turbine is started and allowed to run up to operational speed (while monitoring the supervisory instrumentation for danger signals). The turbine unit is then synchronized and either block-loaded, or loaded to an output load demanded by the system. The on-line monitoring of the unit from this point may be performed by an industrial computer-based system.

While Fig. 1 presents an overall block diagram of the instrumentation and computer controls for a power station, the unit hardware for individual plant components is far more complex. For example, the supervisory controls for the boilers alone will include controls for combustion, feed water, steam temperature, air heater temperature, pump recirculation, pressure reducing, and desuperheating.

Utility boilers. Most of the modern utility boilers are large, high-pressure reheat units that serve turbogenerators of 100-MW capacity and higher. These boilers are usually field erected, and control systems are often purchased separately and tailored to the particular unit.

Boiler controls[3] consist of several multivariable control systems, each designed to regulate some basic function of the unit. They are essentially independent systems that are indirectly related through the process. Thus proper operation of the boiler depends upon the normal function of each individual system, and the malfunction

stem can upset the process. The variables to olled and the complexity of the various control are generally related to the size of the boiler, are influenced considerably by the manner in the boiler must operate.

e basic parameters of controlling fuel, air, and water enerate steam at the required rate, pressure, and temperature are equally applicable to boilers of every size. The equipment required to meet these criteria of operation, and the control requirements of this equipment, account for the wide variety of system configurations. Figures 2(A) and 2(B) show the steam, water, and condensate cycle of the basic utility plant and the basic points of measurement that are common to most boilers.

Utility boiler combustion control system

The combustion control system that will be described and illustrated in this article has been designed specifically for a Combustion Engineering Inc. divided furnace boiler. But, except for minor details, it is applicable in principle to any large utility boiler of over 1 000 000-lb/hr capacity. The key feature of this particular control configuration is that a digital computer will program the steam pressure from start-up to on-line conditions.

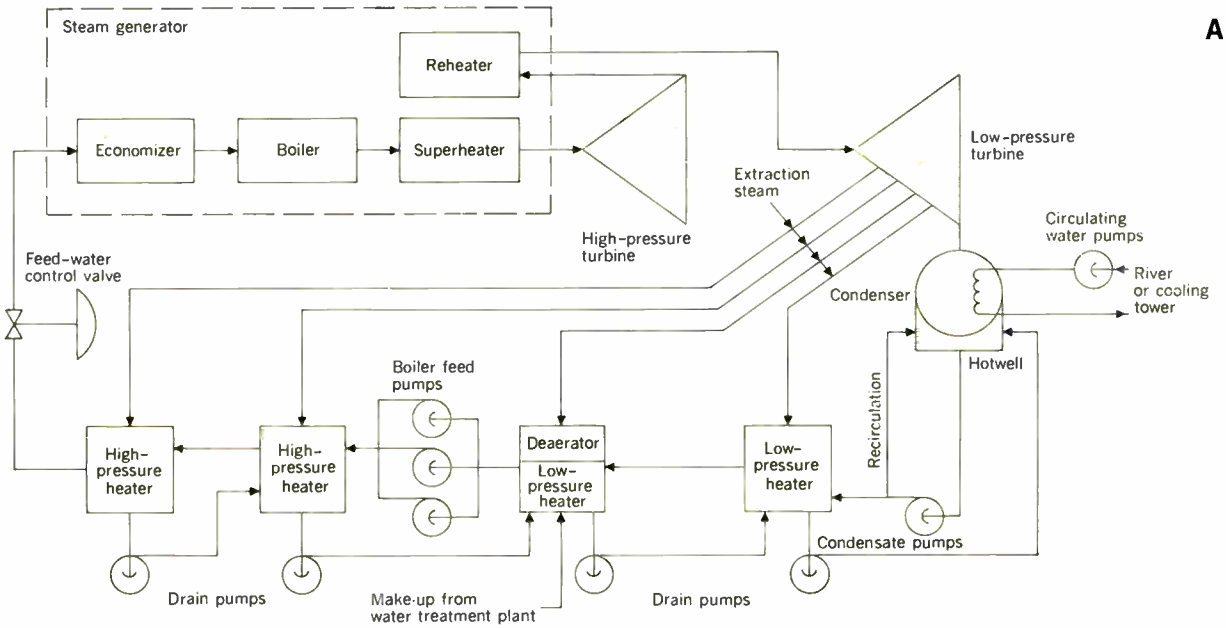
The fuel firing system is for pulverized coal, oil, gas, or combinations of these fuels. Many modern utility boilers operate under positive furnace pressures, thereby eliminating the necessity for induced draft fans and furnace draft control as shown in Fig. 2(B).

Fuel control. As shown in Fig. 2(C), the main steam pressure transmitters vary in function. The wide-range transmitter run is from 200 to 3400 psi so that the control system may be operated on automatic at very low pressures. The computer can change the pressure controller set point, and gradually raise the pressure on the unit from a stored computer program. When the unit reaches the operating condition of 1875 psi, extremely precise pressure control must be maintained; to obtain this precision, a narrow-range transmitter with a run of 1800 to 2000 psi is necessary. This pressure controller has the versatility of proportional and reset control action, plus a manual set point and manual–automatic selection station, and the ability to allow the computer to adjust the set point. When firing gas or oil, the pressure controller calculates the required fuel and transmits its signal—commonly called the *master signal*—through the manual–automatic station, and the oil valve or gas valve will admit either more or less fuel to the boiler.

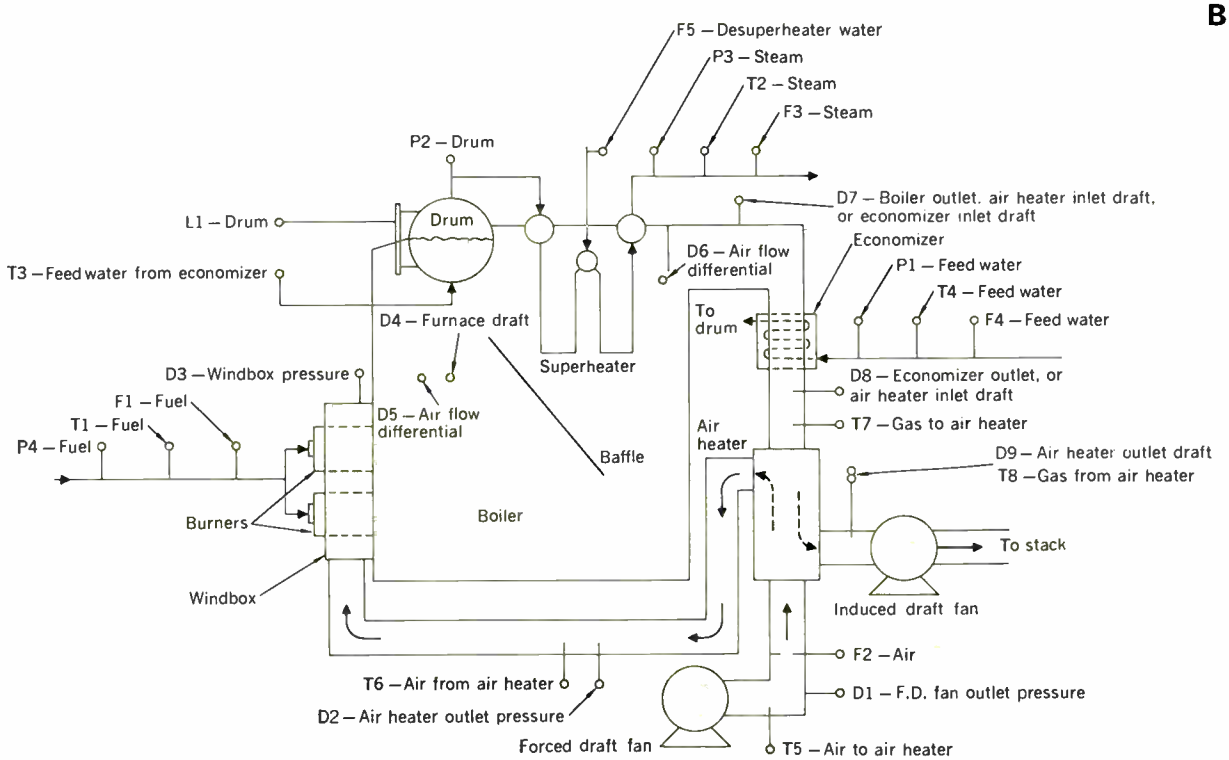
When firing pulverized coal, however, additional control equipment is required since most coal pulverizers have a considerable time lag between the initiation of the increased rate of feed coal to the pulverizer and the release of additional heat from the furnace. Therefore, rate or *derivative* action is applied to the master signal just ahead of the pulverizer manual–automatic stations. This controller determines the *rate of change* of the pressure control signal and, in effect, gives anticipatory action to the feeders. This signal passes through the manual–automatic stations with ratio adjustment directly to the exhauster damper, and then is tapped off beyond the ratio station through a bias manual–automatic station to the feeder speed control. The ratio station will be adjusted dependent upon the number of coal-pulverizing mills in service.

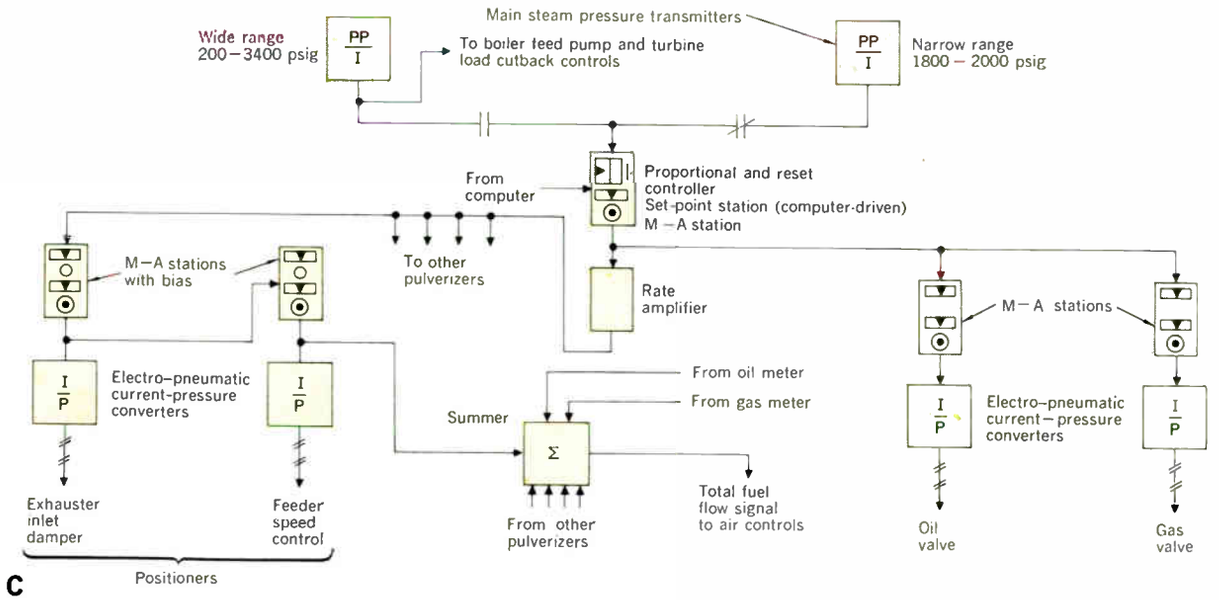
Fig. 2. A—Instrumentation for basic utility plant steam, water, and condensate cycle. B—Basic points of measurement common to most utility boilers. Note that prefix D indicates draft and air points, F indicates liquid and gas flows, L indicates a level (drum), P indicates pressure, and T indicates temperatures. C—Fuel control system for combustion in a utility boiler. D—Forced draft air controls for utility boiler combustion control system.

A

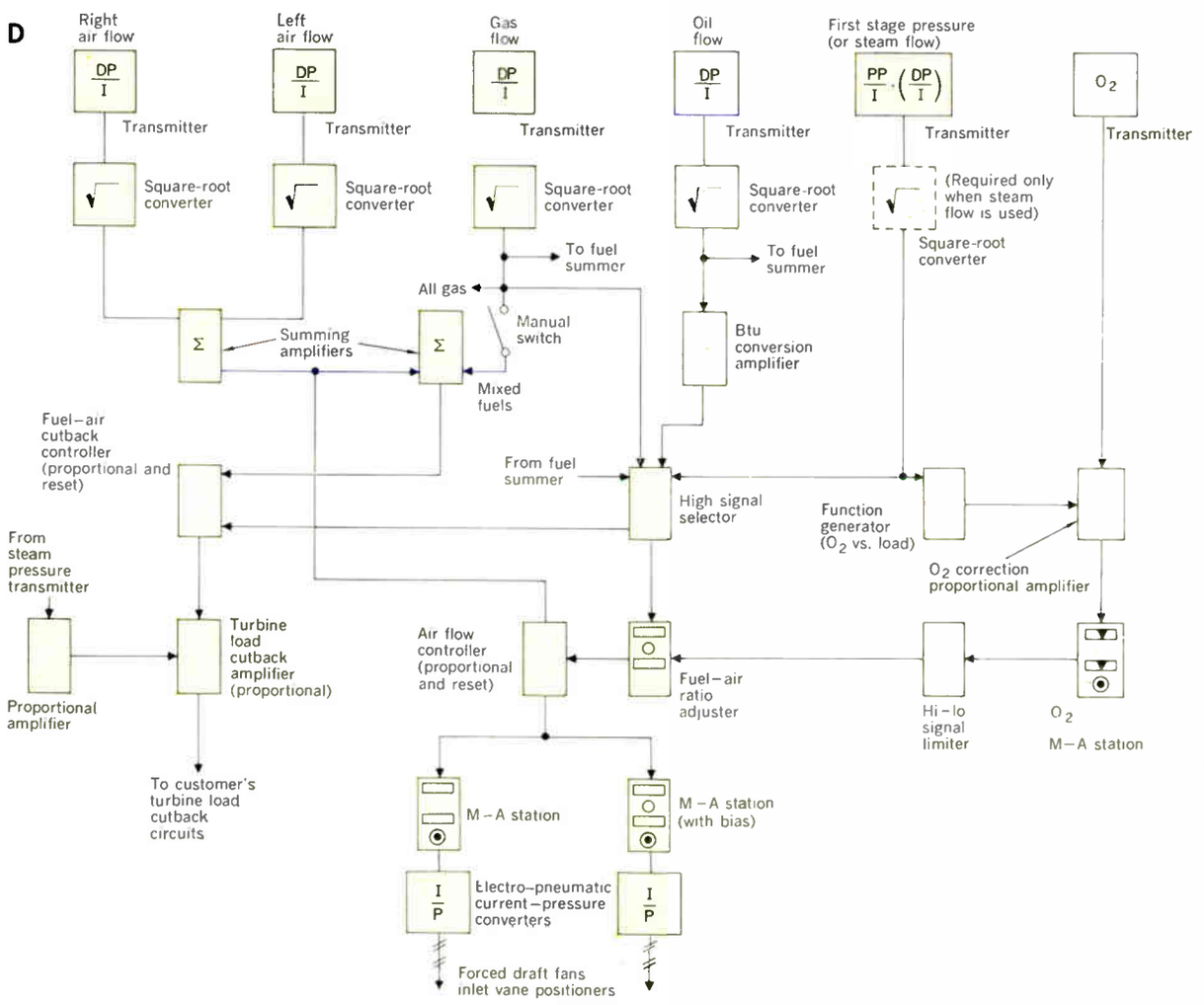


B





C



Air control—forced draft. The air flow is measured across a venturi section in the duct work to each furnace. Two air flow transmitters, two square-root extractors, and a summation amplifier are required. Figure 2(D) is a block diagram for a forced draft air control system. The final signal, indicative of total air flow to the boiler, is transmitted to the air flow controller. It will also be the feedback signal transmitted through the process itself in response to the air flow demand signal. The latter is a function of the "total fuel fired" signal from the fuel summation amplifier, or from the Btu requirement as indicated by steam flow that includes the corrective action of the oxygen system.

Note in Fig. 2(D) that the gas flow transmitter sends its signal to a square-root converter, and then into a high signal selector. The oil flow performs the same function with the added feature of a Btu conversion amplifier. Since it is necessary to convert the fuel oil signals to equivalent Btu's to standardize measurement units, the conversion amplifier is a vital component.

The steam flow on large utility boilers is often calibrated from the first stage inlet of the high-pressure turbine. In some cases there may be one or more steam flow meters, differential pressure transmitters, square-root extractors, and a summation amplifier (see Fig. 2D).

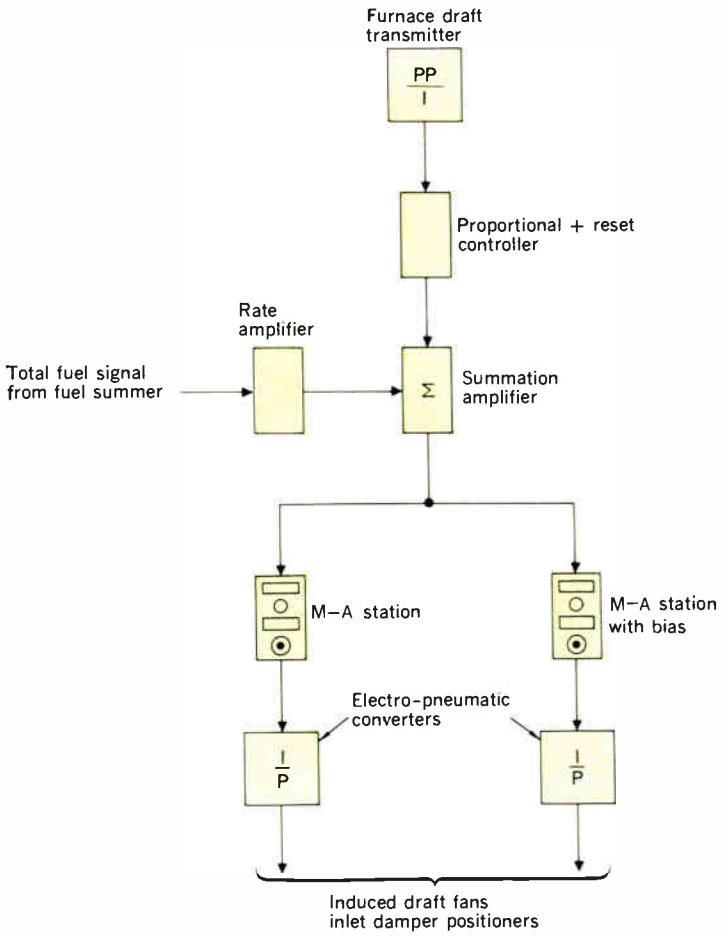
The steam flow signal and the fuel output signal from the fuel summation amplifier are transmitted to the high signal selector. The output signal from the high signal selector will be the *highest* of the four inputs, and will represent the air flow demand. Actually, this signal represents a comparison between the fuel flow signal and the steam flow signal in Btu's.

The fuel–air ratio adjustment station permits the operator to adjust his fuel–air ratio, and it also provides a point at which the oxygen correction signal may be interposed. The output of this station enters the air flow controller as the demand signal for total combustion air. The air flow controller develops a proportional, plus reset action, signal to the forced draft fan manual–automatic stations, one of which is provided with a bias adjustment for the operator to balance the load on the two fans.

In a forced draft boiler, it is desirable to program the excess air content of the flue gases with respect to unit load, as greater excess air is required at low loads for proper fuel and air mixing at the burners. A function generator (Fig. 2D) is included between the steam flow measurement and the oxygen correction amplifier. The generator will be adjusted to generate a curve indicative of the oxygen requirement vs. load. The oxygen transmitter will sense oxygen in the flue gases, and transmit its signal to the oxygen correction amplifier. The latter device will determine any difference between the actual oxygen content of the flue gases and the demand signal from the function generator, and will transmit a corrective signal through its manual–automatic station to the fuel–air ratio adjustment station. The manual–automatic station is placed here to permit the periodic maintenance that is required on the oxygen transmitter sampling system. During such maintenance, the operator will keep the station on manual, and the overall air flow system will operate without oxygen correction.

The sampling lines occasionally become plugged, and the corrective action to the fuel–air ratio system must be correspondingly limited for overall boiler safety. For this reason, a high- and low-limit station, which can be field-adjusted to allow a limited amount of correction from the oxygen influence, is included between the oxygen correction amplifier and the fuel–air ratio adjustment station.

Fig. 3. Induced draft air controls for typical utility boiler combustion control system.



Air control—induced draft. Some large utility boilers, as shown in Fig. 2(B), have induced draft fans that permit the combustion chamber pressure to be maintained at 0.15 inch (water column negative). As shown in Fig. 3, the furnace draft transmitter, sensing this pressure, transmits its signal to the furnace draft controller—a device which incorporates proportional, plus reset, control, and transmits its signal through a summing amplifier to the manual–automatic stations that control the fan damper positioners. One of the manual–automatic stations contains a bias adjustment for balancing the load between the fans. The summing amplifier between the controller and the manual–automatic stations is used to interject a derivative action signal from the fuel system. Large boilers of this type have a huge volume and a long gas run from the input section to the stack. Thus it is desirable to anticipate the rapid changes in fuel firing rate to assist the furnace draft system, and in this circuit, the rate of change of total fuel is employed.

Safety features. The safety or cutback features of this control system will be triggered by the lack of necessary combustion air, low main steam pressure, or loss of one of the two main feed pumps to the turbines.

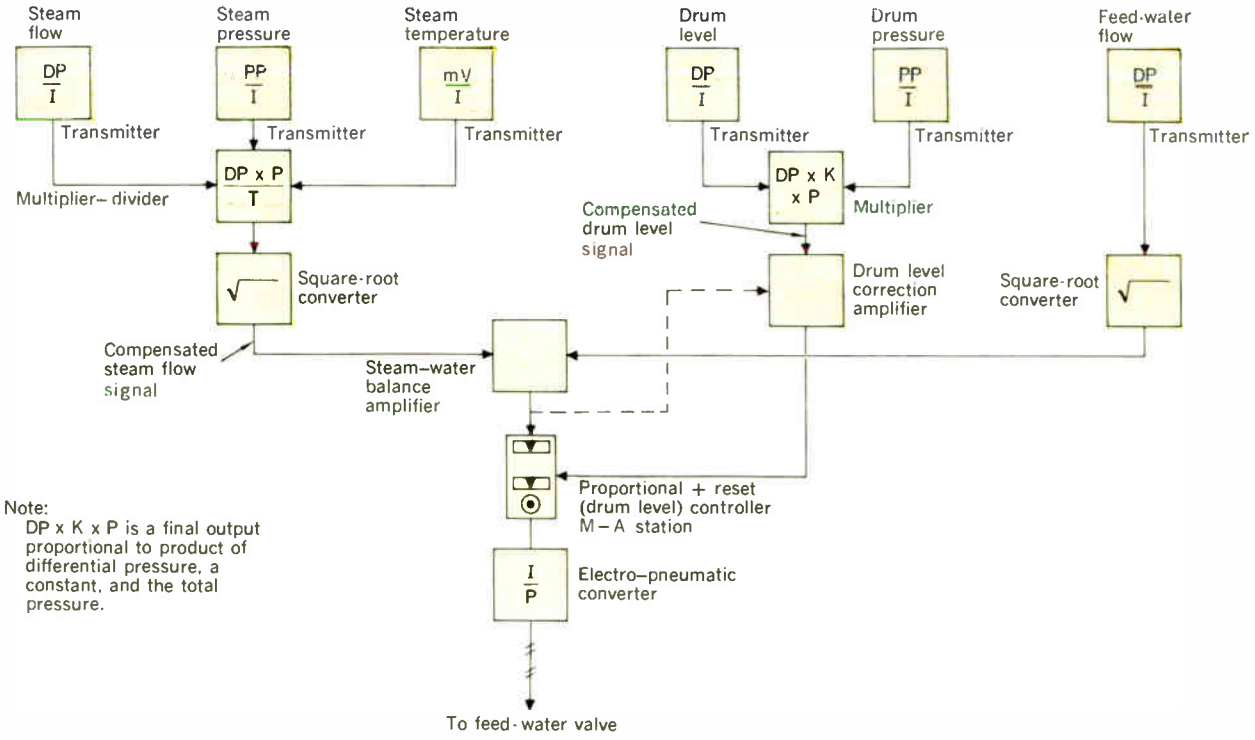


Fig. 4. Circuit configuration for three-element feed-water control, compensated forboilers in a utility plant.

The fuel and air cutback controller, Fig. 2(D) is adjusted so that, if the available air is insufficient for fuel combustion, the controller will transmit a signal to the turbine load cutback amplifier, and thence to the turbine cutback system. This action will decrease the steam requirements of the unit which, in turn, will tend to increase the steam pressure with a corresponding decrease in the fuel demand.

The air requirements for firing oil and coal are about the same, but they are significantly less for gas. Therefore, the summing amplifier is included between the air flow summation amplifier and the fuel–air cutback controller. A manual switch, which allows the selection of mixed fuel or all gas, is associated with this amplifier. The operation of this switch will add a signal in the summing amplifier that will change the point at which the fuel–air cutback controller can operate.

The low main steam pressure cutback amplifier receives a signal from the wide-range steam pressure transmitter, Fig. 2(C), and if this signal falls below a preset pressure value, the amplifier will apply a corrective signal into the turbine load cutback amplifier. The cutting back of the turbine will decrease the amount of steam required and allow the boiler pressure to build up.

Feed-water control for utility boilers

The feed water, or drum level, control for a utility steam generating unit is usually controlled by a circuit configuration (Fig. 4) that involves three elements: drum level, steam flow, and feed-water flow. The circuit is designed to compare the steam flow vs. water flow for equality, compare the drum level to its set point, and then to combine these two influences in a controller to actuate the final control element.

On high-pressure utility units, the drum level transmitter of the feed-water control circuit must be compensated for changes in the specific gravity of the water that are caused by temperature and pressure changes. As the drum level is a critical function in the operation of the boiler, it is desirable to have the drum level control put on automatic long before the unit delivers large amounts of steam to the turbine.

As indicated in Fig. 4, the steam flow meter is pressure- and temperature-compensated—a differential pressure device, a pressure transmitter, and a temperature transmitter are included together with a multiplication–division–square-root extraction network. These may be either static solid-state devices or analog servomechanism computing elements that produce a pressure- and temperature-compensated steam flow signal.

In some high-pressure boilers it is not desirable to use a primary element in the steam leads to the high-pressure turbine, and the accepted method for instrumenting steam flow is to utilize first-stage steam pressure. In such cases, the steam pressure, which is indicative of flow, is temperature-compensated to obtain the corrected steam flow signal for the three-element feed-water system. The compensated steam flow and water flow are generally recorded on the same two-pen instrument.

The water flow is normally measured at the discharge of the feed pumps, and the steam flow–water flow balance amplifier determines any difference between the feed-water flow and steam flow signals. The output of this amplifier is biased so that its signal level is at mid-scale whenever feed water and steam flow are equal.

The drum level is measured by a differential pressure device that transmits its signal, for compensation purposes, into a multiplier. The output from this corrective

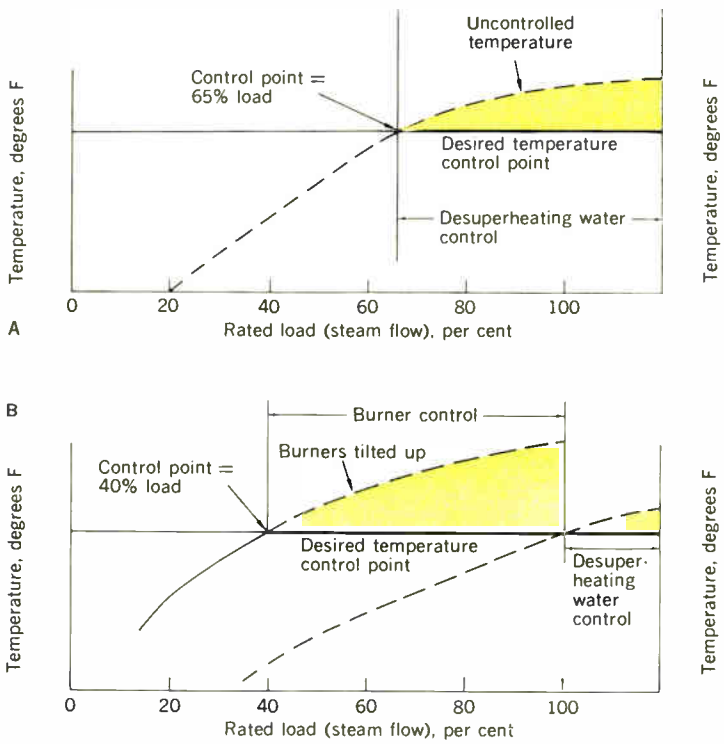
and compensatory multiplier represents the true drum level; the signal is always recorded, and it is often prepared as an individual chart.

Next, the drum level correction amplifier receives the true drum level signal and compares it with a predetermined set point. The output signal from this component is biased to mid-scale whenever the drum level is equal to the set point so that positive or negative readings may be obtained for high or low drum levels respectively.

The signals from the steam and water balance amplifier and the drum level correction amplifier are fed into the drum level controller which, under stable operating conditions, maintains the feed-water valve in a position that will maintain equality between the two signals. Signal equality exists when the steam flow is equal to the feed water entering the boiler and the drum level is normal. An abnormal combination will influence the drum level controller to operate the feed-water valve through proportional, plus reset, adjustments until the valve restores the system to normal.

The manually operated feed-water valves for large utility boilers are usually piston-actuated, and may have a 5-in bore by 10-in stroke cylinder that provides a 2000-pound thrust straight at the shaft. Because of this and the high-pressure drops—on the order of 100 psi—the feed water is often mechanically controlled by a hydraulic coupling on a motor-driven pump, or by steam actuation from an auxiliary turbine-driven pump. If the control configuration utilizes a feed-water valve or the hydraulic coupling between a constant-speed motor and a pump, proportional, plus reset, control is sufficient; however, if the control medium is a steam-driven turbine, some derivative action may be necessary.

Fig. 5. A—Graph showing plot of steam temperature control point on a boiler, with fixed burners and a superheater, that is designed to provide the control temperature within the range of 65 to 120 per cent of the rated load. B—Graphic plot similar to A, but for a boiler with tilting burners. Note that the control range has been extended down to 40 per cent of load. No desuperheating is required until the load reaches 100 per cent of rating.



Steam temperature control

Today, all utility units of 100 MW and higher capacity operate on the reheat cycle. Saturated steam from the drum passes through tube nests, called *superheaters*, which are located in the boiler gas passages and absorb heat from the combustion gases that leave the furnace. Simultaneously, steam leaving the high-pressure turbine is returned to the boiler and is passed through another tube bundle—or reheater—that raises its temperature before it enters the intermediate- or low-pressure turbine. The reheated steam is usually raised to the same temperature as the superheated steam, 1000° to 1050°F on most units.

The basic methods of steam temperature regulation used on modern utility boilers are: flue gas recirculation, tilting burners, by-pass dampers, and condensers. Most of the large American utility boilers are designed so that one of these four methods is employed for the major portion of the superheat and reheat control, with a small amount of spray desuperheating reserved for high load operation and primary control backup.

Figure 5(A) represents the steam temperature control on a boiler with fixed burners and a superheater that is designed to provide the control temperature within the range of 65 to 120 per cent of the rated load. Note that there is a considerable amount of desuperheating required.

Figure 5(B) represents the steam temperature control on the same boiler, but with tilting burners. Observe that the control range has been extended down to 40 per cent load; however, no desuperheating is required until the load reaches 100 per cent of rating.

Reheat boilers usually incorporate one of the four basic methods previously mentioned to control reheat temperature. While the basic method controls the final reheat temperature, it will also have its effect on the final superheat temperature. But superheat control will be by desuperheating spray to maintain superheat temperature at its control point.

Superheat–reheat temperature control. Because of space limitations this article cannot encompass every superheat–reheat temperature control system that is commonly in use. Therefore the writer will attempt to present in detail only one instrumentation and monitoring arrangement—that for utility boilers with tilting burners. In this system the main steam temperature is controlled by spray valves, and reheat steam temperature is controlled by tilting the burners up and down to direct radiant heat against the reheater systems, with spray backup.

In the piping configurations for a superheater and a reheater, the superheated steam section includes a primary superheater, crossover network, and a secondary superheater. Spray water, the cooling medium, is admitted in the crossover section, and an interstage temperature measurement is taken at the inlet to the secondary superheater. The final steam temperature is

measured in the main line to the high-pressure turbine.

The reheat section includes a primary reheater, a multitube connection, and a secondary reheater. The exhaust steam from the high-pressure turbine is fed into the primary reheater section, and the cooling controls for emergency and backup service are the right and left spray valves that are installed ahead of the primary reheater.

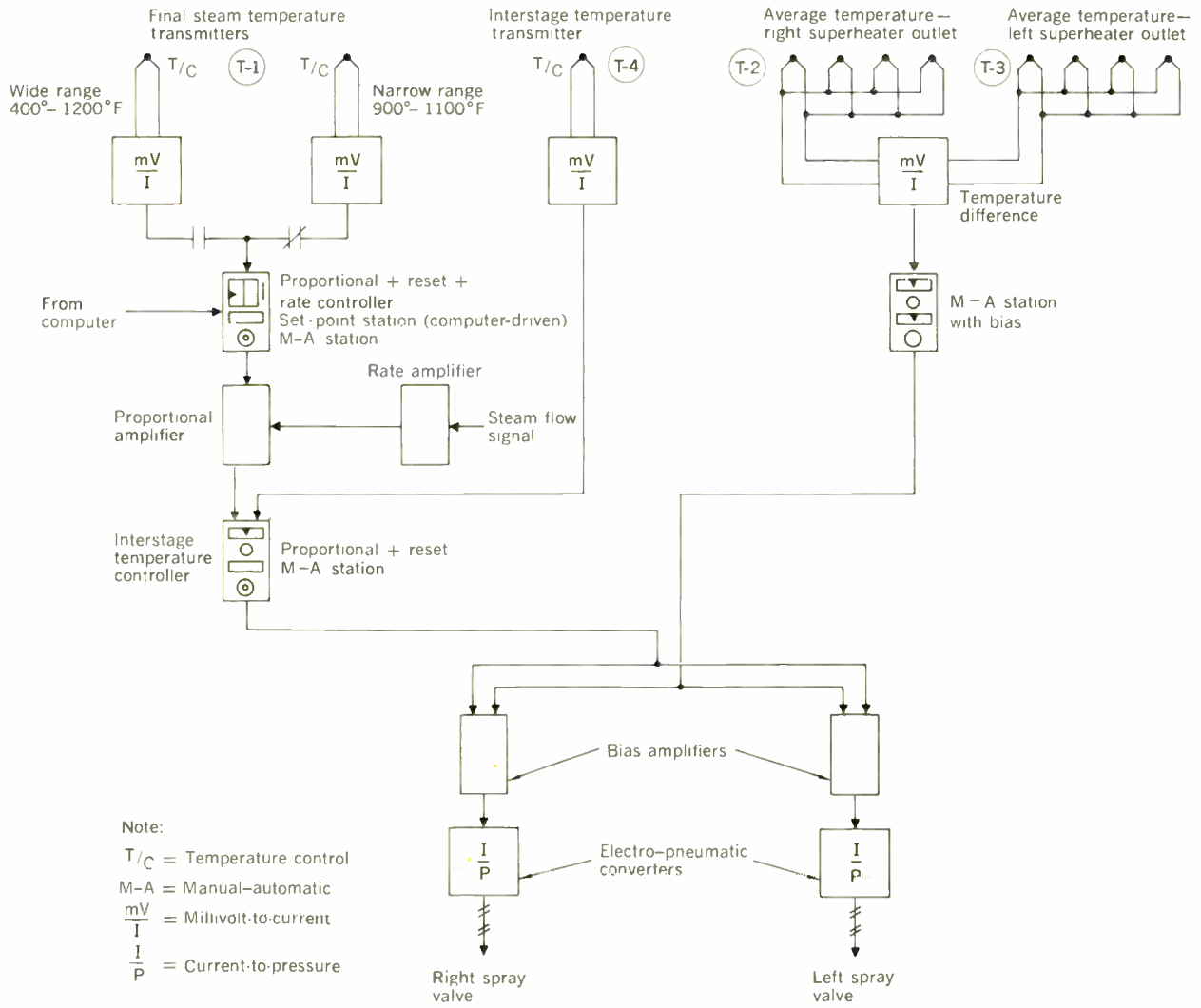
The type of burners under discussion are usually known as *tangential-firing tilting burners*. A mechanical–pneumatic positioner or an electric actuator raises and lowers the tilt angle of the burners to apply more or less heat to the reheat section. The final reheat temperature is sensed in the main steam line to the intermediate-pressure turbine.

Final steam temperature control. Figure 6 indicates the necessary circuitry features for final steam temperature control that are associated with a computer-controlled steam-electric generating station. The main steam temperature control configuration utilizes interstage temperature control that is cascaded from final steam temperature. The right and left superheater average temperature differential signal is used to bias the right and left spray valves to equalize the steam temperature across the secondary superheater section. Since it is necessary to maintain very accurate control of the final steam temperature, a narrow-range transmitter (900° to 1100°F), that is accurate to plus or minus 10°F under all conditions of load change, is cut in when the unit is up to rating and is delivering power. The wide-range transmitter (400° to 1200°F), however, monitors the steam temperatures during the start-up and shutdown procedures. A switch is provided for the selection of the governing signal.

The final steam temperature controller is a computer-controlled set point station that includes proportional, plus reset, plus rate action, and also manual–automatic selection and manual set point. This gives the operator manual override, plus the ability to decide whether the computer can adjust the set point, and it permits him to adjust the final steam temperature set point.

The interstage temperature transmitter sends its signal of interstage temperature at point T-4 to the interstage temperature controller. The set point of this controller is the output of the final steam temperature controller, plus the rate-of-change signal interjected from the steam

Fig. 6. Circuitry features for superheat temperature control system on a tangential-firing, divided furnace utility boiler.



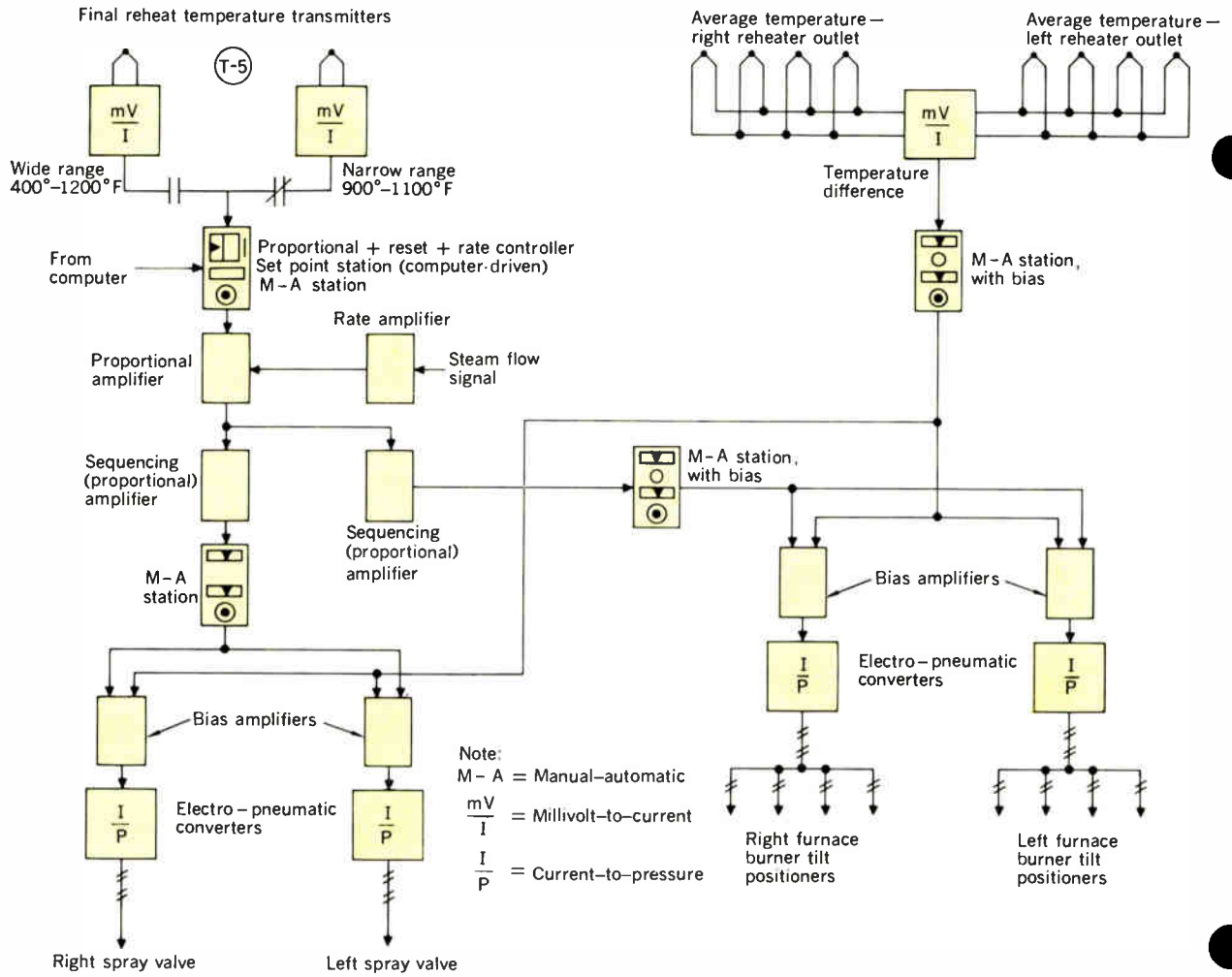
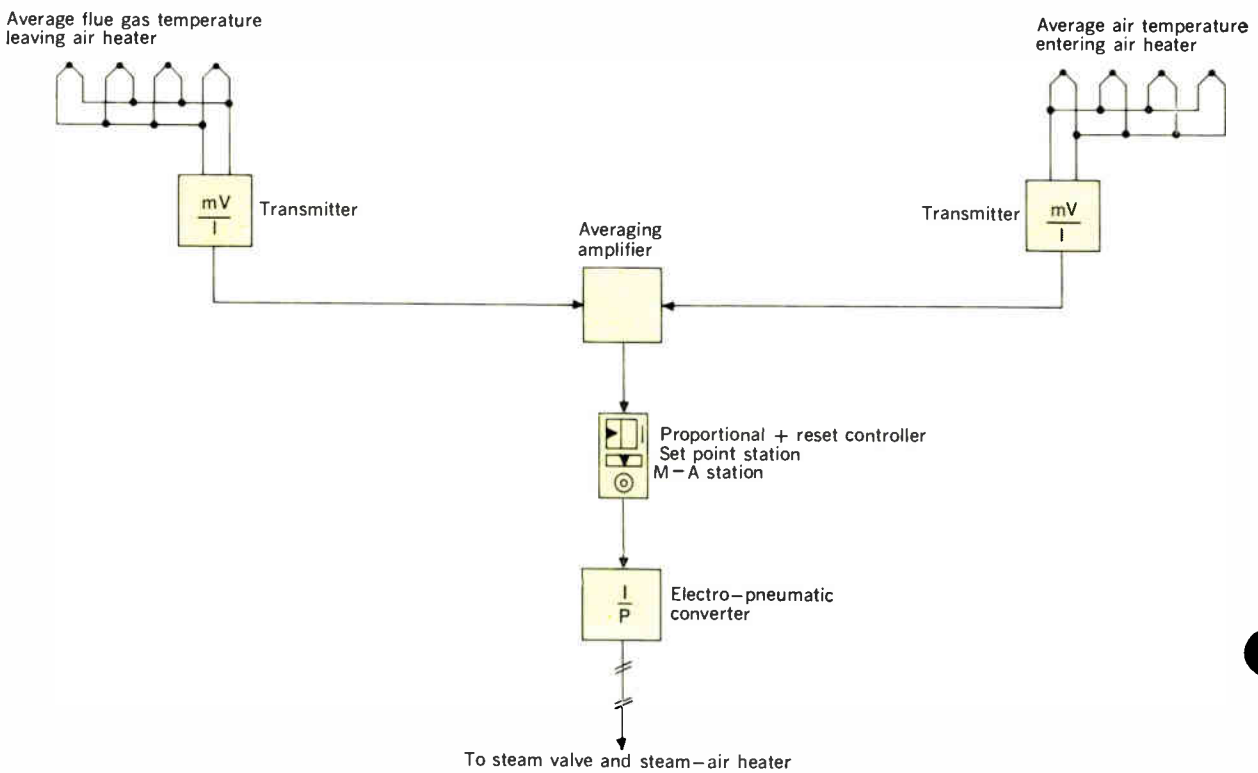


Fig. 7. Circuit configuration for reheat temperature control system for a boiler similar to that of Fig. 6.

Fig. 8. Air heater temperature control circuit.



flow. The output of the interstage temperature controller is transmitted through the bias amplifiers to the electro-pneumatic converters at the right and left spray valves.

The bias amplifiers permit the interjection of a corrective signal from the right and left superheater outlet average temperature system. Here, four or five thermocouples in parallel give the average temperature over several superheater tubes. By combining the signals of these two averaging networks, the temperature difference output signal becomes a function of the difference in temperature at the right and left superheater outlets. This signal passes through a manual-automatic station with bias to the bias amplifiers so that it affects the right and left spray valves in opposite directions. For example, an increase in signal would decrease the right spray valve and increase the left valve, and this has the effect of equalizing the temperature at the superheater outlet. The manual-automatic station with bias is included between the temperature difference amplifier and the bias amplifiers for two purposes: it allows the operator to take this system out of service, and to adjust the temperature system manually for the repair of thermocouples, etc.

Reheat temperature control. The reheat steam is actually the exhaust from the high-pressure turbine that is admitted to the reheat tubes of the boiler for additional heating, from which it is returned to the turbine cycle. Reheat temperature control, as shown in the Fig. 7 circuit configuration, depends upon the burner tilt angle as the main adjustment of steam temperature, with override and emergency control available from the spray valves.

The same overall requirements of control accuracy are necessary for reheat temperatures and for main steam temperature. Thus, on computer-controlled equipment, there are two millivolt-to-current transmitters to provide a narrow-range and a wide-range system. The final steam temperature signal is transmitted to the reheat temperature controller which incorporates proportional, plus reset, adjustments, and also the set point and manual-automatic station. The computer signal (indicated by arrow) can also adjust reheat temperature, but is usually an auxiliary rather than an override control. The primary control signal passes through the proportional amplifier, wherein the rate of change of steam flow can be interjected, and thence into the sequence amplifiers.

Bias amplifiers, which permit the interjection of the temperature difference influence, are included in both the left and right spray valve signal lines and the left and right burner tilt lines. The average temperature in the secondary reheater section—left and right sides—is sensed by the temperature difference transmitter. This signal is then passed through the manual-automatic station with bias, in a manner similar to that for the main steam temperature control. Next, the signal is interjected at the bias amplifiers for both the spray valves and the burner tilts to balance the temperature across the secondary reheater section.

Essentially, this type of boiler comprises two large furnaces with separate fuel firing and adjustable burners. Tangential firing means that each corner of the furnace has one row of vertically mounted burners that fire in a clockwise, tangential direction. To obtain an optimum admixture of fuel and air for combustion, the burner rows can be raised or tilted vertically by a large electro-pneumatic cylinder actuator. By lowering the burners, less radiant heat is absorbed by the reheater section and more heat is absorbed by the furnace water wall tubes. If the burners are raised, the reverse effect takes place. Thus the tilting burners represent the primary method of reheat temperature control.

There is a manual-automatic station with bias in the signal circuit to the burner tilts. It is included to permit the operator to interject a bias influence between the spray valves and the burner tilts. By adjusting the bias knob, he can increase or decrease the burner tilt angle with respect to the spray valves. This has the advantages of allowing the operator to adjust the general range of burner tilt operation, of keeping the burner tilts in range, and of keeping the spray water out of service. In this type of boiler, reheater spray water should not be used more than is absolutely necessary. And as the reheater section picks up more slag, it may be desirable to utilize the bias adjustment to maintain the burners in a normal operating pattern.

Air heater temperature control

Large utility boilers incorporate either tubular or rotating air heaters. These are heat recovery equipment designed to extract some heat out of the stack gases and return it to the combustion air. Conditions can exist, however, where both the combustion air temperature and the stack gas temperature are fairly low. But the efficiency of the heater is very high, thereby reducing the gas temperature—which leaves the air heater to cope with a dangerously low value. And danger is present if the flue gas temperature is lowered below its dew point. Then, sulfur dioxide and other corrosive gases will cause the corrosion of the air heater elements and the stack.

Air heater temperature control is primarily designed to preclude the lowering of flue gas temperature below a given safety point. Since the air temperature into the boiler varies considerably with ambient temperature, and the flue gas temperature varies to a large extent with load, it is believed that the safest operating method is to establish the set point at some average between the air inlet temperature and the flue gas outlet temperature.

Air heater control circuit. The control circuit shown in Fig. 8 indicates four thermocouples operated in parallel to obtain average flue gas temperature. Four thermocouples are also used in parallel to monitor the average air temperature into the air heater. As in some other control systems, transmitters for flue gas and air temperatures send their signals to an averaging amplifier, which is actually a summation amplifier with a gain of 0.5. The average of the two signals is then transmitted to the air heater temperature controller. This unit contains proportional, plus reset, adjustments, a set point, and manual-automatic station. The output is next transmitted to an electro-pneumatic converter mounted on a steam control valve. This valve admits extraction stage steam from the unit to the steam coil air heaters that are built into the inlet of the boiler air heater. For simplicity, the Fig. 8 diagram shows only one control circuit; however, two sets of equipment are normally required for large utility boilers since there are two separate air heaters.

Pressure reducing and desuperheating control

Steam pressure reducing and desuperheating systems are usually installed together. In most instances when

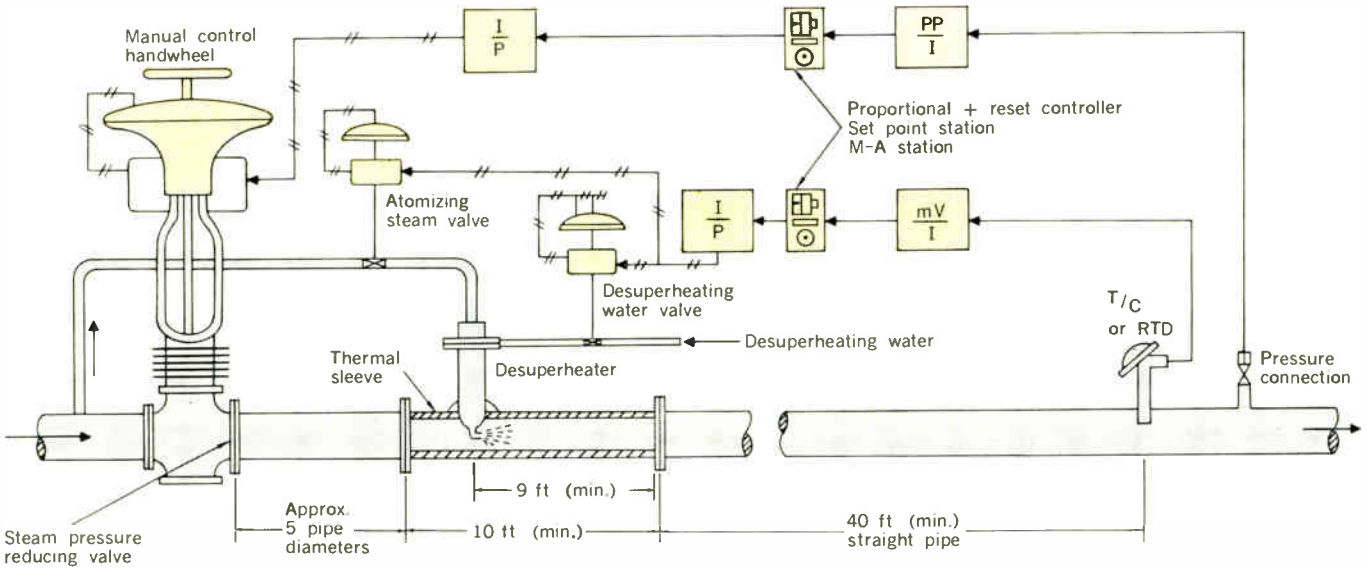


Fig. 9. Steam pressure reducing and desuperheating controls

steam pressure must be reduced for a lower pressure application, a lower steam temperature is required. But there are cases in which either pressure reduction or temperature reduction alone is necessary. Since the two systems are independent, either may be installed without the other.

Steam pressure reducing control. The normal steam pressure reducing control, indicated in Fig. 9, maintains the pressure downstream of the valve at some fixed value. This requires a pressure transmitter, two-mode controller, manual–automatic transfer station, and a control valve.

By connecting the pressure transmitter upstream of the valve, the same system can be used as a spillover control to maintain upstream pressure through the admission of the necessary quantity of high-pressure steam to a lower pressure receiving system.

Desuperheating control. Temperature reduction of steam is accomplished by spraying water into the steam through a nozzle called a *desuperheater*. In the carburetor-type desuperheater, the nozzle is designed to atomize the water for a good mixture, or carburetion, of the water and steam. If the flow range is not too great and precise temperature regulation is not necessary, this type of desuperheater is satisfactory. For more precise regulation and operation over wide changes in flow, however, the steam-atomizing desuperheater, in which high-pressure steam is injected into the unit, affords greater efficiency.

Since there is usually a large temperature differential between the steam that is being desuperheated and the desuperheating water, a thermal sleeve (see Fig. 9) should be installed in the pipe to absorb the thermal expansion shock and to prevent metal fatigue and possible pipe failure.

The temperature sensor must be located far enough downstream of the desuperheater to assume that the water and steam are completely mixed and all the water is vaporized. If the steam contains water particles at the point of temperature sensing, erratic monitoring and poor control will occur.

Desuperheating water must be available at sufficient pressure to overcome the pressure drop across the desuperheating water control valve, plus the drop across the desuperheater nozzle. Desuperheating water may be either boiler feed water or that furnished by a separate desuperheating water pump.

Atomizing steam is normally supplied at a fixed rate, and the atomizing steam valve is controlled on an on–off basis.

The control system components are a temperature detector (T/C or RTD), temperature transmitter (millivolt-to-current transducer), two-mode controller, manual–automatic station, desuperheating water control valve with valve positioner, and atomizing steam shutoff valve with a snap-acting pilot for open–shut service. When measuring the flow in a line that is equipped with a desuperheating station, the primary element should be installed downstream from the desuperheater. The amount of desuperheated steam delivered is the sum of the main steam plus desuperheating water plus atomizing steam.

Pump recirculation control

Recirculation control systems are most commonly used with motor-driven boiler feed-water pumps to ensure sufficient flow at all times and to prevent cavitation. Whenever the flow through the pump decreases to a preset minimum rate, the system will trip and reset at some higher rate, and the flow is automatically increased by recirculation. The overlap between trip and reset prevents frequent reversals when operating near the trip point, and it precludes cycling when the recirculation is part of the total metered flow (see Fig. 10).

Normal operation—no recirculation. In normal operation, the recirculation valve is closed. The pump flow is metered with a flow transmitter that operates a high–low alarm switch, which, through a relay, opens and closes the valve in the recirculation line. With both alarm switches closed, the relay coil and the solenoid coil are energized. The three-way solenoid valve applies full air

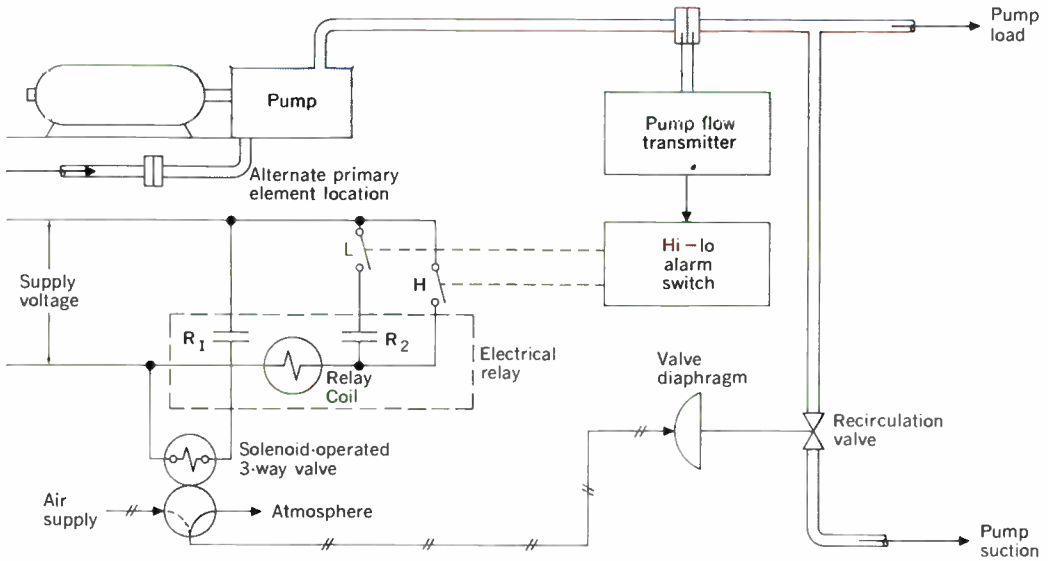


Fig. 10. Pump recirculation control system circuitry.

pressure to a spring-opening diaphragm recirculation valve that keeps the recirculation line closed.

Operation at minimum flow. When the flow decreases, the high alarm switch (H) opens first, but the circuit remains energized through the relay contact (R_2). A further decrease in the flow to the predetermined minimum will open the low alarm switch (L). This de-energizes both the relay coil and the solenoid coil by opening both the R_1 and R_2 contacts. The three-way valve is actuated and vents the diaphragm of the recirculation valve, thereby opening this valve and permitting recirculation.

Operation on increasing flow. Opening the recirculation valve will immediately increase the metered flow and close the low alarm switch, but the circuit will remain de-energized through the open contact R_2; therefore, the recirculation valve will stay open. An increase in pump load is required to augment the metered flow before the high alarm contact will close to re-energize the circuit, and reclose the circulation valve.

As soon as the recirculation valve closes, the metered flow will decrease to some value that is higher than the minimum flow at which the low alarm contact opens. The proper setting of alarm contacts L and H must provide the necessary overlap. Thus the low alarm contact is set at the minimum safe flow rate recommended by the pump manufacturer. The recirculation valve will have a capacity approximately equivalent to the minimum safe pump flow. The high alarm contact will be set at a flow that corresponds to minimum safe pump flow, plus the recirculation valve capacity, plus about 5 per cent of total

I. Panel-mounted recorders—essential for operation

1. Steam flow–air flow, or fuel flow–air flow
2. Drum level
3. Superheat temperature, reheat temperature
4. Turbine throttle pressure, condenser vacuum
5. Excess air, or per cent oxygen

II. Optional panel-mounted recorders

1. Feed-water flow
2. Drum pressure
3. Superheater outlet pressure
4. Reheater inlet pressure
5. Reheater outlet pressure
6. Superheater interstage temperature
7. Desuperheater water flow—superheat
8. Desuperheater water flow—reheat
9. Condensate flow from condenser
10. Boiler feed-water pump suction pressure
11. Boiler feed-water pump discharge pressure
12. Turbine first-stage pressure
13. Air heater air inlet temperature
14. Air heater air outlet temperature
15. Air heater gas inlet temperature
16. Air heater gas outlet temperature
17. Air heater average temperature
18. Economizer gas inlet temperature
19. Economizer gas outlet temperature
20. Feed-water temperature to economizer
21. Feed-water temperature from economizer
22. Fuel pressure
23. Fuel temperature
24. Circulating water inlet temperature
25. Circulating water outlet temperature
26. Conductivity make-up water
27. Conductivity boiler feed water
28. Conductivity saturated steam
29. Conductivity superheated steam
30. pH boiler feed water
31. pH make-up water
32. Generator stator temperatures

III. Panel-mounted indicators—essential as operating guides

Draft Gauges	Vertical Edgewise Indicators	Switchboard Instruments
Forced draft fan outlet pressure	Steam flow–air flow, or fuel flow–air flow	Pulverizer motor, amperes
Air heater outlet pressure	Steam flow–water flow	Boiler feed pump motor, amperes
Windbox pressure	Drum level	Forced draft fans, amperes
Furnace draft, or pressure	Superheat temperature	Induced draft fans, amperes
Boiler outlet draft, or pressure	Reheat temperature	Condensate pumps, amperes
Air heater inlet draft, or pressure	Feed-water pressure	Circulating water pumps, amperes
Air heater outlet draft, or pressure	Drum pressure	Air heater drive motors
Pulverizer operation	Superheater outlet pressure	Other major motors, amperes
	First-stage pressure	
	Condenser vacuum	
	Fuel supply pressure	
	Fuel burner pressure	
	Per cent excess air, or oxygen	
	Desuperheater water flow—superheat	
	Desuperheater water flow—reheat	
	Superheater interstage temperature	
	Turbine extraction pressures	
	Deaerator level, pressure	
	Instrument air supply pressure	

IV. Panel-mounted manual–automatic stations

1. Fuel
   a. Pulverizer positioners
   b. Oil valve
   c. Gas valve
2. Forced draft fans damper positioners
3. Induced draft fans damper positioners
4. Feed-water control valve
5. Steam–air heaters steam valves
6. Steam temperature control positioners
7. Boiler master
8. Condensate recirculation
9. Deaerator level and pressure control valves
10. Miscellaneous loops, if desired

pump capacity for overlap. If the high alarm contact were set at a value equal only to the minimum pump flow, plus the recirculation flow, it is obvious that the system would continuously cycle the valve open and closed whenever the low alarm setting was reached.

Utility plant instrumentation list

Tables I through IV represent a condensed list of typical utility plant instrumentation (recorders, indicators, switchboard instruments, and manual–automatic stations) that relate to the unit controls for the boilers, turbines, and auxiliaries. Note that either all or most of the optional panel-mounted recorders listed in Table II may be omitted if a data logger or digital computer is installed. Similarly, about ten panel-mounted alarm annunciators may be omitted, as they can be incorporated into the computer or other alarm scan-monitoring equipment.

The unit controls, instrumentation, and monitoring we have discussed for boiler and turbine operation in a utility plant have been necessarily simplified for presentation and explanation. They are, however, generally typical of installations that are now operational in the United States and abroad.

The voltage regulator controls are part of the generator control system, and will be discussed—together with automated controls for relays and circuit breakers—in Part II of this article.

Figure 11 is a more detailed diagram of the process control computer elements[4] shown in Fig. 1. As indicated, the entire system is divided into three parts: the base computer—including control, arithmetic, and memory units—plus the input–output control; a peripheral section that handles communications to and from the process, and a programmer's console through which programs are entered, started, debugged, etc.; and the process itself, along with the operators' stations.

Analog inputs are read in a programmed sequence by time multiplexing through one or more input channels. Low-level signals are amplified, and all are converted to numerical form by an analog–digital converter. Contact inputs are read similarly; however, no conversion is necessary since these data are already in binary numerical form.

Contact outputs operate logging typewriters, punches, digital displays, alarm and trend recorders, and control sequencing operations.

Figure 12 shows a simple flow chart that reads an input into memory, converts it to engineering units, checks it against normal operational limits, and performs a simple calculation. The actual assembly language coding is listed in Table V.

Computer-controlled power plants—domestic

There is an ever-growing list of computer-controlled power plants that are in operation, under construction, or in the design stages.

While it is impossible either to enumerate or describe

the features of all such plants, it is believed that the following three plants are representative of typical systems and installations in the United States.

Plant Jack McDonough. Plant Jack McDonough is the Georgia Power Company's newest addition. It consists of two 250-MW tandem-compound, double-flow turbine–generators, rated at 2400 psig, 1000°F with reheat to 1000°F. Steam is supplied by tangentially coal-fired, controlled circulation boilers that are rated at 1.75 million lb/hr of steam. The cycle has five closed, and one deaerating, feed-water heaters.

The innovations that implement the detailed study and control of the plant's operation are

1. An on-line digital computer with stored program.
2. Advanced design electronic boiler controls.
3. A solid-state events recorder.
4. Instrumentation that has accuracy and long-term repeatability.

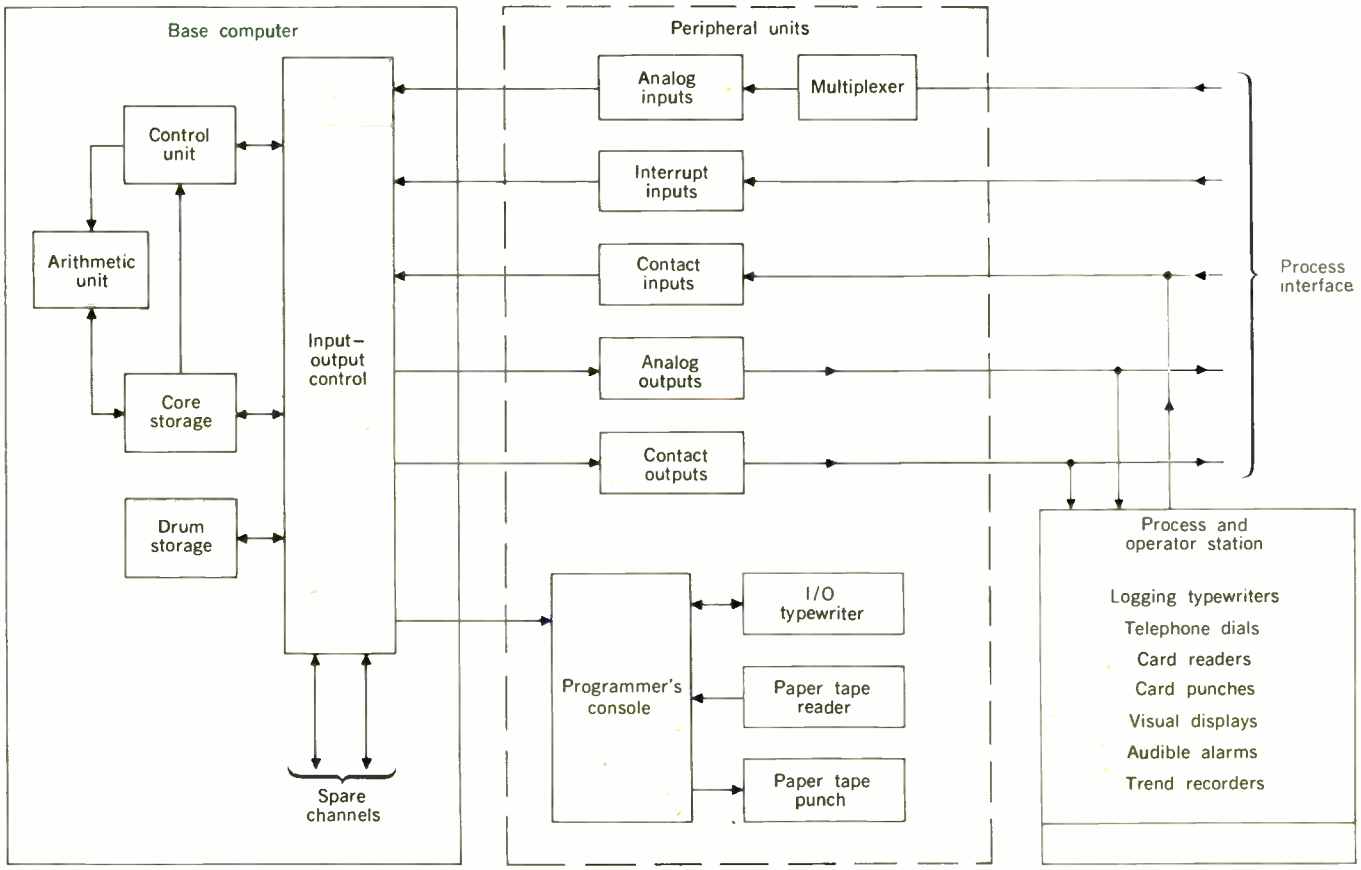
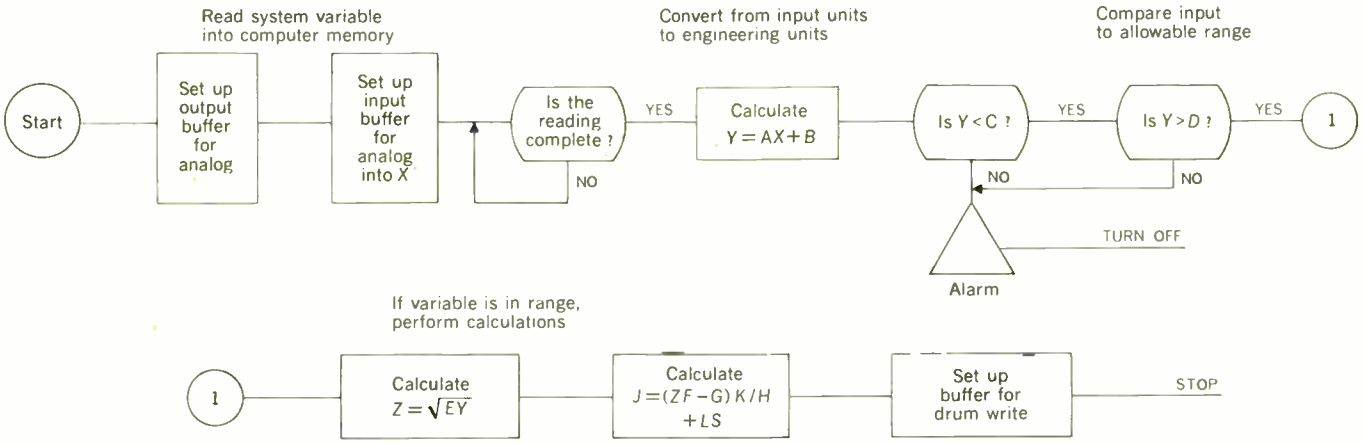


Fig. 11. Detailed diagram of the process control computer elements shown in Fig. 1.

Fig. 12. Simple flow chart. Note that an input is read into memory, converted to engineering units, checked against normal operational limits, and finally a simple calculation is automatically performed.



V. Assembly language coding

Location of Instruction	Operation	Address	Comments
	ORG	512	
START	OUT	6	Set output buffer without interrupt
BUF	VFD	3, 15	
	BUF	0, ADRS + 1	
	BUF	0, ADRS	
	INT	6	Set input buffer without interrupt
	BUF	0, X	
	BUF	0, X	
	TSI	6	Test for channel activity
	DJP	L-1	
	ENL	X	Proceed if channel is inactive
	MPL	A	
	ADD	B	Y in accumulator
	CPL	C	
	JGE	ALARM	Test limits
	CPL	D	
	JEQ	ALARM	Test limits
	JOL	ALARM	Test limits
	MPL	E	
	DRJ	SQRT	Go to square-root subroutine
	STL	Z	Store Z
	MPL	F	
	SUB	G	
	MPL	K	
	DIV	H	
	ADD	LS	
	STL	J	Store J
	EXF	7	Call drum
	BUF	O, YOW + 1	
	BUF	O, YOW	
	OUO	7	
	OUT	7	Store Z and J into first two locations of track XXX
	BUF	O, Z + 2	
	BUF	O, Z	
	TSO	7	
	DJP	L-1	
	STK	32	
Z	OCT	O	
J	OCT	O	
ADRS	OCT	GAINAD	Gain and address of point X
X	OCT	O	
YOW	OCT	600XXX	Code for writing into track XXX

The computer is used to monitor the operation of the plant, and its functions include the monitoring of equipment for off-normal conditions, accurate determination of component and plant performance, the monitoring of the performance of the boiler controls, changing the control set points when necessary, and communication with operating and engineering personnel.

An event recorder with a memory is incorporated into the system to monitor the sequence of events during abnormal operation. There are 256 active inputs, and a 20-point memory that has a time resolution of about one millisecond. Thus the time relationship of abnormal conditions of any 20 of the inputs can be stored. This memory is transferable to the computer at a rate of approximately 300 points per second, and as a memory is cleared, it is instantly available for alarm storage. Therefore, if all 256 inputs were to become off-normal in 20 ms, only the first 20 would be printed out with time correlation; the remaining 236 would be printed out as off-normal, but without time correlation.

The computer inputs required for the monitoring of steam-electric plant performance are of three types—analog, digital, and pulse.

As we have seen in the description of boiler controls, all analog inputs require conversion to digital quantities. Also, the thermocouples require a reference junction, the temperature of which may be either floating or controlled.

The reading of the pressure transducers requires two sequential computer operations. First, a subcontrol output action is essential to select a particular pressure transducer. Then, after a time delay, a priority interrupt signals the computer program to read the already digitized result by a fast digital scanner.

The pulses from the kilowatthour meters are buffered into pulse accumulators that are periodically read and accumulated in the core memory by the computer program.

Data concerning the fuel analysis are inserted by the use of a "chemist's console." The results of the per-

formance calculations, together with supporting data such as the feed-water drain cooler and terminal differences, average throttle and reheat temperatures, pressure flows, etc., are printed out on the console typewriter. Other information, such as boiler efficiency and high-pressure and intermediate-pressure turbine efficiencies, will also be available for display.

The performance calculations are divided into three main areas: boiler efficiency, plant heat rate, and correction to contract or reference cycle. But before data are gathered for any of the calculations, certain test criteria must be met to ensure that the unit is in a stabilized operating state. These tests are of two types: a *start criteria* and a *continue criteria*. The former is used to ensure that the unit is in a normal operation condition, and once this set of criteria is satisfied, the unit is allowed to "soak" in this condition for a minimum of 30 minutes for all transients to settle out. During this soak period, the continue criteria are applied every 30 seconds to ensure that the unit has not moved to a new operating point.

After the soak interval, a data-gathering phase is entered, during which the continue criteria must still be met. If the continue criteria test fails at any time during either the soak or data-gathering phases, the program is recycled to the start of the criteria tests. Upon completion of the data-gathering phase, the program transfers to a calculation phase in which conversion to engineering units and the performance calculations take place.

Each piece of equipment has been analyzed to determine the start and continue criteria. For example, if the main steam temperature is changed and thereafter remains within very narrow variations, it will take approximately 30 minutes for the temperatures of the turbine metals to reach their new levels. If data are gathered for 30 minutes thereafter to average out any minor changes of variables, it will take at least one hour to obtain the results of a turbine test after the set point of the unit has been attained and the transient conditions settle out. This time requirement, however, is not detrimental for turbine testing since complete turbine performance is not needed on a more rapid time interval.

Boiler performance calculations by the heat loss method require a knowledge of the chemical constituents and heating value of the fuel, but an accurate coal analysis is difficult. Time-scheduled calculations are made, using an average fuel analysis, and these are refined periodically. For the accurate calibration of boiler efficiency that is needed to obtain plant heat rate, coal samples are obtained from the pulverizers at the same time that the computer is gathering other data. At the end of the test period, the computer punches its data on paper tape. The fuel sample analysis and other data are read into the computer, and the results are calculated. By this method, accurate plant heat information is available about 24 hours after actual performance.

The use of an on-line digital computer and accurate instrumentation for Plant Jack McDonough is part of a continuing effort to utilize available technology to improve system performance.

The Georgia Power Company is one of the four parent companies of the Southern Company system, whose transmission systems are interconnected. The power generation requirements of the system are determined, and in most cases automatically controlled, by an economic dispatch computer located in Birmingham, Ala.

Cholla Power Plant. The Arizona Public Service Company's Cholla Power Plant[5] is the first coal-fired steam-electric station in the state, and the plant is situated in the middle of Arizona Public Service's 300-mile-long 345-kV transmission line.

Coal of about 10 600 Btu/lb is mined near Gallup, N. Mex., and it is shipped by rail to the plant site. The station was originally designed for two units, but it is presently operating with a single 115-MW turbine–generator unit.

Many new labor-saving devices are incorporated into the station; they include semiautomatic coal handling facilities and a monitoring and results computer system (MARC). The MARC system not only assists the plant operating personnel but also monitors and logs the 230- and 345-kV substation and transmission line data.

The application or design philosophy of the computer system at Cholla is to provide the operators with the maximum amount of information and operating aids without controlling any plant functions. Within this framework of definition, the computer provides

1. The scanning of approximately 375 analog inputs and 75 contact closure inputs.
2. An alarm system for abnormal operating conditions, with relay outputs for 67 annunciator alarm windows.
3. The automatic logging of plant information.
4. Special on-demand data gathering, and the reporting of functions as called for by the operator.
5. Performance calculations at three-minute intervals to guide the operator toward efficient plant operation.

Information from all the plant sensors and calculated performance is immediately available to the operator in the control room. Since the digital computer is completely dependent upon accurate information, the plant sensors become a vital link in the overall system.

Figure 13 shows the major components in the Cholla MARC system. Note that the termination cabinets for all sensors are located in the cable vault below the control room. The cabinets contain the thermocouple inputs, low-level (below 100 mV) sensor inputs, and the contact inputs and outputs, plus the alternating current inputs. This arrangement completely isolates the low-level and high-level signal terminations.

The plant computer system consists of four cabinets to contain the digital computer, which has 8192 words of core memory and 16 384 words of bulk drum memory (this unit is a single address, parallel binary logic, and has a 20-ms word time); one analog scanner cabinet; one mercury-wetted relay matrix cabinet; one power distribution cabinet; and one digital fast scanner cabinet.

A separate computer room contains the computer console, a 100-character-per-second paper tape reader, a printer, and a 100-character-per-second paper tape punch. Also, a second on-line parallel printer is mounted on a separate desk.

The computer console was designed for both program debugging and program maintenance. The console allows single-stepping through a program either by instructions or by word times. All registers in the computer and the analog section may be displayed by using a

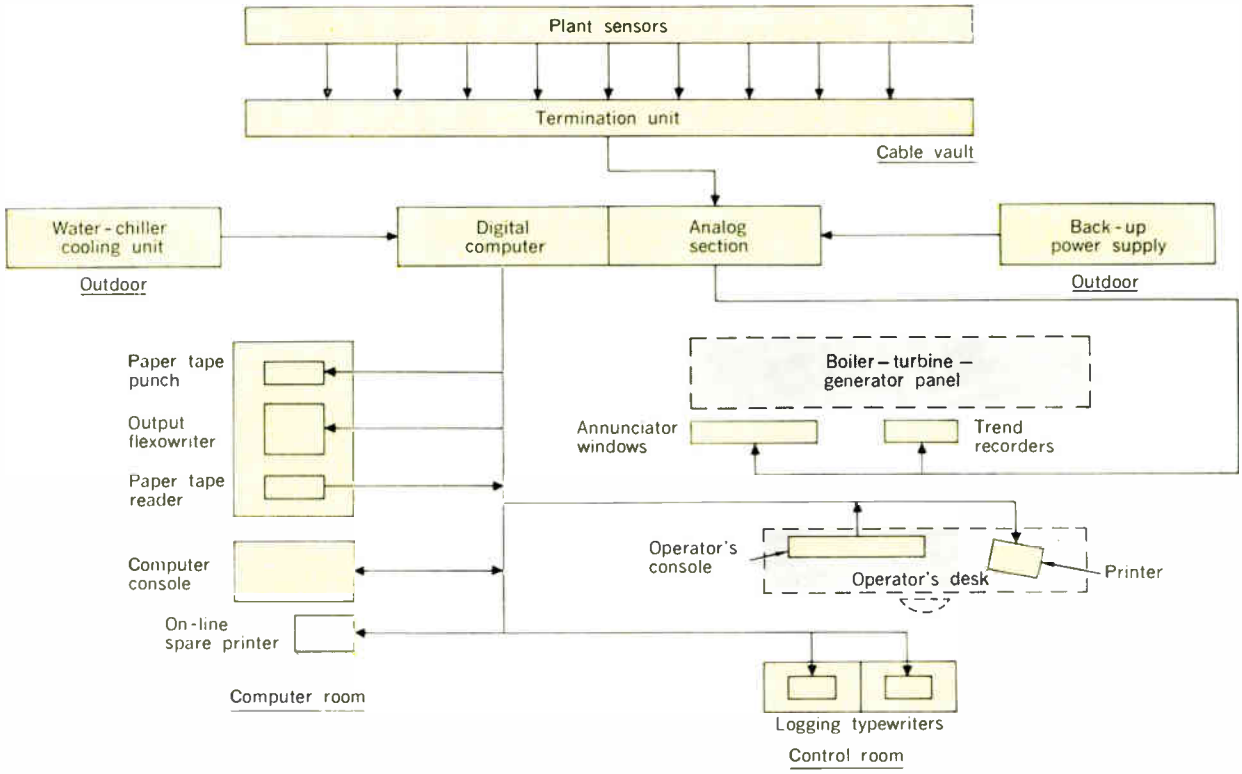
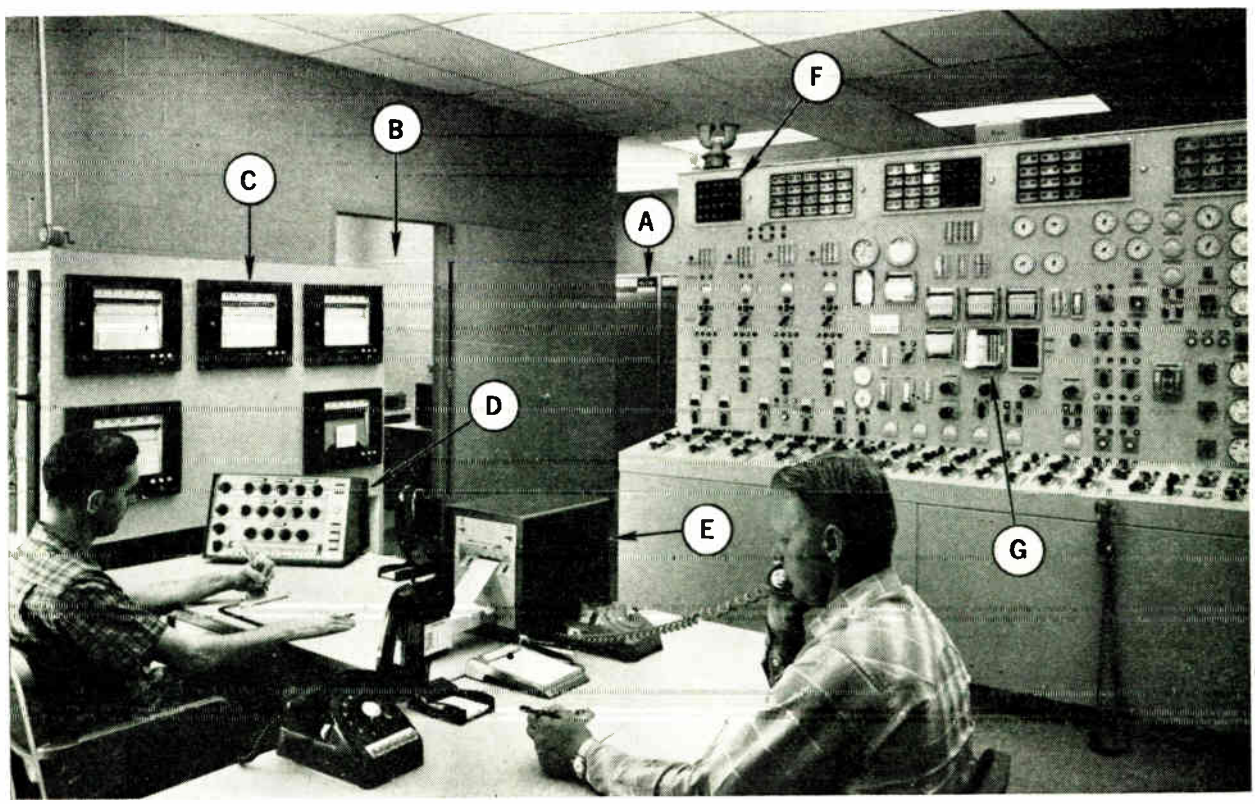


Fig. 13. Block diagram of the Cholla Power Plant computer system

Fig. 14. Control room arrangement at Cholla. The eight computer cabinets (A) are located behind the large boiler-turbine-generator board, and the computer room is at (B). The point recorders (C) were provided as temporary instrumentation during the interim period between the initial roll date and installation of the computer system. They have since been removed. The operator's console (D), with its associated parallel printer (E), are located on the operator's desk. The trouble location annunciator (TIA) window groups are at (F). Finally, the location of the two MARC-controlled trend recorders is shown at (G).



selector switch; this includes the priority interrupt, memory, drum transfer, input–output, and all other internal registers. A key is provided to lock the console so that the computer program functions cannot be accidentally disrupted.

Operational data from more than 360 sensors, plus supplementary computational values, converge in the control room (Fig. 14) in the form of flows, temperatures, rates of change, differences, and efficiencies to provide instantaneous and continuous operational data for each major component of equipment associated with the turbine–generator unit.

Temperatures make up about 75 per cent of the system inputs. Monitoring of the turbine–generator, motor, and pump bearing; boiler drum, tube, and header metal conditions; plus the steam and water cycle status, proceed on a continuous basis. The remaining 25 per cent of the system inputs are distributed among pressures, orifice differentials, levels, turbine supervision, water analysis, and plant and connected switchyard electrical values.

Three basic scan classes are contained in the system. Each sensor is assigned to a given class as a function of its priority to the overall system, and all sensors within a given class scan are read once during the assigned time interval, as shown.

SCAN CLASS	INTERVAL
1	4 s
2	10–20 s
4	1–2 min

Scan class 1 is utilized to dampen inherent cycle oscillations from high-priority inputs such as the feed-water orifice differential reading.

Scan class 2 contains sensors that are used as inputs for the efficiency computations, the data from which are averaged over a three-minute time interval.

Scan class 4 is the monitoring and alarm cycle.

The basic communication link between the operator and the MARC system is the point identification (Point I.D.).

Five alphanumeric characters constitute the general format, and are arranged as

where: Y = major plant group (alphabetic)
I = unit number (numeric)
XX = group item number (numeric)
Z = type character (alphabetic)

Inputs (sensors) and calculated values (efficiencies, etc.) are divided into two basic groups or point series for identification purposes. For example, the 100-series Point I.D. numbers encompass all inputs, and the 500-series Point I.D.s represent all calculated values. The following are illustrative of this alphanumeric coding:

SENSOR
A091T—Ambient air dry bulb
- (plant)
- Type symbol (temperature)
- Group sensor number
- Denotes plant common sensor
- Auxiliary group

SENSOR
C110P—Feed-water orifice differential
- Type symbol (pressure)
- Group sensor number
- Unit no. 1 sensor
- Condensate and feed-water group

CALCULATED VALUE
D525 —Boiler efficiency
- Blank
- Group calculation number
- Unit no. 1 calculated value
- Draft system calculation sub-group

VI. Sample of alphanumeric listing of sensors

Code	Description	Units	Scale	Alarm Limit -S-V-	Scan Class -1-2-4	Logged L TLA-MB-F-T-G
C121	Cond to IP htr	F			4	
C122	Cond from LP htr	F			4	
C123	Cond to LP htr	F			4	
C124	Cond from SJAE aftercond	F			4	
C125	Cond from cond hotwell	F			2	
C126	Condensate storage tank level	FT	V/0.1	S	4	F1–9
C128	Deaerator storage TK temperature	F			2	
C130	Crossover heater orifice diff.	Corr. MV	V/100		2	
C131	Hot water heat exchanger orifice diff.	Corr. MV	V/100		2	
C132	Crossover htr drain temp	F			2	
C133	HP htr drain temp	F			4	
C134	IP htr drain temp	F			4	
C135	LP htr drain temp	F			4	
C136	Gland steam cond drain temp	F			4	
C137	SJAE drain temp	F			4	
C139	Hot water heat exch drain temp	F			2	
C140	BF pump disch. header pressure	PSIG			2	
C141	Superheat spray orifice diff.	Corr. MV	V/100		2	
C142	Reheat spray orifice diff.	Corr. MC	V/100		2	

The explicit identification of a given Point I.D. is made by either a schematic input–output diagram, or an alphanumeric listing of sensors and calculated values as shown in Table VI.

A preprinted log format provides operational and historical data of the unit performance. Fixed log data are printed out each hour, with selected values either totaled or averaged over a 24-hour period. Performance data are included for the overall unit, turbine, generator, boiler, condenser, auxiliaries, and the analysis group—plus system flows, fuel totals, and energy accounting totals.

Included with each hourly printout are three indexes that are related to the plant and computer performance:

1. The accuracy of the computer logic, and the accuracy of analog to digital conversions.
2. An index of the magnitude of generator load variations.
3. A reference list of boiler transient conditions.

The dispatch portion of the hourly log (punched paper tape output) provides data on the tie lines, plus plant energy accounting.

The MARC system at Cholla represents an important step toward centralized control. Plant operators now have available, on a continuous and instantaneous basis in the control room, all pertinent data relative to the plant status. This initial step provides supporting data, plus performance computation to assist in optimizing plant operation.

The final phase of the evolutionary process will lead to the controlled operation of the steam-electric unit entirely by the digital computer.

Keystone steam-electric station. The Keystone Station will be a mine-mouth generating plant to take economic advantage of contiguous deposits of high-grade bituminous coal in sufficient quantities for a 30-year plant life expectancy.

Conveyors will bring about 60 per cent of the coal from four new mines to be opened on the site and in the immediate vicinity. Also, conveyor belts will transport coal to and from a 500 000-ton fuel reserve storage area that will be serviced by a large-capacity, automated stacker and reclaimer unit of advanced design. The remainder of the fuel will be brought by rail and by truck. Keystone is being built in Armstrong County in Pennsylvania. The site is 30 miles northwest of Johnstown, and about 35 miles northeast of Pittsburgh (see Fig. 15). Participating in the plant construction are the Pennsylvania Power & Light Company and six major investor-owned utility companies of the Pennsylvania–Jersey–Maryland (PJM) interconnection. More than 700 miles of 500-kV EHV transmission lines will be used eventually to interconnect the power system of seven states.

Two 6 350 000-lb/hr generators—the largest "single furnace" steam generators ever built—will supply supercritical-pressure steam to two (Units 1 and 2) 900-MV cross-compound turbine–generators.

Two digital computer systems will be used for performance monitoring, alarm scanning, data logging, and efficiency studies on the no. 1 and no. 2 generating units. Additionally, the systems will provide sequence monitoring and operational guides for start-up and shutdown of the units.

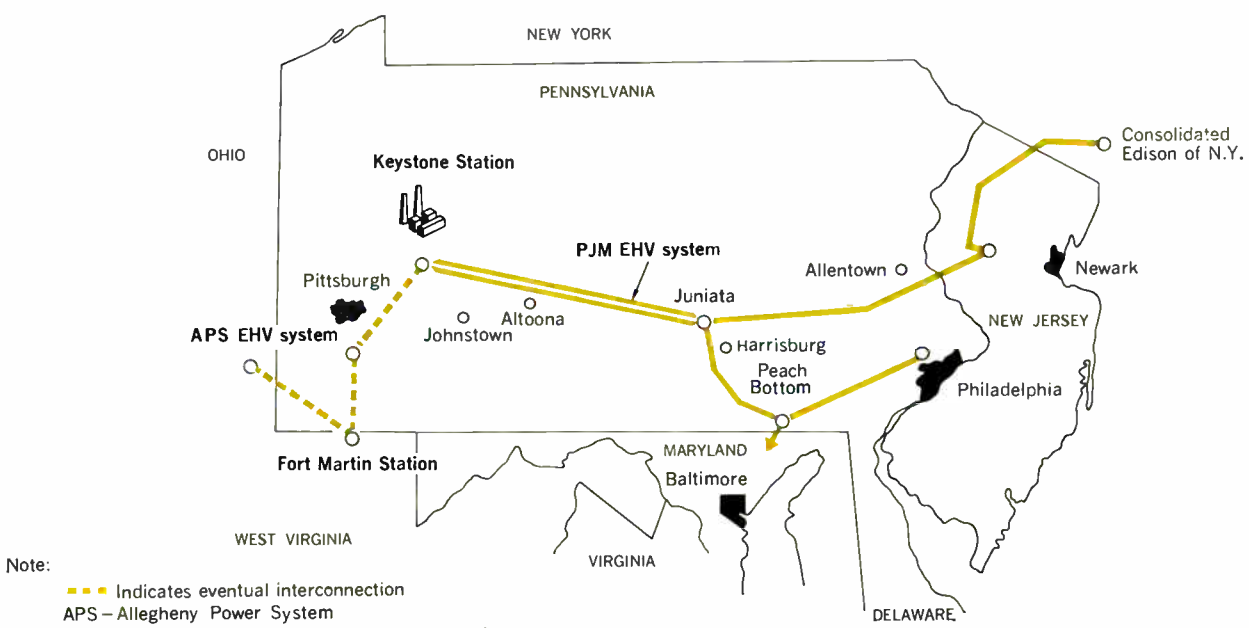
The boiler control system will utilize the direct-energy-balance (D-E-B) method of control. With this technique, the automatic regulation of combustion and feed-water supply to the boiler is coordinated with the turbine governor to furnish the desired electrical output.

The Keystone Unit no. 1 is scheduled to go on the line in the spring of 1967, and Unit no. 2 will follow in 1968.

An overseas computer-controlled power plant

The General Electricity Generating Board of Great Britain has ordered a complete automation system to control the new 2000-MW oil-fired steam-electric station now being built at Fawley, England.

Fig. 15. Area map showing location of the Keystone steam-electric station and the major grid interconnections it will eventually serve.



The unique and extensive application of automation to a British oil-fired power station is based on a newly developed computer that has capabilities beyond the range of machines designed for the usual industrial process control. Four of these computers will control the start-up and run-up to full load of the station's four boilers and four 500-MV turbine–generators. The computers will then monitor the running plant and supervise operations on a 24-hour daily basis.

The digital and display equipment will provide instantaneous warning of any malfunction and correct the fault automatically; or, if the failure is serious, the affected component or unit will be shut down. Also, the computers will continuously record the on-line performance of all equipment units, and will provide accurate performance figures for the entire power station for economic efficiency computations.

During these operations, television screens in the power station control room will show, in writing, every action taken or warning produced by each computer.

Each computer will monitor more than 2500 sensor points that provide operational information, and precise readings will be taken from each point at intervals varying from one-half to five seconds.

Now let's look at the bugs

Near the beginning of this article a section was devoted to the advantages of automatic control and supervisory systems. Thus the reader was forewarned that where there are explicit advantages, the disadvantages—and other problems—may be implicit.

First of all, computer equipment represents a sizable initial capital investment. This, in itself, may be a major disadvantage to the smaller private or public utility groups. Considerable practical knowledge, however, has been acquired—and some lessons have been learned—from operational computer systems at a number of power plants.

Thorough and adequate advance planning is an essential prerequisite prior to the actual construction of a computer-controlled plant. Both the complexity of power plant automation and the probability of many changes in the intended operation must be realized. The design of a computer program and procedural strategy must be flexible enough to accommodate such revisions. When the preliminary planning has been accomplished, a firm schedule must be established to ensure the timely factory check-out and shipment of the computer system to permit its installation, together with ancillary components, at the plant during construction. And all tests should be completed before the plant goes on line.

Adequate field tests should be conducted both to verify pre-established parameters of operations and to ensure that hardware and software perform as intended. Finally, plant operators and supervisory personnel must be thoroughly proficient by training and preliminary "dry run" experience to preclude as much human error as possible when the automated plant becomes operational. This last point is particularly important, since operational experience has revealed some predictable shortcomings in the method of communication between computers and operators.

Operational problems. A composite list of the "feedback" from several computer-controlled plants reveals that during testing phases:

1. It was necessary to use a long-range speed control function for initial turbine roll and for acceleration to synchronous speed, since the computer-controlled simulation of the operator's manual method of rolling the turbine off turning gear was found to be too inflexible to meet every contingency.

2. Flexibility of shutdown procedures had to be increased by transferring the burner shutdown program phase to separate subroutines that are actuated by the main shutdown program. A similar program change increased the operational flexibility of the boiler circulating pumps during start-up and shutdown.

In the computer system it was found that

1. During a rapid restart, a conflict could occur between the emergency shutdown routine and the start-up program because of the simultaneous use of common computer coding.

2. The scanning rates for some analog inputs were found to be initially too slow for trimming specific control functions, and overshoot resulted.

And in the unit control–computer intertie, it was discovered in one instance that the procedure for transfer from the fuel control mode that is used in warm-up to the normal pressure control mode for on-line operation produced a temporary increase in furnace burner firing rate.

In most instances, however, all or most of these operational problems have been overcome by remedial action and adjustments of both hardware and programming.

Some final thoughts and conclusions

It is quite apparent that computer-controlled power plants, despite the sizable costs of the initial investment in analog and digital equipment, have achieved noteworthy economics in both power generation and plant maintenance. And the advantages certainly outweigh the disadvantages. Reports from private and public power groups throughout the United States and abroad indicate an accelerating trend toward at least supervisory control and monitoring of boiler–turbine–generator units within individual steam-electric generating plants. Many of the large utility combines, with extensive and complex interconnections, have already expanded the concept to elaborate load dispatch area control systems that include analog–digital computers in on-line operation with microwave and telephone channel communications networks. Part II of this series will attempt an in-depth exploration and discussion of area control systems.

The author wishes to acknowledge the courtesy and cooperation of English Electric Corporation, the General Electric Company, and the Westinghouse Electric Corporation in making available equipment data and illustrations for the preparation of this article.

REFERENCES

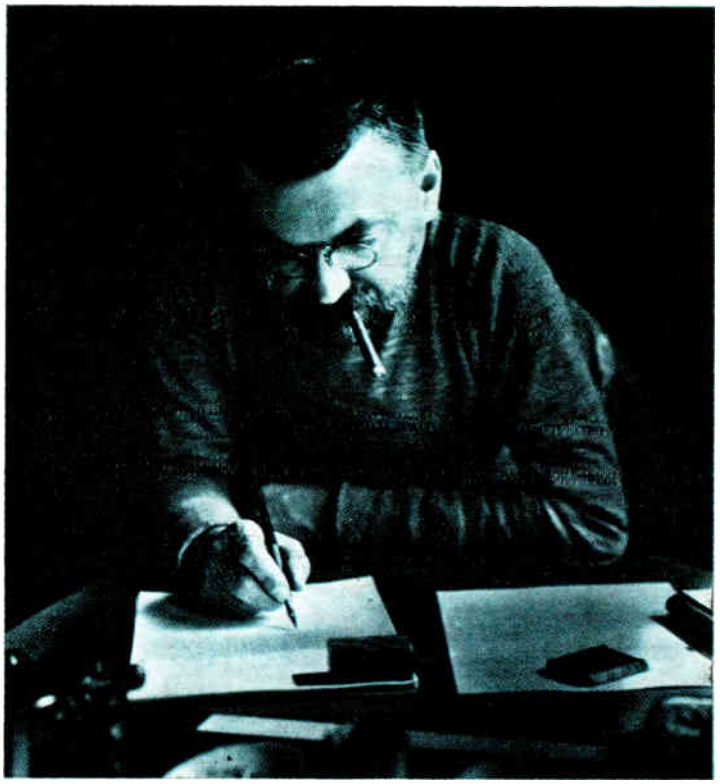
1. Ayers, C., "Automation in the Power Station," *English Elect. J.*, vol. 19, no. 5, Sept.–Oct. 1964, pp. 12–13.

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

3. Paulson, R. E., "Instrumentation and Control for Steam Power Plants," General Electric Co. Manual, 1961.

4. Enns, Mark, "Automation in the Power Industry," Electric Utility Engineering Dept., Westinghouse Electric Corp.

5. Burdick, E. J., and Tickle, D. E., "An On-Line Computer System at the Cholla Power Plant Aids Plant Operation," reprinted from *Proc. of Power Industry Computer Applications Conf.*, Phoenix, Ariz., April 1963.



Charles Proteus Steinmetz

One hundred years ago this month, Karl August Rudolph Steinmetz was born in Breslau, Germany. When he died 58 years later —as Charles Proteus Steinmetz—he had achieved little popular fame. But he had contributed substantially to the electrical engineering art. On the occasion of this anniversary of his birth, we present a three-part tribute to Dr. Steinmetz. The first two parts give an insight to Steinmetz the man; the third, first published in 1922, is indicative of Steinmetz' engineering foresight. Our appreciation is extended to the General Electric Company and, particularly, to Professor P. L. Alger, who has permitted us to use material from the forthcoming book on Steinmetz (Steinmetz—His Philosophy: The Philosophy and Social Views of Charles P. Steinmetz) that he has co-authored with Ernest Caldecott.—Ed.

Steinmetz revisited: The man and the myth

*C. D. Wagoner**

Who was the Steinmetz I knew? a self-satisfied electrical wizard? a twisted hard-bitten man, isolated from his fellow man? Although a common stereotyped image, this is not the Steinmetz I knew.

The man who created artificial lightning was deathly afraid of electricity; the slightest shock from a harmless low-voltage circuit made him jump with alarm; yet this same man, who could not swim a stroke, spent a major portion of his summer days paddling and drifting on the Mohawk River in a 12-foot tippy canoe with no apparent fear.

These were some of the peculiarities of Dr. Charles Proteus Steinmetz, world-famed electrical wizard, who spent a lifetime investigating the various ramifications and mysteries of electricity. So far as we know, he never experienced any narrow escapes in his experiments. His knowledge of electricity's dangers probably created the fear. Rain or shine, he always wore thick-soled rubbers when working in his laboratory. He never explained why, but his associates thought he wore them as an insulator should he accidentally come into contact with an electric circuit.

On the other hand, he liked the solitude of the water. It afforded him the opportunity to read or work out a mathematical problem unmolested. He frequently explained, "It is so quiet and peaceful on the water." And there he solved many of his important mathematical problems.

A friendly acquaintance begins

As a newspaperman in Schenectady, I knew Steinmetz. But our long acquaintance really began when I became a publicity man for General Electric. There Steinmetz was chief consulting engineer and often referred to as the Supreme Court of the Company. I had been with the Company only a few months when my boss called me into his office. He explained that he had a page-one story: The Company had just sold the largest turbine ever undertaken—a 60,000-kw machine for the Commonwealth Edison Company in Chicago.

I left the boss's office thoroughly bewildered, wondering how the big news could get more than a paragraph or two on the financial pages of the newspapers. I thought: There's nothing glamorous about a turbine. Steam goes into one end and electricity comes out the other. You see no action, no wheels turning, nothing but a huge metal casing emitting a low humming sound.

Electricity is something most people take for granted. Only the technically educated minority knows or cares what constitutes a kilowatt of electricity. I was among the vast majority ignorant of the term.

My problem: Write the story with popular appeal. An engineer friend suggested I see Steinmetz. In my newspaper days I had heard how this great man thrilled neighborhood children with explanations of what made the stars appear to twinkle, what made the wind blow, what made flowers bloom. But I had no occasion to consult him on a personal problem. Now with some fear and concern, I made my way to his office. There kneeling on a leather-padded stool, elbows resting on his desk, this shaggy-bearded hunchback with a twinkle in his eyes greeted me with, "Hello, what's new?"

I explained my problem and shall never forget how quickly he sensed my predicament—and how quickly he came up with an answer. Without a moment's hesitation, he jotted down some figures on his pad, explaining them as he progressed.

"One kilowatt of electricity equals 1.34 hp. So 60,000 kw would be equivalent to about 80,000 hp. And one horsepower is equivalent to the muscle work of 22½ men." Steinmetz remarked, looking up from his pad to see if I understood his explanation. "To carry it further, 80,000 hp would equal the muscle power of 1,800,000 men. But a turbine requires no rest; it works 24 hours a day, three 8-hour shifts. And so we multiply the 1,800,000 by three, and what do we get? Energy equivalent to the muscle work of 5,400,000 men. That's greater than the combined muscle power of all the slaves in the United States before the Civil War."

His quick, clear explanation amazed me. In jotting down these figures he had used no reference books. I thought he might be wrong on the slave population angle; and after leaving his office, I checked the figure, only to find that the slave population in 1860 was 4,700,000. Armed with his simple explanation, I fulfilled my boss's request for a page-one story.

That interview began a friendly acquaintanceship that continued until his death, 34 years ago. I visited him frequently at his home and his summer camp as well as at his office. His ability as a mathematician was uncanny. He could multiply and divide huge numbers between cigar puffs as readily as I could add two and two. For instance, one day an associate, thinking he might trick Steinmetz, put this problem to him:

"If you bore a hole two inches in diameter through a circular solid rod two inches in diameter, how much material is removed?"

Again, without hesitation, jotting a few figures on his pad, he came up with the answer:

"Exactly 5.333 cubic inches."

The inquiring friend returned to his office and spent almost an hour figuring out that Steinmetz was correct.

Maker of lightning

He made numerous worthwhile contributions to the electrical field yet is undoubtedly best known to the public as the first man to create artificial lightning. His gnome-like appearance suggested the type of person one might imagine as the hurler of man-made thunderbolts.

* C. D. Wagoner, deceased, retired from General Electric in 1954. This article is a somewhat condensed version of one that appeared in *General Electric Review* in July 1957.

But why make lightning when nature supplies far more than the average person likes?

Before he perfected his Jovian machine so that electric equipment could be laboratory tested and made to withstand the destructive forces of natural lightning, homes were often thrown into darkness during a thunderstorm. No one knew how to make electric equipment withstand a lightning stroke. You just built it as well as you knew how, placed it in service, and hoped for the best during a thunderstorm.

Steinmetz closely duplicated nature in producing his artificial lightning. For thunderclouds, which store up the electric energy of the raindrops, he coated several glass plates with metal foil. These were arranged on wooden racks that acted as insulators. When voltage was applied, each plate stored up energy until it could retain no more. Then an instant discharge, a quick flash, and simultaneously the plates discharged their energy.

Steinmetz figured that the destructive force of such flashes represented one-million horsepower during the hundred-thousandth second of their duration. It was sufficient to shatter huge blocks of wood, heavy porcelain insulators, or whatever was desired in testing equipment against the natural forces of lightning. His lightning, though not equivalent to natural lightning, proved most useful. Later, General Electric built what has become one of the world's best-known lightning laboratories, where flashes extending more than 50 feet and carrying 15-million volts are produced.

His human side

Although best known as an electrical wizard, Steinmetz had a human side—equally as interesting though little known. He loved to play cards, calling his poker club the Society for the Equalization and Distribution of Wealth. In pinochle he remembered every card until the last one was played.

Though careful not to hurt people's feelings, he enjoyed outwitting them. He once talked the president and board of trustees of Union College in Schenectady into granting his fraternity permission to build a new home on the most coveted spot on the college campus. And later he helped them to have the only private tennis court on the college grounds. During 10 of his years at General Electric, Steinmetz also served as professor of electrical engineering at Union.

Pet problems

Of all his queer pets, his alligators caused the greatest commotion in Schenectady. He had seven. One day they escaped and took refuge in the Erie Canal, which then passed through the center of Schenectady. For hours people thronged canal banks watching the rescue operations. Eventually six were recovered. A passing canalboat crew probably picked up the seventh.

Steinmetz had other pet problems: His two crows brought home all sorts of shiny articles—often pieces of jewelry—that he was forever trying to return. He also had trouble with his owls. He didn't realize they were of the cannibalistic species until one morning he discovered all his baby owls missing and the plump mother owl sitting on a limb, feathers still protuding from her mouth. Later his pet raccoon evened the score by eating the mother owl!

His reputation as a keeper and friend of strange animals was far-reaching, as evidenced by an exchange of greetings when Marconi visited Schenectady a year before he died. His first question to Steinmetz was:

"Doctor, how's that Gila monster of yours?"

"Oh, he got too lazy to eat and died," Steinmetz replied. Then he told how he tried to force raw eggs and other food down its throat, but in vain.

Summer solitude

When he sought solitude, he would go to his crudely built summer camp, hidden from view along the Mohawk River, about eight miles from Schenectady. It had few conveniences other than electric lights operating from a large storage battery that was recharged at the laboratory about once a week. When at camp, his office secretary reported him out-of-town to visitors, unless she thought it important enough to disturb him. Today this camp occupies an honored spot in Henry Ford's Greenfield Village, Dearborn, Mich.

He spent much of his time while at camp in his floating office—a canoe with a couple of boards stretched across the gunwales as a desk, an old tin box as the receptacle for his pencils and slide rule, and a cushion to kneel on. But he was not always engaged in working on some technical problem. I remember my last visit to his camp, a few months before he died. I found him in his canoe, stripped down to a sweat shirt and shorts, with elbows resting on his improvised desk, reading a book.

"What are your reading today, Doctor?"

"It's called *The Lunatic at Large*."

"Who wrote it?"

"I don't know. The first eight pages have been torn out."

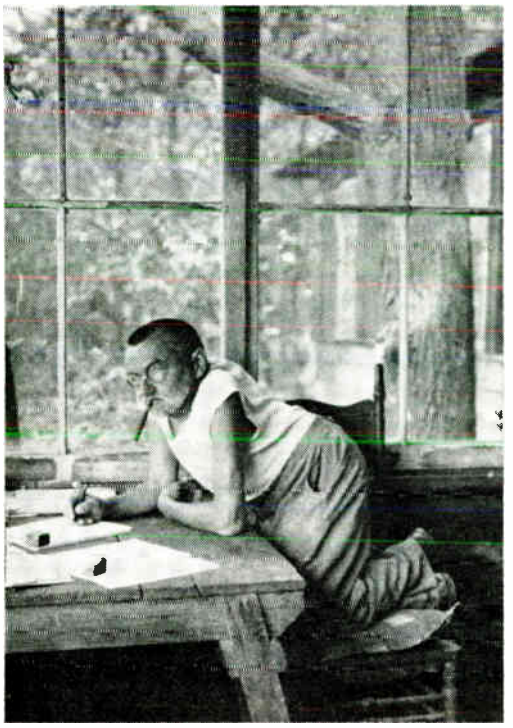
In this way he relaxed his busy mind. However, his camp wasn't just a place to escape work. Here he wrote his entire series of electrical textbooks, still considered among the best fundamental authorities in teaching modern electrical engineering at many colleges.

An event at his camp prompted his investigations of lightning. On a summer afternoon in 1920, lightning struck his camp, shattering timbers, fusing wires, burning out his storage battery, and breaking a large mirror in his bedroom into hundreds of small pieces. He carefully reassembled the broken glass, studied the path of the lightning on the silver coating, remarking "Just the evidence needed to begin some laboratory investigations." A year later he had built his artificial lightning generator at General Electric.

Practical jokes

Steinmetz enjoyed his pranks even at camp. One day he had a friend in his canoe. When a few feet from shore, he deliberately tipped it over, throwing both men into the water. In the commotion, Steinmetz disappeared from sight. Although not a swimmer, he could hold his breath a long time under water. On this occasion he sank to the bottom, crawled to shore under water, and hid in the underbrush until excitement reached a high pitch.

Another time, when I had made an appointment to bring a newspaper friend for an interview, I found him apparently swimming with a lighted cigar in his mouth. Actually, he was crouched in the shallow water, moving about on his hands and knees in hopes of creating comment and laughter.



Steinmetz in two of his favorite working positions at his campsite in upstate New York and Steinmetz with Edison.

Steinmetz enjoyed cooking and prepared most of the meals on a small alcohol stove. When some particular guest was present at camp, he would prepare a chicken or steak, serving vegetables from cans. Another favorite dish was *eier kuchen*, or egg pancakes. I never learned their ingredients, but they tasted good.

He never liked to wash dishes, generally allowing them to pile up until he had some friend at camp with whom he would bargain: "I'll cook the meal, if you'll wash the dishes." He would then slip in his accumulation of dirty dishes with those used for the meal.

Exploding some myths

Many stories are told about Steinmetz. Perhaps the best known is the "No smoking, no Steinmetz." To meet an insurance regulation, General Electric posted "No Smoking" signs. As the story goes, when Steinmetz saw the signs, he left the plant and was absent three or four days. Found at home, he explained between puffs on his cigar that if he couldn't smoke, he couldn't work. It's almost too good a story to refute. The truth is Steinmetz never paid any attention to the signs but kept on smoking; no one ever stopped him.

Another anecdote, also without foundation, claims that Steinmetz never drew a salary from General Electric. According to the story, he was given a Company check book, wrote whatever checks he wanted, and the Company paid them. Actually, Steinmetz received a definite salary, just as every other employee of General Electric. Being a socialist, money was never too great an object with Steinmetz so long as he could have all the necessary tools to work with, and these the Company readily supplied without cost.

Although owner of six different automobiles and co-organizer of the Steinmetz Electric Motor Car Corporation, Steinmetz drove a car only once in his life. When he died, he owned a Lincoln and a Detroit Electric but relied on his friends to do the driving. The Electric was arranged so that he could sit in the front seat, in what appeared to be the driver's seat, while a man in the rear seat operated the car.

Steinmetz seldom resorted to dictation in answering his

mail. Instead he wrote replies across the face of letters, using his own peculiar style of shorthand, a sort of German script and Pitman combined. Only his secretary and one or two others closely associated with him could transcribe it. I once asked him why he used this peculiar style. He had a reason and to me it seemed a good one. He went to his files, picked up some of his shorthand written at least 20 years previously, transcribed it as readily and easily as though reading from a daily newspaper. 'Show me the stenographer who can pick up 20-year-old notes and do the same," he said.

On one occasion our conversation drifted to remembering names of people.

"Do you have that trouble?" I asked.

"No. I don't have that trouble because I don't try to remember names," was his ready response. "I remember people as the one who said this or that or the man who did something worthwhile. That's all that counts."

Personal philosophy

On another occasion, when I was seeking information for a story on his advice to young people, he made this remarkable statement that seems as sound today as when he gave it more than 30 years ago:

"If a young man goes at his work as a means of making money only, I am not interested in him. However, I am interested if he seems to do his work for the work's sake, for the satisfaction he gets out of doing it.

"If I were able to bequeath one virtue to every young man, I would give him the spirit of divine discontent, for without it the world would stand still. The man hard to satisfy moves forward. The man satisfied with what he has done moves backward."

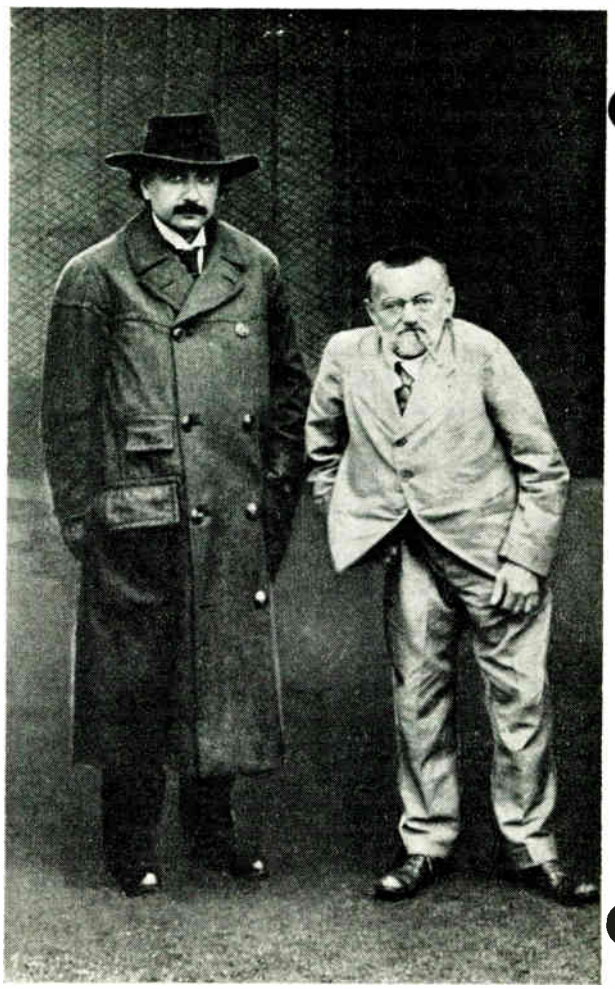
Although Steinmetz didn't belong to any church and had often been accused of being an atheist, this wasn't true. He may not have believed in the Supreme Being, since he once asked me, "What evidence do you have for one?" Yet he did believe some supernatural force governed the world. Roger Babson, well-known economist, asked him: "What line of research will see the greatest development during the next 50 years?" With a moment's careful thought he replied:

"I think the greatest discovery will be made along spiritual lines. Here is a force which history clearly teaches has been the greatest power in the development of man and history. Yet we have been merely playing with it and never seriously studying it as we have physical forces.

"Someday, people will learn that material things do not bring happiness and are of little use in making men and women creative and powerful. Then the scientists will turn their laboratories over to the study of God and prayer and the spiritual forces which, as yet, have hardly been scratched. When this day comes, the world will see more advancement in one generation than it has in the last four."

Steinmetz—born in 1865 in Breslau, Germany, of ordinary parents—was reared by his *grossmutter*, or grandmother. His mother had died soon after his birth. His crippled body prevented participation in games with other children, who gradually shunned him.

When he reached college age, he had largely overcome, through his brilliant mind, the feeling of others against him. He knew the answers when other students didn't, so naturally they turned to him for assistance. His socialistic tendencies at college forced him to flee to Switzerland in his senior year to escape imprisonment. He entered the Zurich Polytechnical School and while there met a fellow student who persuaded him to join him on his return trip to America.



Einstein and Steinmetz.

Arriving at Castle Garden with practically no knowledge of the English language, he possessed no visible means of support. A bad cold had caused his face to swell, adding to his deformed appearance. Immigration officials first decided to refuse him admission to this country. Only when his friend produced a roll of bills, declaring "half belongs to Steinmetz," and guaranteeing, "he will not become a public charge," did they change their minds.

First U.S. job

Steinmetz had heard and read much about Thomas Edison. He felt that if he could get a job in Edison's plant, he would be afforded the opportunities he sought. But he got no further than a short interview with the chief engineer. Downhearted but not discouraged, Steinmetz went to Yonkers the next day. He had a letter of introduction to Rudolph Eickemeyer, a manufacturer of hatmaking machinery and electric devices.

Eickemeyer, like Steinmetz, was a German. Speaking their native language, Eickemeyer quickly realized that here was a top-flight engineer and mathematician. He told Steinmetz to report for work as a draftsman the next morning for $12 a week. Before long Eickemeyer transferred him to electrical research work. Here, at last, he was in his element. He soon designed a new streetcar motor, a marked improvement on any then available.

While with Eickemeyer, Steinmetz assumed what he termed a good American name. He had been christened Karl August Rudolph Steinmetz. Learning that Charles was English for Karl, he started using it. Then one day an old German friend—a man who had attended school with him—dropped in to see him.

"Well, if it isn't Proteus!" exclaimed the former classmate.

It was the nickname conferred upon Steinmetz by one of the student societies. At the time he didn't like it. He knew the *Odyssey* from beginning to end; Proteus was the old man of the sea. If you caught him, he'd change in your hands to a hurricane, to a fire, or to a sea serpent. If you kept a firm hold on him, he would change back to his real shape, a wrinkled old deformed person, and tell you all the secrets of the world.

But now he felt differently. He had assumed no middle name and felt that all Americans required one. As names, he liked neither August nor Rudolph; both were too German. His friend's greeting gave him an idea. He picked up his pencil and wrote, "Charles Proteus Steinmetz."

"There, that's a good American name," he said—and used it ever after.

Joins General Electric

Steinmetz' work at the Eickemeyer factory attracted wide attention. One day he had a most distinguished visitor, Edwin Wilbur Rice, chief engineer and later president of General Electric. Rice offered him a job. But Steinmetz was reluctant to leave his good friend Eickemeyer. Then General Electric purchased a part of the Eickemeyer business that included Steinmetz.

Four weeks after Steinmetz had reported for work at General Electric, a friend noticed he didn't appear too well. He asked if anything was the matter. With some hesitation, Steinmetz replied that he hadn't been paid since he joined the Company, that his funds were low. As a result he had not eaten too regularly. When asked why he had not said something, he explained that he had hestiated for "fear that maybe General Electric thought the experience he was getting was enough—without pay." Nothing had been said about salary in his talk with Mr. Rice. The Company quickly rectified the mistake. A clerk had neglected to add his name to the payroll.

His achievements

Steinmetz was not a great inventor. His discoveries carried no popular understanding and appeal as did Edison's. Perhaps his greatest contribution to the electrical art during his 30 years with General Electric was his formula on the laws of alternating current, which made possible a much wider use of alternating current. The formula solved hundreds of problems that had puzzled engineers for years in the designing of transformers, motors, generators and in the distribution of electricity for far greater distances at higher voltages.

His artificial lightning machine made possible the development of dependable lightning arresters, protectors of transmission lines, and electric equipment that withstood the ravages of natural lightning. To General Electric he was probably most valuable as chief consulting engineer, the Supreme Court to whom all engineers with difficult problems came for advice. The Company had implicit confidence in his decisions.

Steinmetz' death at the age of 58 came suddenly and unexpectedly. Only a week earlier he had returned from a combination vacation and business trip to the Pacific Coast. He had enjoyed his first visit to Hollywood where he met his favorite movie actor, Douglas Fairbanks, Sr. He even drove an ostrich cart at the Los Angeles Ostrich Farm and asked that a photograph be taken for his scrapbook. He enjoyed retelling this unique experience.

The exertion of the trip apparently was too great. Returning to Schenectady, Steinmetz' doctor advised him to remain at home resting for two weeks. On the morning of October 26, 1923, he awoke at the usual hour, cheerful as ever, and ordered his breakfast. But when the tray was brought to his room a few minutes later, he had passed on, as quietly as an electric motor stops when it ceases to receive electric current.

Steinmetz examining fragments of tree branch damaged by lightning from his lightning generator.

Who was the Steinmetz I knew? Not a hard-bitten man of science, but a patient friend to mankind. Not a self-satisfied man of fame, but a friend to the underdog. Having pulled himself up to the height of giants, he paused to give mankind and science a lift.

Two men summed up this dual devotion to man and science...

"He always wanted to help everybody."—Alfred E. Smith.

"Mathematics to Steinmetz was muscular strength and long walks over the hills and the kiss of a girl in love and big evenings spent swilling beer with your friends."
—"Proteus," *USA*, by John Dos Passos.

Reminiscences of Dr. Steinmetz

Philip L. Alger

I first became aware of Dr. Steinmetz at the age of 17, when my father gave me a copy of Steinmetz' just-published book on *Engineering Mathematics*. In this book he developed the use of complex quantities, or "general numbers" as he called them. This was a new concept, which has proved to be invaluable in the solution of all kinds of problems, especially in the fields of alternating currents and electrical engineering. The book strongly increased my early interest in mathematics and led me to dream about some day going to work for the General Electric Company, where Dr. Steinmetz held forth.

When I actually came to Schenectady as a General Electric engineer in 1919, I met Dr. Steinmetz for the first time, and began to hear all kinds of stories about him. While I only spoke with him a few times, I read many of his books and shared in the universal admiration for his achievements. In the course of time I acquired at either first or second hand many legends and anecdotes about him, some of which may well serve as inspiration to young people even at the present time.

These stories are associated in my mind with a vivid memory of seeing him walk up Wendell Avenue on many occasions and sit down on a boulder in Corlaer Heights overlooking the broad vista of the Mohawk valley. In the year after his death, my wife and I bought the plot of land with this boulder on it and built our present home, 1758 Wendell Avenue, with the idea of enjoying the view that Dr. Steinmetz appreciated so much. Unfortunately, on an occasion when I was out of town, some men employed by the General Electric Company treasurer, Henry W. Darling, took the boulder and installed it upside down at the lowest level on the left-hand side of a gate in the wall he built around his property, at the corner of Wendell Avenue and Rosa Road. I still hope some day to recover the boulder from its present position and restore it to its original site on my front lawn.

Steinmetz as a teacher

Many years ago, one of the leading engineers of Stone and Webster told me that "at a meeting of the American Institute of Electrical Engineers when Dr. Steinmetz speaks, no one else has spoken." This was his way of saying that anything Dr. Steinmetz said was far more interesting and important than anything said by anyone else.

Dr. Steinmetz had a peculiar talent for engaging the interest and close attention of his audience and holding them spellbound for a good portion of his address. With his bent form, his small stature, his large head, and intense manner, he was always a striking figure, who became the center of attention as soon as he entered a room. When he began an address, speaking without notes, he would start off with some very simple statement, which every one present could immediately accept as being true, if not obvious. Having thus put his audience at ease and in motion with him, he would gradually develop the subject, making statements which, though clear, were more and more advanced. After a while, some of the audience would be unable to follow, but others would keep up; and toward the end of the period only a few would still be fully understanding. However, every one present had the feeling that he had understood for a certain distance and that others around him had carried on much further, so evidently it could be done. The listeners felt like hitchhikers to whom Dr. Steinmetz had given a ride far along a road that they wanted to follow. When they went away, they felt that they had also acquired a momentum that inspired them to undertake studies of their own, until they had mastered the ideas that Dr. Steinmetz had held before them.

Dr. Steinmetz had another vital quality of a teacher—he was most sympathetic and understanding with young people. It was his custom on Easter Sundays to attend services at the Unitarian Church, of which he was a member, and to take part afterward in the annual ceremony of giving flowers to the children.

Those who were present still remember with pleasure a lecture on soap bubbles that Dr. Steinmetz gave in the old Edison Club on Washington Avenue. He was pointing out that a soap bubble always assumes a shape that gives the least possible stress in the film, so that these shapes provide useful information to engineers for the design of structures. In the course of his talk, Dr. Steinmetz was interrupted by a question. He immediately seemed lost in thought, then turned to the blackboard and began to write out longer and longer equations, entirely forgetting his audience and his open-mouthed questioner. Suddenly, on hearing a titter from the audience, he turned and said with a smile, "I see I have been a 'leetle' too technical"—then resumed his popular talk.

"What a baby wants and what it thinks it wants, are two different things; and exactly the same can be said of most adults."

For many years it was Dr. Steinmetz' custom to hold open house for young men, particularly General Electric engineers, at his home on Sunday afternoons. Dr. R. E. Doherty, later President of Carnegie Tech, was one of those who learned a great deal from the Doctor on these occasions. Doherty came to General Electric in 1909, after graduating from the University of Illinois, and became a designing engineer in the old Alternating-Current Machinery Department. Here he soon found problems that he did not understand, and he sought out the Doctor for help. With the Doctor's encouragement, Doherty became a regular attendent at the Wendell Avenue Sunday afternoon meetings, and soon found that he could apply Dr. Steinmetz' ideas to his practical design problems.

In this way, and because of his own abilities, Doherty before long became recognized as the leading engineer in the A-C Department, and he was called upon to an-

swer all kinds of questions. Increasingly the problems that came to him were outside of his assigned work, so that he spent more and more time working in fields which, strictly speaking, were none of his business, even though they were of vital importance to the company as a whole. At last the day came when H. G. Reist, the head of his department, called Doherty in and said, "Bob, whom are you working for? You must be spending half of your time on things that do not concern my department." As a result of this conversation, soon afterward Doherty moved up to work directly with Dr. Steinmetz in the Consulting Engineering office; and when Dr. Steinmetz died in 1923, Doherty was immediately recognized as his successor. With the aid of Dr. A. R. Stevenson, Mr. Doherty created and launched on its long career the famous General Electric Company's Advanced Engineering Program, in which so many leading engineers of General Electric have developed their full powers.

Early days in Schenectady

When Steinmetz came to America, he spoke very little, if any, English, and had no money, but he found a position as a draftsman in Yonkers with his father's friend, Rudolph Eickemeyer, who had a small manufacturing business. In due time, Steinmetz went to various meetings and became widely known. As stories spread about his ability, these came to the ears of E. W. Rice, then Chief Engineer of General Electric. Mr. Rice went to see Steinmetz in Yonkers, but Steinmetz merely told him that since he worked for Mr. Eickemeyer, Mr. Rice should see him. When Mr. Rice asked Eickemeyer if he could hire Steinmetz, Mr. Eickemeyer said, "No, but you can buy my company." Rice did buy the Eickemeyer company, and that is the way that General Electric acquired Dr. Steinmetz.

The rough and ready ways of the electrical industry in those days are illustrated by a story about Elihu Thomson, the famous inventor, then the leading figure in the electrical world. On one occasion he went to call on a customer, with the hope of selling him some electric equipment. To gain favor for his apparatus, Mr. Thomson explained how the electrons went in here and came out there in the circuit. As soon as Thomson had left, the customer called in his competitor, J. J. Wood, and asked him, "Do you have electrons in your equipment?" Mr. Wood replied, without hesitation, "Of course not. Them damn things?—we wouldn't have them in our equipment." He got the order. Incidentally, not long afterward the General Electric Company secured the services of Mr. Wood by buying his company, which became the General Electric plant in Fort Wayne, Ind.

In Schenectady Dr. Steinmetz wrote many articles and books, spoke at many meetings, became President of the American Institute of Electrical Engineers, and also became a professor at Union College. In these ways his fame spread far and wide. This brought many able young men to Schenectady to work for the General Electric Company and to learn from Dr. Steinmetz, just as moths are attracted to a flame.

Two of these young men were Dr. Ernst Berg and his brother, Eskil, from Sweden. These two, with W. L. R. (Bill) Emmet, and Dr. Steinmetz were great friends, and they shared many activities. On one famous occasion, they decided to hold a swimming race in the Mohawk River. The conditions were that Dr. Steinmetz should paddle his canoe and act as judge, while Emmet and the two Berg brothers should swim for a specified distance, fully dressed in evening clothes—complete with tails, dress shirt, tie, and top hat. The winner, who finished the race in full costume, was Eskil Berg.

Another great engineer who came to work with Steinmetz was Dr. E. F. W. Alexanderson, later famous for his pioneering in transatlantic radio and his many inventions. Soon after Alexanderson arrived, Dr. Steinmetz decided to find out if he was worthy to enter

"When a man thinks only of the dollars he is getting, he is not apt to get very many."

his circle of associates. He invited Alexanderson to join him in a canoe trip along the Mohawk. When they had arrived at a point some distance from the Steinmetz camp, the Doctor suggested that Alexanderson get out and walk back to the camp along the shore, explaining that he would enjoy the scenery and remarking also that he would not need his shoes. Alexanderson obediently got out and walked, but he found that the way was stony and very hard indeed on his bare feet. However, on arriving at camp he made no mention of his sufferings, and the Doctor accepted him without further question.

Another anecdote, concerning Dr. Berg, shows how well Dr. Steinmetz understood the characteristics of the men who worked with him. It appears that a transformer failed somewhere out West, and Steinmetz asked Berg to investigate. He gave him careful instructions regarding what to look for, and said that on no account was he to put the transformer on full load. In due time, Steinmetz received a telegram from Berg, asking, "Have I your permission to put the transformer on full load?" The Doctor wired back, "I see you have already done so—let me know the results."

The Doctor's versatility

Many distinguished visitors with all sorts of interests came to see Dr. Steinmetz. It was his usual custom on

"Watch the man whose only motive for being good is the Heaven he hopes to get in payment. He may be good, but he won't be good for much."

these occasions to ask the visitor what his field was, or in what he was particularly interested. Then, whatever the subject—whether history, or Latin, or nature, or politics, or engineering—the Doctor would talk about it, and generally the Doctor proved quite as well informed as

his visitor. This characteristic of Steinmetz is illustrated by the story of an unexpected visit he made to Buffalo on one occasion. On arriving he called upon an old friend, who was surprised to see him and said he would like the Doctor to come to dinner with him but that he had to attend a meeting afterward. As an afterthought, he said," Perhaps you would like to come too." The Doctor said he would be glad to, and asked what the meeting was to be about. His host said, "It is a Masonic meeting." After dinner, Dr. Steinmetz went with his host to the meeting and was introduced. For courtesy's sake, the Doctor was asked if he would care to say a few words. He replied in the affirmative and launched forth on a long talk about the Masonic order. To the astonishment of his hearers, whom he held spellbound, he told them far more about Masonry than they knew themselves.

After the meeting, his friend asked the Doctor, "How in the world did you know so much about Masonry?" "Oh," he replied, "you have a good library in Buffalo."

As this illustrates, Steinmetz had an excellent memory as well as wide interests, so that he was a student of and a contributor to many different fields of thought. One of his famous sayings is that in the process of learning, at first, the more things you learn the more difficult it is to remember them. However, in due time, as the process of learning continues, it gradually becomes easier; and finally, the more you learn the easier it becomes, for the reason that each new fact fits in with and serves to recall other facts that are related to it!

The White Revolution* *Charles P. Steinmetz*

Heat is energy; light is energy; motion is energy, or power, as we often call it. That is, there is a thing called "energy," or "power," which in the form of motion turns the wheels of industry and propels the railroad train or the steamship, the trolley-car or the aeroplane; in the form of light converts night into day; in the form of heat cooks our food and makes our houses habitable in winter; in the form of chemical energy makes steel out of iron ore and converts clay into aluminum. Without an ample supply of energy or power our civilization would quickly come to a standstill.

We cannot create energy, so we have to take it where we find it in nature. Practically, all of our energy supply comes from the sun, in the form of light. Aeons ago the sunlight shining on the primeval forests of tree ferns was stored up as chemical energy in coal, and is now, when we take the coal out of the mine and burn it to produce heat, recovered by us as light or power.

It is less than a hundred years since coal has been used extensively as fuel, and now the annual consumption has become almost inconceivably large. The amount of coal mined in one year in the United States alone, if used as building material, would build a wall all around the United States larger than the famous Chinese wall by which in bygone centuries China attempted to protect its northern frontier. The power contained in this coal would be sufficient to lift the wall two hundred miles high. The amount of coal existing in the earth is so vast that this huge annual consumption makes a barely appreciable inroad upon it. But even if our annual coal supply should last for hundreds of years, some day it will be exhausted. What then? Will civilization come to a standstill and man lapse back into barbarism? Or will other sources of power be found? There is a second great source of energy available, which has been opened up by the electrical engineer during the last generation: the water powers of the country. It is to these we shall have to look for help when our coal supply begins to fail.

They are inexhaustible so far as in using a water power we do not consume it, because the sun continuously replenishes it. The heat of the sunlight evaporates the moisture, and the vapor rising in the atmosphere condenses as clouds and comes down as rain on the hills, and, gathering in the river, turns the machinery of the water power station and gives us back the energy of the sunlight as electric power.

But how much energy is available in the water powers of the country? Is it enough, when coal is exhausted, to supply the power demand of our civilization? The total available water powers of the United States have been variously estimated as from fifty to one hundred million horsepower. Such estimate is uncertain and open to the possibility that, when need compels, future advance of engineering may enable us to utilize water powers which cannot now be used. We can however estimate a maximum possible value, beyond which we could never hope to reach. We know the rainfall and the elevation of the different parts of the United States, and from rainfall and elevation we can estimate the foot-pounds of energy represented by the total rainfall of the country. Allowing then for the amount of water required by agriculture, and a minimum loss by seepage and evaporation, we get the maximum possible amount of water power which could be produced if every raindrop which falls anywhere in the United States were collected and all the power which it could give on its travel down to the ocean developed. This amounts to about three hundred million horsepower.

Thus the maximum power which could ever be developed in our country from water, if every drop of rainfall were utilized, must be less than three hundred million horsepower. That is a vast amount of power. But it is just about the total which we get out of our present coal consumption for all purposes—power, light, heat, metallurgical work, and so on. Thus we see that the total water power of the country, even if all of it could be developed—which is not possible—would be only just enough to replace our present coal consumption, leaving nothing for future increased needs of power. This is

* Reprinted from *Survey*, March 25, 1922, pp. 1035–1037.

rather unexpected because people have always hoped that in the future when the coal will have been used up, the water powers of the country will take the place of the coal. But we know already that all the water powers of the country would not be enough.

Unless, then, new and as yet unavailable sources of energy be made accessible, the hope of the future must lie in saving power and in increasing the economy of energy consumption, in order to get along with the limited supply. This is to a very great extent possible, as our present use of energy is extremely wasteful; three-quarters or more of the energy of coal is really wasted, and at present our engineering knowledge is not sufficient to avoid all of this waste.

Consider the largest single user of coal, the railroads. If the railroads replaced the steam locomotive by the far more efficient electric traction, from one-half to two-thirds of the coal would be saved, smoke and dirt would vanish, and without any increase of track a material increase of traffic could be handled, due to the quicker starting, better control and greater power of the electric locomotive.

The huge steam turbine turns out power with half of the coal consumption of the average smaller steam engine. If we replace all the small power sources throughout the country with electric motors receiving their power from huge steam turbine stations, another vast saving would result. This is being done increasingly. Often we could save all the coal, by deriving the power from a hydro-electric station.

In heating houses we really use more than ten times as much coal as necessary. We could save nine-tenths of it, but we probably will not do so until forced to it by a failing coal supply, for to save would require a radical change of building construction. There is, for instance, a great waste of heat in the gases going through the chimney. There is a waste of heat through walls and doors and windows. But the greatest waste is due to our present inefficient methods of ventilation. We must have sufficient ventilation. That is, we must let the foul air out and take in fresh air. In most private houses we do this in a haphazard fashion by periodically opening some door or window. In large buildings we have properly arranged ventilating systems. In either case however we throw out the warm foul air and take in cold fresh air, and so over and over again we have to heat new masses of fresh air and all this heat is thrown away in the foul air which we exhaust. Usually over 90% of all the heat from our furnaces is lost. Although we must replace the foul air with fresh air, there is no reason why we should throw away with the foul air all the good and valuable heat which it contains, and not supply from it new heat to the fresh air. If it were properly arranged, we should take the heat out of the foul air before we exhaust it, turn it into the incoming fresh air, and so heat the incoming cold air by the heat of the outgoing warm air. This recovery of the heat is called a "regenerative system of heating." We could do this by passing the outgoing warm air around the outside of the pipes which bring in the fresh air, and so warm the latter by the furnace. Then all the heat which our furnaces would have to supply would be that lost through the walls, doors and windows. By proper building construction this loss could be made slight, so that a very small amount of heat would warm the house—so small indeed that it would be economical to heat houses electrically, economical to own a "house without a chimney," even though the price of electric energy must always remain many times greater than the price of coal containing the same amount of energy.

The solution here as well as elsewhere is, "do it electrically." Electricity is a form of energy which can be produced economically from hydraulic energy, as well as from steam energy; can be sent over long distances, that is, transmitted with high efficiency; can be distributed and supplied and put to any use with very small losses, in a simple manner by means of apparatus which are reliable and require little or no attention. This is the reason we are now entering the electrical age.

The problem which the engineers have to solve to protect our civilization from a collapse by a failure of the energy supply is therefore a vastly greater one than most people realize. It is not merely to develop the water powers of the country, the "white coal," and to transmit and distribute this power electrically. All the water powers would not, as we have seen, generate enough energy even for our present needs. It is the vastly greater problem of organizing all the country's energy supply, from coal mine and waterfall, from oil well and all other sources. It is the problem of transmitting and distributing and supplying the energy to the places where it is needed, and of converting it into the form in which it is used, and all this in the most economical manner. Whether as light or as heat, as a small desk fan motor giving a cooling breeze on hot summer days, or a huge ten thousand horsepower motor in the steel mill crushing ingots into structural steel rails as if they were soft putty—all the way through, the guiding principle must be economy, the saving of energy to make the best of the limited supply. And this is being done electrically by converting nature's energy supply, whether coal or water power, into electricity, and then by transmitting and distributing it as electric power. If, when traveling through the country, you see electric transmission lines pass by you in increasing numbers, you then realize that they are doing what the coal trains have done and are still doing, that is, carrying the country's energy supply, without which our civilization would cease, from source to user.

But when, in the distant future, with the exhaustion of the coal supply and an increasing energy demand of an increasing civilization and increasing population, our energy supply will have become insufficient in spite of all economics which the engineer can devise, civilization will not fail; we may trust that the engineer will find new energy supplies. There is a source of energy a thousand times larger than coal and water power together—the energy of the sunlight. As yet we know no way of collecting and concentrating it into usable form, and the mechanical methods thus far tried (by reflection, or by collecting the heat under glass) appear hopeless. But methods at present unthought of may be devised. Perhaps some biological engineers, some future Burbank, may by progressive selection through centuries develop a plant life which, fed by the carbon dioxide gases of combustion, will grow with great rapidity and in growing will collect in its structure the energy of sunlight in the form of chemical energy. When burned, this may make available the energy as heat energy, and return the matter as carbon dioxide gas to raise further "energy crops," thus in a closed cycle perpetually supplying unlimited energy from the sun's rays.

On the nature of the electron—Part II

The story of the electron, introduced in the last issue, is brought up to date. In the realm of nuclear physics, for example, we see how the classical principle of parity conservation is being violated

J. L. Salpeter *Adelaide, South Australia*

The first part of this article was concluded with the observation that the theory of relativity demands a pointlike electron, as a result of which the electrostatic energy of the electron e^2/r appears to become infinite. We have said that physicists somehow learned how to live with these infinities, and in this second part we shall try to explain how they contrived to do so.

Quantum electrodynamics

Electrostatic energy becomes infinite because the Coulomb's force becomes infinite when r tends to zero. In "quantum electrodynamics," which is the branch of physics dealing with the motions of the electron in an electromagnetic field, the concept of force has been all but abolished and replaced by the interaction of electrons and photons. In wave mechanics, the electrons appear less real than objects of our everyday life such as a chair or a desk, because we cannot measure simultaneously the position and the velocity of the electron. In quantum electrodynamics, on the other hand, mutual repulsion of two electrons is pictured as a result of mutual bombardment of the electrons with photons, like two tennis players bombarding each other with tennis balls. Instead of computing the acceleration of the electron as a result of the action of a force, we apply the principles of conservation of energy and momentum to the collision between an electron and a photon.

Originally Coulomb's force meant "action at a distance," which is somewhat difficult for our minds to accept. This difficulty has been mitigated to some degree by Michael Faraday, who created the concept of "lines of force," which he imagined as something very real. The lines of force originate and terminate at the electric charges and are responsible for their mutual repulsion and attraction. Later this concept was elaborated by Maxwell, who calculated the tension along the lines of force and the compression across them necessary to account for the ponderomotive effect. This system of tensions and stresses would be difficult to imagine in empty space if it were endowed with solely geometrical properties. That was one of the reasons for the concept of the "ether," which was said to permeate the entire universal space. However, with the advent of the theory of relativity, the ether concept became impossible to accept—and the difficulties of tensions and stresses in the empty, geometrical space reappeared.

Quantum electrodynamics removed these difficulties completely by replacing the "lines of force" with the mutual impact of electrons and photons. However, the reason for the general acceptance of quantum electrodynamics today is not so much the fact that the impact of particles is easier to visualize than action at a distance, but that quantum electrodynamics agrees better with some measurements than the conventional theory does. This does not mean that Maxwell's electromagnetic field theory is obsolete; it is valid for instance when we want to calculate the electron orbits inside an electron-accelerating machine. Yet, if we try to calculate electron orbits and energy levels in spaces of atomic size, the results do not agree with our observation if we apply conventional theory. Although the discrepancies are very small, the accuracy of spectroscopic and other measurements is of such a high degree that they are nevertheless very real, and it was a great triumph of quantum electrodynamics to have explained such differences by reference to the analogy of tennis players bombarding each other with tennis balls.

In 1900 Planck introduced the quantum theory into 20th century physics, postulating that energy can neither be emitted nor absorbed in amounts smaller than one quantum ($= h\nu$). However, the process of emission and absorption was imagined as a continuous one. The question then arose as to what happens if the emitter or ab-

This second part of a two-part article begun in March will also appear in the May issue of the IEEE STUDENT JOURNAL.

sorber is destroyed before a whole quantum has been emitted or absorbed. This awkward question was answered by Einstein in 1905, who revived the Newton corpuscular theory of light for his quantum theoretical interpretation of the photoelectric effect. While light propagates in space as an electromagnetic wave completely in accordance with Maxwell's theory, emission and absorption take place in the form of particles which Einstein called "photons." Thus the dual wave–particle aspect of photons has been established and generally accepted, with the reservation that subsequent research will somehow elucidate and resolve the strange contradiction of a wave–particle. Common sense would say that something is either a wave or a particle, but not both at the same time—and that much has been conceded. At one time it is a wave and at another a particle, although it can be doubted whether this "it" is the "same" in both cases. In any event, the mystery defied any effort at reconciliation with "common sense" physics and when, 20 years later, de Broglie originated the dual aspect even of the electron, the mystery deepened instead of being resolved. Today, after another 40 years and with the coming of age of another generation of physicists, the dual aspect is accepted as one of the facts of life.

However mysterious the photon may have appeared at the beginning of this century, one aspect of it was very real—the "recoil" of the atom that has emitted a photon. The "rest mass" of the photon is zero, but due to its energy $h\nu$ it has a mass m and consequently a momentum $p = mc$, which can be computed by means of the well-known Einstein equation: $E = mc^2$. It is $p = h\nu/c$ and the recoil of the atom is the same as that of a gun that has discharged a bullet of the momentum p. It will be remembered that Mössbauer, a few years ago, discovered a recoilless emission (for which he was awarded the Nobel prize). However, it was not really recoilless; the recoil is taken up by the lattice of the crystal and thus made negligible. Ordinarily, recoil is suffered by only the atom that has emitted the photon and is not shared with the other atoms. This recoil accompanying the process of emission and the analogous impact accompanying the process of absorption are the basis for quantum electrodynamics, in which the concepts of force and acceleration have been replaced by recoil, impact, and conservation of momentum.

It will be useful to consider the process of recoil in a little more detail. Let us recall Bohr's theory of the atom, according to which only certain electron orbits around the nucleus are permitted (without radiation). If an electron falls from a higher orbit to a lower one it loses some amount of energy (one quantum), which is being radiated in the form of a photon. Since it is the electron that loses energy (emits a photon), the recoil is suffered primarily by the electron itself. It is only because the electron is tightly coupled to the atom that we can speak of the recoil suffered by the atom.

The recoil (and impact) will be more conspicuous if the electron is not coupled to an atom; for instance, in the case of conduction electrons in a metal or of free electrons in space or in a gas. In an antenna the conduction electrons are being accelerated by the electromagnetic field within the wire and the accelerated electrons emit photons at the expense of the energy of the field. However, at this point we do not wish to contemplate acceleration due to a force, but rather to consider the charges that are responsible for the generation of the field. The simplest case concerns two electrons in space—and nothing else. Each electron moves in the field generated by the other electron and, classically viewed, is being accelerated by this field. In the new theory we ignore continuous acceleration and rely instead on bombardment with photons. If an electron emits a photon, at whose expense does it happen? If we ignore the field and the energy stored in it, it could only be the other electron that supplies the energy. But the other electron is as poor as the first one and waits its turn to be bombarded—and thus under these conditions mutual bombardment will never start. The way out of this impasse is supplied by Heisenberg's uncertainty principle.

It will be remembered that Heisenberg's principle maintains that we cannot know exactly at the same time both the position and momentum of an electron. If we denote by Δp and Δx the uncertainties in momentum and position, respectively, the product $\Delta p \Delta x$ can be at best equal to a certain constant. This principle is not just a wave mechanical curiosity but has a positive value, inasmuch as it makes possible important predictions in certain circumstances. For instance, if we compress matter to an extreme degree, the radii of the orbiting electrons become smaller and smaller and accordingly Δx for each electron becomes smaller also. As a result the Δp's become larger, i.e., the electron velocities increase until their energies exceed the ionization potential and the electrons become detached from the nuclei. This argument has found useful applications in astrophysics. In terrestrial laboratories it has been found that under high compression insulators (for instance, diamond) become metallic for similar reasons.

The uncertainty principle can be formulated in another way; i.e., the product $\Delta E \Delta t$ can at best be equal to a certain constant, where E and t stand for energy E at a certain time t. That means that if we specify the time exactly ($\Delta t = 0$), then we cannot know the energy at all. This has an important bearing on the conservation of energy of an electron. It is useless to demand the conservation of energy of an electron to an accuracy higher than we can measure. This fact will help us with the initiation of the mutual bombardment of two electrons with photons.

An electron may emit a photon of the energy ΔE, if this energy is returned to it not later than after a time $\Delta t = \text{constant}/\Delta E$. Such a process is referred to as a "virtual" process because we cannot be sure that these actions, although permitted, do take place. How is the elec-

tron to know when the radiated energy will be returned to it? Still, comparison with experience tells us that the virtual processes have an important bearing on the properties of the electron. One consequence is that the electron is always surrounded by a cloud of photons. The electron itself, without the cloud, is referred to as the "bare electron"; this however does not occur in nature, as the real electron is always surrounded by photons.

What about the total energy? Since there is no bottom limit for Δt, there is no upper limit for ΔE. Let us imagine that the electron is a customer of a bank, which grants each customer an overdraft ΔE provided he returns the money within the time period $\Delta t (= \text{constant}/\Delta E)$. No sooner is the money returned than the electron draws another overdraft, and so it is no wonder that the account is always in the red. What is more, the amount tends toward infinity. Luckily the electron has a fixed deposit, the electrostatic self-energy e^2/r, and when this is taken properly into account by means of the so-called "renormalization" process, the balance is slightly in the black.

What it all amounts to is that we have subtracted infinity from infinity and obtained a finite difference, a procedure abhorred by the orthodox mathematician. However, our result agrees with the measurements whereas that obtained by the older, conventional theory does not. For instance, the magnetic moment of the electron is one Bohr magneton for the bare electron, while for the photon-cloud-surrounded electron it is 1.00116 magnetons; this is in excellent agreement with experimental results. Another example is the Lamb shift in the hydrogen atom. For the bare electron the two levels $2s_{1/2}$ and $2p_{1/2}$ have exactly the same energy, while for the photon-cloud electron there is a shift of 1052 Mc/s (experimentally 1057 Mc/s), again in excellent agreement with experiment.

If experiment confirms the "virtual" processes of photon emission and absorption, we can either say that the electron knows in what time it will be able to repay the loan, or—less anthropomorphically—that the effect precedes the cause, which would mean that the causality ceases if we consider extremely short time periods.

A way out of this difficulty has been suggested by Heisenberg, the very author of the uncertainty principle that lies at the bottom of those infinities. His argument is as follows: A complete system of physical units contains three independent dimensions, for instance, M, L, T (mass, length, time). Similarly, a complete system of universal constants should consist of three universal constants. For the time being we have two universal constants: c, the velocity of light, and h, Planck's quantum of action. Missing is a constant representing the shortest possible time (or length, since time and length are connected by the velocity of light). If we could postulate that no shorter time than, say, T_0 or length shorter than r_0 is accessible to our measurements (or even imagination), all our difficulties would disappear; neither e^2/r nor constant/Δt would tend toward infinity, and if cause and effect both happened within the time T_0 we could not say which occurred first, because T_0 is indivisible. However, for the time being there is no experimental evidence for the existence of a fundamental T_0 or r_0 and so we must keep our minds open for future discoveries.

Violation of parity conservation

The source of electrons employed in electronic engineering is almost exclusively thermionic emission. A metal wire (for instance, tungsten), bare or coated with barium oxide, emits electrons if heated in a vacuum. The electrons emitted were originally atomic electrons orbiting around the nucleus.

Another source of free electrons is the nucleus itself. If the nucleus happens to be unstable, it will decay and in the process may emit fast electrons called beta particles. The first radioactive element was radium, discovered at the end of the 19th century by Madame Curie. With the discovery of uranium fission came a new branch of physics—nuclear physics—whose object is the study of nuclear reactions. In the course of such studies a great number of radioactive isotopes have been discovered, all of which emit alpha, beta, or gamma rays, and atomic reactors supply plenty of radioisotopes. Some of these radioisotopes are already being used in electronic engineering—for instance, in X-ray tubes of very small size and in methods of monitoring metal castings, or even in uniform packing of cigarettes. For this reason the electronic engineer may wish to become acquainted with the alternative source of free electrons, the beta particles.

In 1957 it was announced that in beta decay the principle of the conservation of parity is being violated. So exciting was this news to physicists that many high-energy physicists stopped whatever work they were doing to investigate parity in beta decay. The two young physicists responsible for the discovery were awarded the Nobel prize immediately, whereas usually discoverers must wait for years for such recognition. The rest of this article will be devoted to an explanation of why this discovery was so exciting, and of what parity means.

To understand the meaning of parity is not as difficult as to understand why the violation of its conservation was so disturbing to physicists. Briefly, parity refers to nondiscrimination between "right" and "left," and conservation of parity refers to nuclear reactions in which nature treats "right" and "left" with equal favor.

Let us draw a right hand and a left hand on paper and cut them out. Let us now try to make the two hands coincide by shifting them in the plane. We will never succeed until we turn one of the hands over. Yet, in turning it over we step out of the two-dimensional world. In three dimensions we can never succeed in fitting a right-hand glove on a left hand because we cannot step out of the three-dimensional world. If we look into a mirror, we observe that our right-hand glove would fit the left hand of our mirror image. We would be inclined to refer to the mirror image of our left hand as the right hand of the person in the mirror, but then this person would have his heart on his right side. We observe that it is impossible to define right and left mathematically; we can define them only by enclosing a sample. On earth the matter is made easier because every human being carries samples of right and left in his own body, but let us try to communicate to intelligent beings on other planets the message that we have our heart on the left side. This will be impossible unless we find something that by nature can occur only in one variety, right or left. It is true that there is no subspecies of man with the heart on the right side, but we are convinced that this is so because of the common origin of man and heredity.

In the realm of living things there are many examples of helicity, but nature does not seem to prefer either of the two possibilities (except for heredity). The original and its mirror image are absolutely equivalent in every respect. In inorganic materials, wherever helicity occurs it occurs in both varieties; in biological substances the occurrence is somewhat limited by evolution and heredity. We can now understand how disturbing it would be if we discovered a phenomenon in which nature prefers fundamentally—not just by chance or heredity—right or left, and this is exactly what happened with parity in beta decay.

Before we proceed, let us discuss a familiar phenomenon which seemingly violates parity of right and left. Let us consider a long, straight wire, through which we can let a current flow; see Fig. 1. Underneath the wire we place a magnetic needle in the direction of the wire with an N pole and an S pole. Before the current flows, the situation is perfectly symmetrical with regard to right and left. However, as soon as a current flows through the wire the needle will be deflected, with the N pole to the left. Let us now take a mirror image of the arrangement before and after the deflection and let the plane of the mirror be underneath the needle and perpendicular to the plane of the paper. Before the deflection the mirror image of the current will flow in the same direction as in the original and so will show the polarity of the needle. After the deflection the direction of the current will still be the same, but the N pole will be deflected to the left in the original and to the right in the mirror image. The image does not tell the truth. If we consider the situation before the deflection as the "cause" and that after the deflection as the "effect," the effect in the mirror image is different from the effect in the original; it is something that cannot, and does not, happen. By definition parity is "conserved" if cause and effect in the mirror image are in the same mutual relation as in the original. In classical physics all phenomena obey the principle of conservation of parity without exception. The instance of the wire and the magnetic needle would be an exception if the needle truly consisted of a north pole and a south pole; in other words, if "true magnetism" did exist. Actually, what we refer to as magnetism is generated by "moving electricity," either by spinning or orbiting electrons.

If we replace the magnet bar by a number of current loops and take a mirror image of the arrangement, the polarity of the equivalent magnet is reversed; see Fig. 2. It follows that the arrangement with current loops conserves parity, while that with a magnet bar does not. This was known for a long time and physicists were inclined to think that it is not just coincidence that a true magnet bar violates parity conservation—"true" magnetism does not exist.

Mirroring is mathematically expressed by reversing the sign of the space coordinates x, y, z. If we replace $+x$, $+y$, $+z$, by $-x$, $-y$, $-z$, we obtain the mirror image of the situation in $+x$, $+y$, $+z$. We can also simply replace $+z$ by $-z$ and leave x, y unchanged; we then have the x, y plane as the mirror plane.

The sequence of the axes x, y, z defines a helicity ($x \rightarrow y$ is a rotation, $+z$ a translation). If x, y, z defines a right-hand helicity, then x, y, $-z$ would define a left-handed one. If true magnetism did exist, transition from an x, y, z coordinate system to an x, y, $-z$ system would convert a north pole into a south pole, which means that such a transition would not be permissible. We would have to decide once and for all whether we would use a right-handed or a left-handed coordinate system, and Maxwell's equations would be valid only with a unique coordinate system. It was felt, however, that the laws of nature could not depend on whether we use a right-handed or a left-handed coordinate system. Since there were no experimental indications of the existence of true magnetism, everything was considered in good order and the conservation of parity—at least in classical physics—was taken for granted and considered of universal validity, just as the conservation of energy and mass. It should be kept in mind that all the conservation laws are to be regarded as a result of experience and not as a priori truths. Physicists would not hesitate to discard any of these laws found by experiment to be invalid.

It is interesting to note that after the discovery of the violation of parity in beta decay, physicists had second thoughts about "true" magnetism. Dirac's relativistic wave mechanics, the same mechanics that predicted the existence of the positive electron (positron), also predicted the possibility of magnetic monopoles (isolated single north poles and south poles). This prediction was not

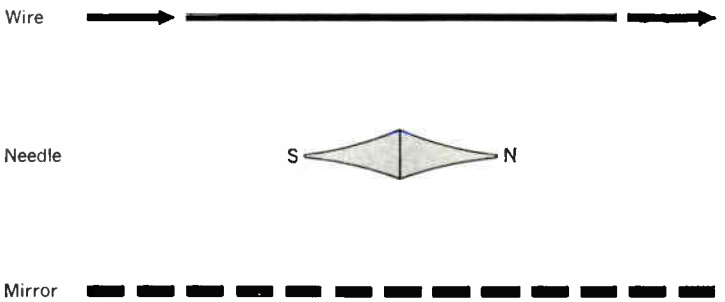
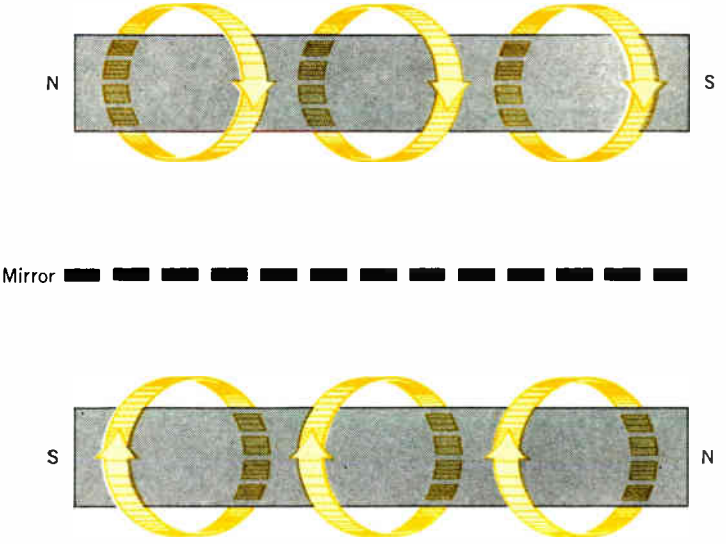


Fig. 1. Arrangement of wire and magnetic needle which demonstrates an apparent violation of the parity of right and left.

Fig. 2. Setup in which the magnet is replaced by current loops, demonstrating the conservation of parity.



acted upon until recently when high-energy physicists started a search for magnetic monopoles. The result has been negative so far and it is not very likely that it ever will be positive, but it shows how physicists refuse to be restricted by preconceived ideas.

The idea that nature does not discriminate between right and left is difficult to reject. If such a discrimination were established, how does it come about? Is perhaps the very space, empty space, partial to right or left? For this reason the principle of conservation of parity was taken for granted until about 1956, and in classical physics the principle is still valid today—but nuclear physics is not a continuation of classical physics. Consider a chemical reaction, which belongs in the main to the domain of classical physics. We know what atoms were there before the reaction, and if we inspect the matter after the reaction we will find the same atoms, but in a different arrangement.

In a nuclear reaction the units after the reaction are not all the same as before the reaction. Let us consider the case of the neutron. The nuclei of the elements all consist of protons and neutrons (a proton is the nucleus of ordinary hydrogen, while a neutron has approximately the same mass as the proton but without charge). It is possible to obtain free neutrons, liberated from nuclei, but the neutrons are not stable; they have a lifetime of about 1000 seconds, after which they decay into a proton and an electron. It would be incorrect to say that a neutron consists of a proton and an electron—the "diameter" of the electron is many times larger than the diameter of the neutron. It is better to say that the neutron is annihilated and, at the same time, a proton and an electron are generated. We cannot observe what happens between the annihilation and generation, and therefore we can only hope, but not be quite sure, that the various quantities such as mass, energy, charge, and momentum are conserved in the reaction. (Until 1956, it was also hoped that parity was conserved.) Here we should mention that there are two kinds of nuclear reactions—"weak" interactions like the beta decay of a neutron, and "strong" interactions like the fission of a uranium nucleus, in which energies of a much higher order of magnitude are involved. In strong interactions parity is definitely conserved. In this article we are concerned with weak interactions only.

The cobalt experiment

In the years 1954–56 a particular meson—an unstable nuclear particle—was investigated. Two variants of it seemed to exist that were identical in every respect but parity. This was quite unusual. A simple explanation would have been possible if it had not been necessary to assume conservation of parity in weak interactions; namely, that the two variants were actually only one kind. It occurred to Yang and Lee[1] that although everyone believed that parity was being conserved, no experimental evidence of right–left symmetry in weak interactions existed. They suggested the following experiment, which utilizes the radioactive cobalt isotope Co-60 (the same isotope that now is used in hospitals instead of X rays). The nucleus of Co-60 consists of 27 protons and 33 neutrons (atomic mass = 27 + 33). This isotope has a lifetime of 5.3 years and disintegrates into Ni-60 (28 protons and 32 neutrons) and an electron. The nucleus of Co-60 has a spin of 5, and it is known, from quantum theory, that the axis of this spin serves as a reference direction for any directional event in the nucleus.

(It may not be necessary but, to be on the safe side, we should point out that a spinning object does not, in itself, violate the right–left symmetry. A spinning wheel rotates clockwise or counterclockwise, depending on the direction from which we view it, but by turning the wheel by 180° we can reverse the sense of rotation. We can never make a right-hand glove fit the left hand, but we can equalize the sense of rotation of two wheels, one of which rotates clockwise and the other counterclockwise. A right–left choice will have to be made only if, besides spinning, we let the wheel progress linearly in the direction of its axis. It is correct to draw a circle with an arrow on its periphery to indicate the sense of rotation of the nucleus Co-60. But to draw the axis of the spin and provide this axis with an arrow would be our own choice, not the choice of the nucleus—and this is quite an arbitrary choice.)

The nucleus Co-60 decays into Ni-60 and an electron. We wish to observe how the directions in which the electrons are emitted are distributed with respect to the axis of the nuclear spin. Since this axis is not provided with an arrow, by reason of symmetry we would expect that this distribution, whatever it is, be symmetrical with respect to both sides of the axis. The experiment suggested by Yang and Lee, and performed by Wu and her collaborators,[2] showed a distinct lack of symmetry—more electrons were emitted on one side than on the other. The nucleus Ni-60 resulting from the decay of Co-60 is in an excited state and reverts to the ground state with an emission of gamma rays. The distribution of these gamma rays was measured and found to be symmetrical with respect to the spin axis of the nucleus Co-60.

Let us consider the experimental arrangement in more detail. The Co-60 sample was spread over a small disk, above which a beta-ray detector was placed. Due to the thermal motion of the atoms the axes of the nuclear spins are randomly distributed and the directions of the emitted electrons of the whole sample will be centrally symmetrical no matter what the distribution for a single nucleus is. In order to study the intrinsic distribution of the directions a magnetic field is applied perpendicularly to the plane of the disk. This magnetic field will orient the nuclei with the spin axis all in one direction. To reduce the thermal motions, the entire setup is placed in a cryostat and the ambient temperature reduced to near zero absolute. The strength of the magnetic field is so low that it has no bearing on the paths of the beta rays; it is just sufficient to orientate the Co nuclei.

To obtain the distribution of the directions of beta-ray emission, we should rotate the beta-ray detector around the Co disk, but it is more convenient to leave the detector in a fixed position and instead rotate the magnetic field over 360° (which is tantamount to rotating the Co nuclei). In discussing the results of the experiment, however, it will be assumed that we rotate the beta detector.

We are particularly interested in the two positions just above and below the disk, *B* and *A*. We have already pointed out that right–left symmetry would demand that an equal number of electrons arrive at *A* and *B*. Instead it has been found that the amounts are definitely not equal, whereas the analogous measurement with gamma rays showed equality, within the errors of measurement.

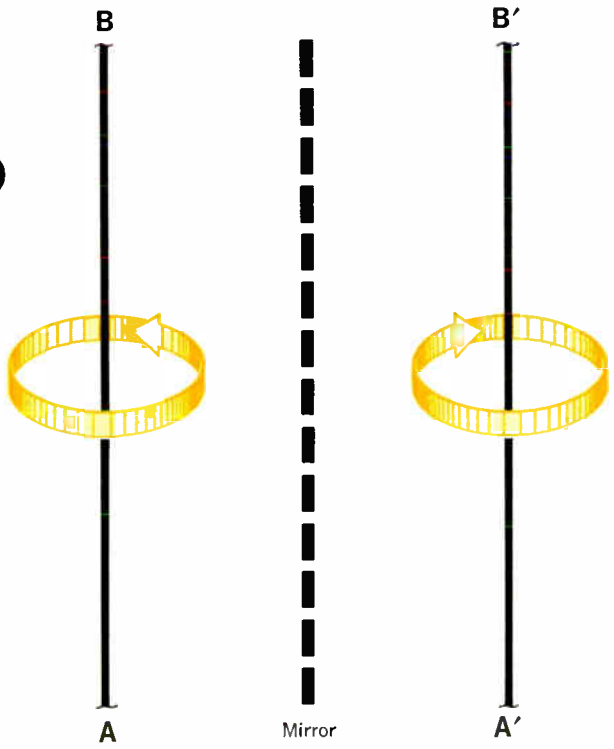


Fig. 3. Original and mirror image of spin of Co-60 nucleus.

Let us express this result in terms of "original" and "mirror image." We see (Fig. 3) that mirroring reverses the spin of the Co nucleus but leaves A and B unchanged, yet the experiment shows that reversing the nuclear spin reverses the inequality of A and B. If in the original we had $A > B$, the arrangement portrayed by the mirror image would in real life yield $A < B$. The mirror image does not tell the truth; parity is not conserved.

Beta rays have been known to exist for more than 60 years, but it was as late as 1956 that the question of conservation of parity arose. The reason was that before the discovery of uranium fission and closer study of cosmic rays, not many nuclear reactions were available for investigation and the quantities of radioactive substances obtainable were rather small.

With the discovery of the asymmetry of the beta decay of Co-60 we obtained a means whereby we could communicate to intelligent beings on other planets how we define positive–negative charge and right–left. We could refer to the beta decay of Co-60; however, we would have to be sure that their world is built of ordinary matter and not of antimatter.

After the discovery of the peculiar behavior of Co-60, attempts were made to find an explanation. No real explanation has been found, but an argument has been put forth that might pacify the physicists—that there are instances where, in beta decay, not electrons but positrons (positive electrons) are emitted. In those instances there is the same lack of symmetry as with Co-60, but with left and right reversed. We could then guess that charge somehow involves helicity or "handedness." If we could find a mirror in which an electron could see its own image, what would it see? The answer is—a positron.

This fact is reminiscent of the magnetic needle, which violates parity conservation. As soon as we replace the needle by a current loop, parity conservation is restored. Similarly, perhaps we should replace charge by a loop of something; but this argument has two weak points. First, if in strong interactions parity is conserved, why should it not be conserved in weak interactions? If there were any hidden helicity in charge, we would have to explain why it is manifest in weak interactions but not in strong ones. Second, the asymmetry is not 100 per cent. More electrons arrive, say, at A than at B, but still some do arrive at B. The ratio of the numbers of electrons arriving at A and B, respectively, depends on the velocity of the beta particles. With increasing velocity the asymmetry increases, approaching 100 per cent with the electron velocity tending towards the velocity of light.

Polarization and spin

In conclusion we should like to express the asymmetry in terms of polarization, which is easier to visualize. Let us provide (arbitrarily) the axis of the electron spin with an arrow. The quantum rules then demand that the linear velocity of the electron be either parallel or antiparallel to the spin axis; no other directions are permitted. The electron progresses either like a right-hand screw or a left-hand one. We say that the electron is polarized, either right- or left-handed.

Let us now return to our nucleus Co-60. We have said that this nucleus has a spin of 5, while the nucleus Ni-60, into which Co-60 is being transmuted, has a spin of 4. Conservation of angular momentum demands that we account for the missing spin. We shall find this spin in the half spin of the emitted electron and the half spin of the emitted neutrino. So far we have neglected the tiny neutrino to keep our story from becoming too involved. The neutrino originally was postulated by Pauli in order to save the principles of conservation of energy and angular momentum in beta decay. A few years ago the existence of the elusive neutrino was confirmed by an ingenious and elaborate experiment. The story of the neutrino is fascinating, but here we must content ourselves with the statement that the neutrino is a constant companion of the electron in beta decay and that it has a spin of one half.

The spin axis of the emitted electron will be the same as the spin of the Co nucleus (because only in this case will we have $5 = 4 + \frac{1}{2} + \frac{1}{2}$) and its velocity will be directed either toward A or toward B (Fig. 3). In one case the electron will be right-polarized, and in the other, left. The asymmetry in beta decay means then that beta particles are mostly right-polarized. Thus expressed, the asymmetry in beta decay is most striking.

It was pointed out in the previous part of this article that it is not possible to observe the spin of a single electron, and the same applies to its polarization. It is only possible to measure the distribution of polarization statistically, as was done in the experiments of Wu *et al.* It is also possible to separate right-polarized electrons from left-polarized ones. Strangely enough, no engineering uses have yet been found for either free electron spin or electron polarization.

REFERENCES

1. Lee, T. D., and Yang, C. N., "Question of Parity Conservation in Weak Interactions," *Phys. Rev.*, vol. 104, Oct. 1, 1956, pp. 254–258.

2. Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D., and Hudson, R. P., "Experimental Test of Parity Conservation in Beta Decay," *Ibid.*, vol. 105, Feb. 15, 1957, pp. 1413–1415.

Authors

B. M. Oliver (F), 1965 IEEE President, received the A.B. degree in electrical engineering from Stanford University in 1935 and the M.S. degree in electrical engineering from the California Institute of Technology in 1936. He then spent a year in Germany as an exchange student under the auspices of the Institute of International Education. In 1940 he received the Ph.D. degree, magna cum laude, in electrical engineering from the California Institute of Technology. As a member of the technical staff of the Bell Telephone Laboratories in New York City from 1940 to 1952, he worked on the development of automatic tracking radar, television, information theory, and efficient coding systems. In 1952 he joined Hewlett-Packard Company as director of research and in 1957 was appointed vice president of research and development.

Dr. Oliver holds over 40 U.S. patents in the field of electronics. He was elected Director-at-Large of the IRE in 1958 and has served as chairman of the San Francisco Section of IRE and as a member of the Board of Directors of WESCON. Following the merger of IRE and AIEE, he was elected Vice President of the newly formed IEEE for 1963 and 1964. He is also a member of the American Astronautical Society.

I. P. Kaminow (M) is a member of the Crawford Hill Laboratory, Bell Telephone Laboratories, Holmdel, N.J., where he is engaged in research on optical communication systems. He has also worked on investigations of magnetic and dielectric materials, ferrite devices, and microwave antennas. He joined Bell Telephone Laboratories in 1954 in Whippany, N.J., as a member of the company's communication development training program after working for two years as a member of the technical staff of Hughes Aircraft Company's Microwave Laboratory, Culver City, Calif., where he was concerned chiefly with slot-array antennas. In 1956 he received a Bell Laboratories Communications Development Training Program Fellowship for Harvard University, and in 1960 he received the Ph.D. degree in applied physics from Harvard. He also holds the B.S.E.E. degree from Union College (1952) and the M.S.E. degree from the University of California at Los Angeles (1954). Dr. Kaminow is a member of Sigma Xi and the American Physical Society.

Philip L. Alger (F, L) has been adjunct professor of electrical engineering at Rensselaer Polytechnic Institute since 1958. He was associated with the General Electric Company, in Schenectady, N.Y., from 1919 to 1959. He received the B.S. degree from St. John's College, Md., in 1912, the B.S.E.E. degree from Massachusetts Institute of Technology in 1915, and the M.S. degree from Union College in 1920. During World War I he was a lieutenant in the U.S. Army Ordnance Department. His books include *The Nature of Polyphase Induction Machines* (Wiley, 1951), *Mathematics for Science and Engineering* (McGraw-Hill, 1957), *The Nature of Induction Machines* (Gordon & Breach, 1965), and, with S. P. Olmsted and N. A. Christensen, *Ethical Problems of Engineers*, to be published by Wiley in June 1965. He was formerly a Director of AIEE and received the AIEE's Lamme Medal for 1958. He is a Fellow of ASME, the American Society for Quality Control, and the American Association for the Advancement of Science, and an Eminent Member of Eta Kappa Nu.

J. L. Salpeter. A biographical sketch of Dr. Salpeter appears on page 197 of the March issue.