THE BELL SYSTEM TECHNICAL JOURNAL

A JOURNAL DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

TABLE OF CONTENTS AND INDEX

VOLUME XIX

1940

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PRINTED IN U. S. A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XIX, 1940

Table of Contents

JANUARY, 1940

APRIL, 1940

JULY, 1940

OCTOBER, 1940

Index to Volume XIX

V

W

VOLUME XIX JANUARY, 1940 NUMBER 1

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XIX *January, 1940* *No. 1*

The Physical Basis of Ferromagnetism

By R. M. BOZORTH

After an introductory review of the general nature of the theory of magnetic phenomena and the magnitudes of the atomic forces involved, there is a discussion of Ewing's theory, its results and limitations. The later theory of Weiss is then given briefly in order to fix the concept of the molecular field. In order to elucidate the nature of this field a digression is made to discuss the atomic structure of the ferromagnetic elements and elements having similar structures. With this as a basis the physical nature of the molecular field is discussed at some length. Its relation to the structure of domains, particularly the nature of the boundaries between domains, is brought out.

Finally there is a review of the gyromagnetic effect, its significance for magnetic theory, the principal expermental method for its determination, and the numerical results supporting the idea that the spin of the electron and not its orbital moment is responsible for ferromagnetism.

Introduction

IN THE last five or ten years the theory of ferromagnetism has shown indications of maturity. For the first time a plausible story can be told concerning the ultimate magnetic particle, the essential nature of the atom of a ferromagnetic substance, the kind of forces which determine the properties of magnetic crystals, the effect of strain on magnetic materials and the manner in which these various phenomena combine to determine the properties of commercial materials. It is true that the story is largely qualitative, and that there are still many points that are uncertain or missing entirely, but nevertheless it is possible to describe the major features with some confidence.

The fundamental magnetic particle is the spinning electron. One might think that the orbital motions of the electrons in the atom would also contribute to ferromagnetism, owing to their magnetic

moments, but it has now been established that when the magnetization is altered all that changes is the direction or "sense" of the spin of certain of the electrons in the atoms—the orbital motions remain practically unchanged.

The electrons that are responsible for the magnetic properties of iron, cobalt, nickel and their alloys lie in a definite "shell" in the atom. As shown in Fig. 1, there are four shells or regions, more or

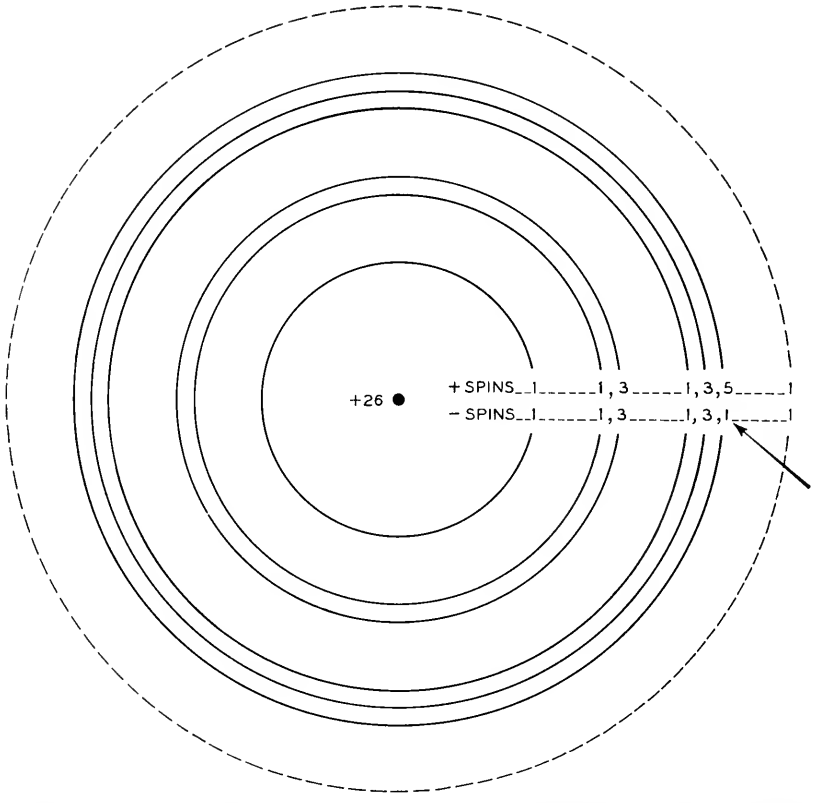


Fig. 1—Electron shells in an atom of iron. The arrow indicates the incomplete sub-shell that is responsible for ferromagnetism. The numbers specify how many electrons with each spin are in the corresponding sub-shells.

less well defined, into which all the electrons circulating about the nuclei of these atoms may be divided when the atom is separated from its neighboring atoms, as it is, for example, in a gas. Some of these shells are subdivided as shown. When the atoms come closer together as they do in a solid, the fourth or outermost shell of each becomes disrupted, and the two electrons which comprised it wander from atom to atom and are the "free" electrons responsible for

electrical conduction. The electrons in the outer part of the third shell are those responsible for the distinctive kind of magnetism found in iron, cobalt and nickel. Some of these electrons spin in one direction and some in the opposite, as indicated, so that their magnetic moments neutralize each other partially but not wholly, and the excess of those spinning in one direction over those spinning in the other causes each atom as a whole to behave as a small permanent magnet.

The well-established kinetic theory of matter tells us that if each atom were to act independently of its neighbors, the atoms would be vibrating and rotating so energetically that they could not be aligned even with the strongest field that can be produced in the laboratory. To explain the kind of magnetic properties found in iron, therefore, it is necessary that there be some internal force capable of making the magnetic moment of a group of neighboring atoms lie parallel to each other—the small atomic "permanent magnets" of each group must point in the same direction so as to provide a magnetic moment great enough to permit a realignment when subjected to external fields. Recently it has been shown by independent means that there is such a force in just those elements which are ferromagnetic, and it is from this force that the difference between magnetic and non-magnetic materials arises. The force is electrostatic in nature and is called "exchange interaction" by the atomic-structure experts, the wave mechanicians, who have shown its existence and calculated its order of magnitude. This force maintains small groups of atomic magnets parallel against the forces of thermal agitation. (When the material is heated so hot that the disordering action of the agitation becomes strong enough to overpower the forces of "exchange interaction" the material loses its ferromagnetism; in iron this happens at 770° C.)

But why then is not every piece of iron a complete permanent magnet? For some reason not understood at present, at ordinary temperatures the electrostatic forces of exchange interaction maintain the elementary magnets parallel only over a limited volume of the specimen. This volume is usually of the order of 10^{-8} or 10^{-9} cubic centimeters and contains a million billion atoms and is of course invisible. Such a volume is said to be saturated because the atomic magnets are all pointing in the same direction, and has been given the name "domain." Thus a magnetic material at room temperature, before it has been magnetized by subjecting it to the influence of a magnetic field, is divided into a great many domains each of which is magnetized to saturation in some direction generally different from that of its neighbors. The net or vector sum of the magnetizations is zero, and externally the material appears to be unmagnetized but in

reality the magnetization at any one point is very intense. When a magnetic field is applied by bringing near the metal a permanent magnet or a coil of wire carrying a current, the magnetization of the material as a whole is increased to a definite value. We believe that what then takes place is simply a change in the direction of the magnetizations of the domains. If we represent the magnetization of any domain by a vector, the effect of the externally applied field will be represented by the rotation of these vectors—rotations not accompanied by any changes of length.

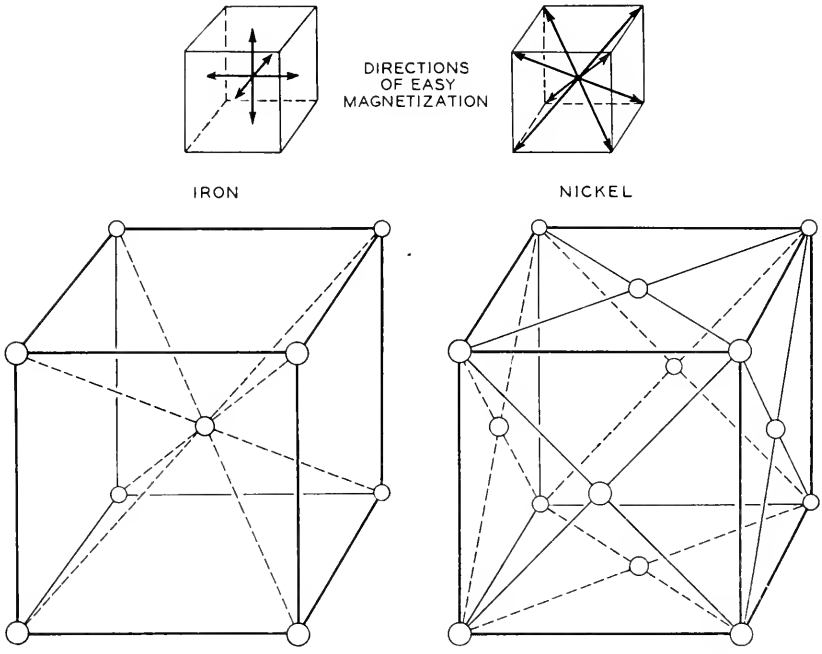


Fig. 2—The positions of the atoms and the directions of easy magnetization in crystals of iron and of nickel.

Recently much has been learned about the magnetic properties of materials by a study of single crystals. Ordinary metals are composed of a great many crystals often too small to be seen easily by the naked eye. But in the last few years methods have been found for making large crystals of almost all the common metals, crystals as large as the more familiar ones of rock candy and even of quartz. Experiments on such crystals of iron show that they are much more easily magnetized in some directions than in others.

This dependence of ease of magnetization on direction is illustrated in Fig. 2 for iron and nickel in relation to the positions of the atoms in

the crystals. The circles represent the positions which centers of atoms take up on an imaginary framework or lattice. Because of the smallness of atomic dimensions only a small fraction of the atoms in a crystal of ordinary size are shown, but the same pattern, the unit of which is outlined by solid lines, extends throughout the whole of the single crystal. The arrows indicate the directions of "easiest" magnetization, which are different for the two materials as may be noticed.

In order to give a notion of the absolute and relative sizes of crystals and domains and atoms with which magnetic processes are concerned, it may be pointed out that a piece of ordinary iron a cubic centimeter in volume may contain about 10,000 single crystals, and that each crystal contains on the average 100,000 domains each with from 10^{14} to 10^{15} atoms.

Although this article is not concerned primarily with the details of the changes in magnetization that occur when a magnetic field is applied, a brief description of such changes is desirable. In a crystal of iron the directions of easy magnetization are parallel to the cubic axes, that is, they are the six directions parallel to the edges of the cube which represents the structure. When such a magnetic material is unmagnetized as a whole a portion of one of the crystals in it may be represented by the highly schematic Fig. 3(a). As shown, each of the domains, represented by the arrows, circles and crosses, is magnetized in one of the directions of easy magnetization, equal numbers in each of the six directions. When a weak field is applied in the direction indicated and its strength gradually increased to a high value, the magnetizations of the domains change suddenly and their directions approach coincidence with that of the magnetic field. This is usually accomplished by the displacements of domain boundaries, these moving so that some domains grow at the expense of others in which the magnetization lies in a direction further from that of the field. When the field has been increased to such a strength that practically all the domains are oriented as shown in (b) and the crystal is really just one large domain, a second process commences: the magnetization changes slowly in direction until finally it is parallel to the field, and then changes no more. The material is then said to be saturated, as shown in (c).

Figure 3 is drawn to illustrate the changes in magnetization that occur in a single crystal of iron. Iron as we ordinarily see it is composed of a great many minute single crystals, but the changes in magnetization that occur in each one of these crystals are just those which have been described, the magnetization of the whole polycrystalline material being the sum of the magnetization of the parts.

The most definite evidence of the existence of domains is the Barkhausen effect. To produce and detect it, a piece of magnetic material is wound with wire the ends of which are connected to a vacuum tube amplifier. When the magnetization of the material is changed, as *e.g.* by moving a permanent magnet near it, a rustling sound or a series of clicks may be heard in phones or in a loud speaker

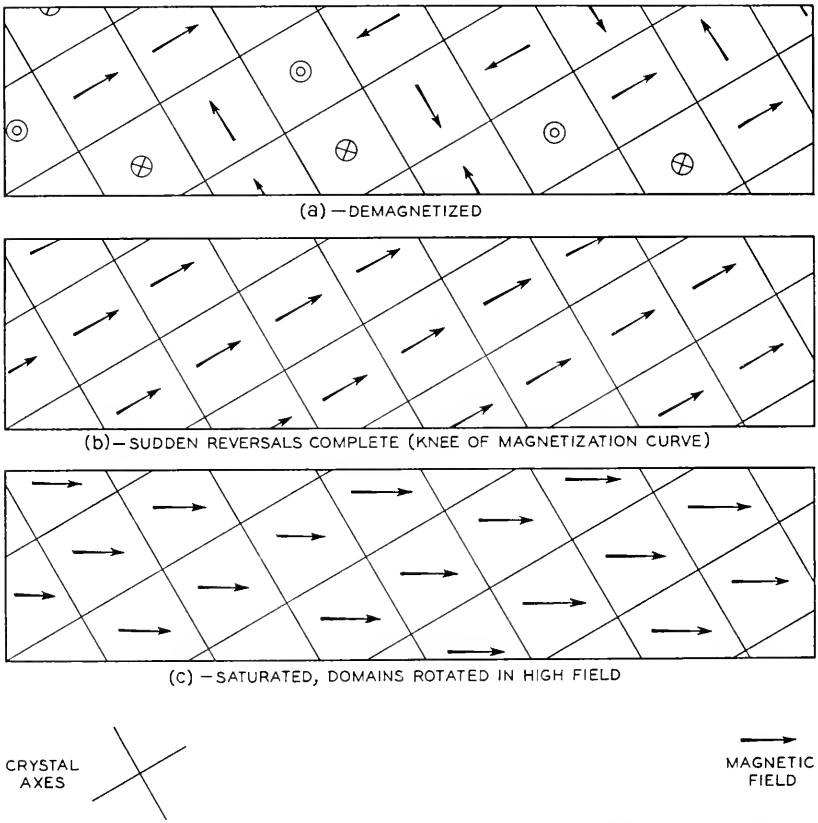


Fig. 3—Domains in a single crystal of iron. As the magnetic field increases in strength the magnetic moments first change suddenly (*a* to *b*) by displacement of the boundaries between them, then rotate smoothly (*b* to *c*).

connected to the output end of the amplifier. Every such click is ascribed to the sudden change in direction of magnetization in a single domain, and from measurements of the sizes of the clicks we get our best estimate of the sizes of the domains. Even more direct evidence of the existence of domains and the changes that they undergo has been obtained recently by spreading colloidal iron oxide over the surface of a magnetic material and looking at it under a microscope.

The regular pattern observed [1] is similar in nature to the familiar one obtained when iron filings are sprinkled near a permanent magnet; the fine colloidal particles are necessary in this case because the whole scale is small. This micro-pattern changes when the applied field changes, and the difference is attributed to the redistribution or reorientation of groups of domains. These patterns are obtained only on magnetic materials and are found on them even when the material is unmagnetized; such a one is shown in Fig. 4.

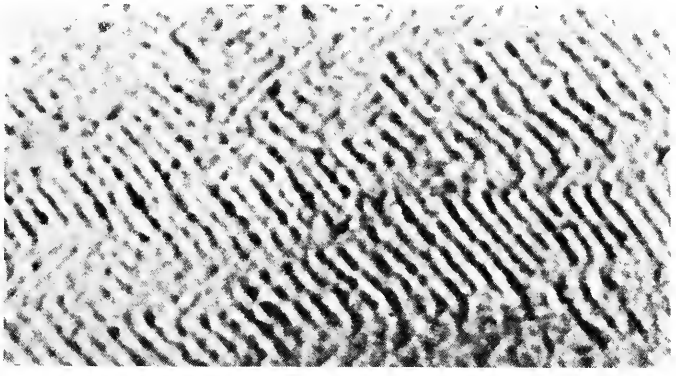


Fig. 4—The powder pattern produced by colloidal iron oxide on the surface of a demagnetized silicon-iron crystal, showing the presence of inhomogeneous magnetic fields. Magnification about 1000.

Magnitudes of Magnetic Forces

Ferromagnetic theory has been made difficult by the fact that the magnetic forces between the electrons in an atom are small compared to the electrostatic forces. The latter force between two electrons of charge e (in e.s.u.), a distance a apart, is equal to

$$e^2/a^2.$$

The magnetic force between the same electrons depends on the speed of the charges as well as on their magnitudes, and, when the direction of motion is perpendicular to the line joining them, is equal to

$$\frac{e^2}{a^2} \cdot \frac{v^2}{c^2},$$

where v/c is the ratio of the speed of each electron to the speed of light. Since v/c is usually of the order of 0.01, these magnetic forces

[1] L. W. McKeehan and W. C. Elmore, *Phys. Rev.*, *46*, 226–228 (1934). See also the earlier experiments by F. Bitter, *Phys. Rev.*, *41*, 507–515 (1932). See also the account by Elmore in F. Bitter's Introduction to Ferromagnetism, McGraw-Hill, New York, 55–66 (1937).

are about 10^{-4} of the electrostatic forces. The difference is even greater when electrostatic forces between electrons and nuclei, or between nuclei, are compared with magnetic forces. The magnitudes of these forces for a specific hypothetical arrangement are shown in Fig. 5.

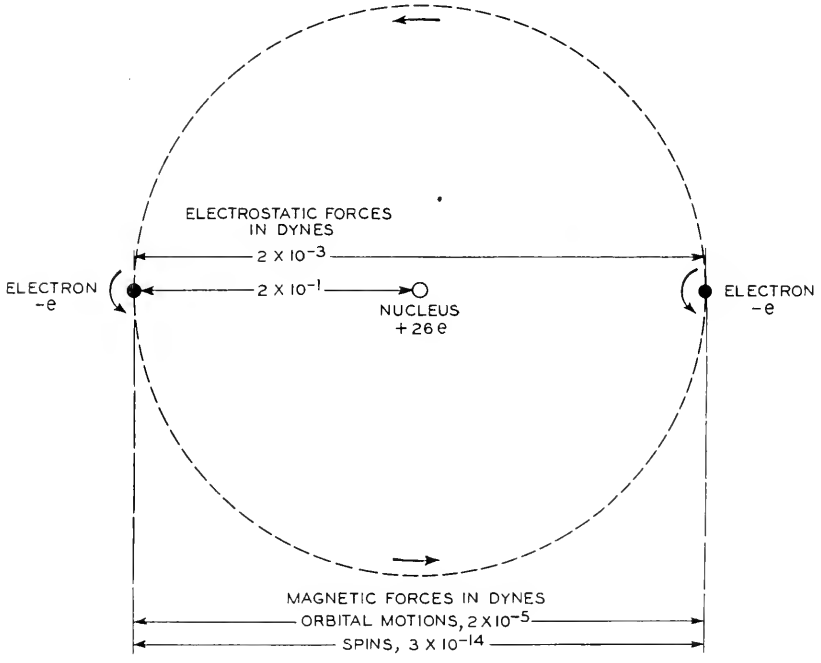


Fig. 5—The magnitudes of the forces in a hypothetical iron-like atom, showing that electrostatic forces are more powerful than magnetic forces.

Consider the magnitude of magnetic forces from another point of view. The magnetic energy of a permanent magnet of moment μ_A in a field of strength H is

$$E = -\mu_A H,$$

when μ_A and H are parallel. In a magnetic substance we may regard the atomic magnets as being held parallel by a fictitious field H_i. When the material is heated to the Curie temperature, θ, the energy of thermal agitation ($\approx k\theta$) destroys the alignment of the atomic magnets by the fictitious or "internal" field H_i. Then

$$k\theta \approx \mu_A H_i.$$

For iron, $\theta = 1043°$ K. and $\mu_A = 2.04 \times 10^{-20}$ erg/gauss, thus the

energy per atom is

$$k\theta = 1.4 \times 10^{-13} \text{ erg} = 0.09 \text{ electron-volt}$$

and the internal field

$$H_i = 7{,}000{,}000 \text{ oersteds.}$$

Although this field is much stronger than any so far produced in the laboratory, the energy involved is small compared to that which controls chemical binding. For example, the energy of ionization of the helium atom is about 25 electron volts. Another way of showing that the magnetic forces are small compared to the electrostatic forces holding atoms together, is to compare the Curie temperature with the temperature of vaporization.

The calculation of magnetic forces by theory is thus extremely difficult, because they are but small additions to the electrostatic forces which themselves cannot usually be calculated with much precision.

Ewing's Theory

Ewing [2] was one of the first to attempt to explain ferromagnetic phenomena in terms of the forces between atoms. His theory will be described briefly here, since many physicists today, when thinking about magnetic phenomena, still go back to Ewing's ideas of fifty years ago. He assumed with Weber that each atom was a permanent magnet free to turn in any direction about its center. The orientations of the various magnets with respect to the field and to each other were supposed to be due entirely to the mutual magnetic forces. The I, H curve and hysteresis loop were calculated for a linear group of such magnets and were determined experimentally using models having as many as 130 magnets arranged at the points of a plane square lattice.

The calculations for a linear chain show that as the field is gradually increased in magnitude from zero there is at first a slow continuous rotation of the magnets, then a sudden change in orientation and finally a further continuous rotation until the magnets lie parallel to the field. The I, H curves calculated for such a group of magnets resemble in general form the actual curves of iron: they show a permeability first increasing then decreasing, and saturation and hysteresis.

A magnetization curve and a hysteresis loop obtained [3] with a model of 130 magnets in square array, are shown in Fig. 6. Experi-

[2] J. A. Ewing summarized in "Magnetic Induction in Iron and Other Metals," The Electrician, London, 3d ed. (1900).

[3] J. A. Ewing and H. G. Klaassen, *Phil. Trans. Roy. Soc., 184A*, 985–1039 (1893).

ments with the model showed a variety of other phenomena including rotational hysteresis loss and its reduction to zero in high fields, the effect of strain on magnetization, the existence of hysteresis in the strain *vs.* magnetization diagram, the effect of vibration and the existence of time lag and accommodation with repeated cycling of the field.

Ewing's general method may be illustrated by calculating the magnetization curve and hysteresis loop for an infinite line of parallel

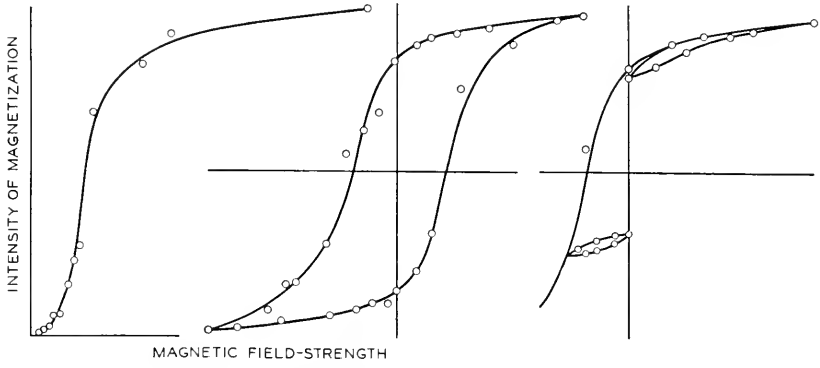


Fig. 6—A magnetization curve and hysteresis loops of a Ewing model of 130 pivoted magnets in square array.

equally spaced magnets (Fig. 7a). It is done most simply by considering first the magnetic potential energy [4] of a magnet of moment μ_A and length l, in the field of a similar magnet:

$$W = -\frac{\mu_A^2}{r^3} P_2(\theta) - \frac{\mu_A^2 l^2}{r^5} P_4(\theta) - \frac{\mu_A^2 l^4}{r^7} P_6(\theta) - \cdots. \qquad (1)$$

Here r is the distance between the centers of the magnets and the $P(\theta)$'s are Legendre functions of the angle, θ, between the direction of the moment of the magnet and the line joining the magnet centers.

$$P_2(\theta) = (1 + 2\cos 2\theta)/4,$$
$$P_4(\theta) = (9 + 20\cos 2\theta + 35\cos 4\theta)/64,$$
$$P_6(\theta) = (50 + 105\cos 2\theta + 126\cos 4\theta + 231\cos 6\theta)/512.$$

The potential energy per magnet, W_1, for an infinite straight row of magnets can easily be obtained by summing W for all pairs.

$$W_1 = -\frac{2\mu_A^2}{r^3}[1.20P_2(\theta) + 1.04P_4(\theta)(l/r)^2 + 1.01P_6(\theta)(l/r)^4 + \cdots]. \qquad (2)$$

[4] G. Mahajani, *Phil. Trans. Roy. Soc., 228A*, 63–114 (1929).

The behavior of the line when subjected to a field H may be found by adding to W_1 the energy term $-H\mu_A \cos(\theta_0 - \theta)$, where θ_0 is the angle between the line of centers and the direction of the field, and

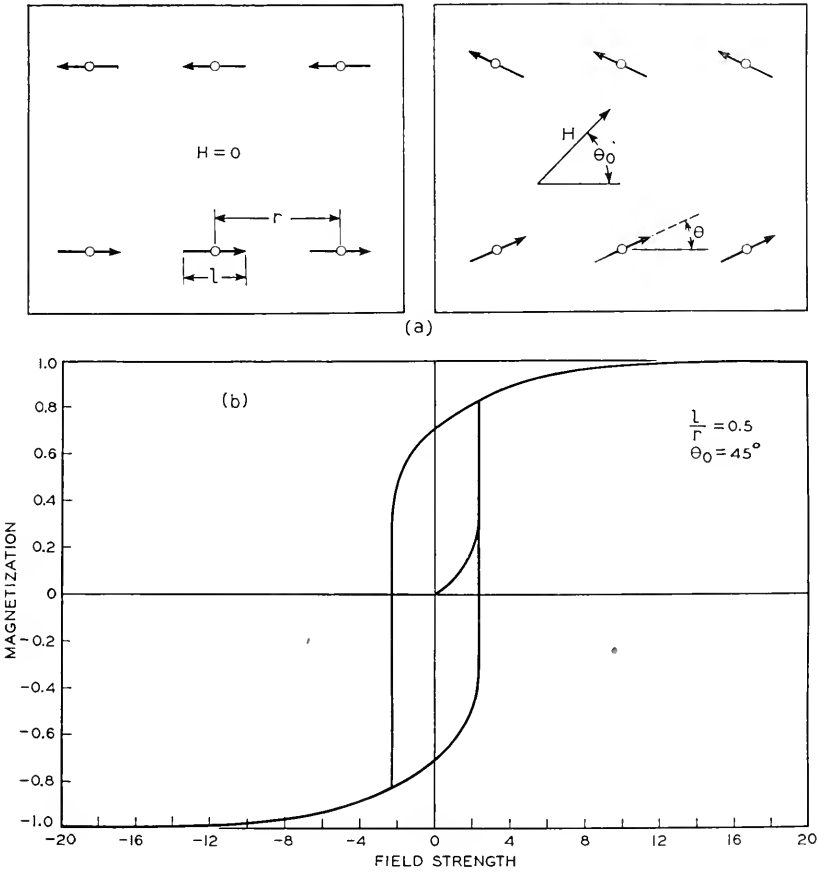


Fig. 7—A magnetization curve and hysteresis loop for an infinite line of equally spaced magnets originally "demagnetized."

finding the value of θ which makes this total energy a minimum for given values of θ_0 and H:

$$\frac{d}{d\theta}[W_1 - H\mu_A \cos(\theta_0 - \theta)] = 0.$$

This gives

$$H = \frac{(d/d\theta)W_1}{\mu_A \sin(\theta_0 - \theta)}.$$

The component of magnetization parallel to H is

$$I = I_s \cos (\theta_0 - \theta),$$

where I_s is the saturation magnetization. By starting with half of the line of magnets pointing in a direction opposite to that of the other half, the initial magnetization is zero and an unmagnetized or demagnetized material is simulated. Thus a magnetization curve and a hysteresis loop of this assemblage are obtained by plotting H against I. Such a plot is shown in Fig. 7(b), with the scale of H determined by the magnitudes of μ_A and r. The curves are obviously similar to those for real materials.

Limitations of Ewing's Theory

So far, this calculation is equivalent to what Ewing did over four decades ago. But now we know the crystal structure of iron and in particular the distances between the atoms. We also know the magnetic moment of each iron atom and know, therefore, the value of μ_A/r^3 which determines the scale of H. Using the appropriate values $\mu_A = 2.0 \times 10^{-20}$ erg/gauss and $r = 2.5 \times 10^{-8}$ cm, the coercive force H_c for $l/r = 0.1$ is found to be 4600 oersteds. This is affected somewhat by the ratio l/r, but in any case H_c is found to be of this order of magnitude unless l/r is very close to unity. This magnitude of H_c is greater by a factor of 10^5 than the lowest value obtained experimentally, 0.01. Similarly the initial permeability, μ_0, according to the model is about unity while observed values for iron range from 250 to 20,000. Adjustment of l/r to higher values decreases μ_0.

This calculation of the magnetization curve and hysteresis loop are based on a very much idealized model, and it is difficult to estimate the error to which it may lead. One factor that has been completely neglected is the fluctuation in energy. A much better approximation would be to calculate the magnetic potential energy of a group of magnets arranged in space in the same way that the iron (or nickel) atoms are arranged in a crystal. This has been done by Mahajani [4] who showed that application of Eq. (1) with $l = 0$ (but summed to account for the effects of all magnets in the structure) leads to the result that the magnetic potential of the space array is independent of θ, in other words one orientation of the dipoles is as stable as any other and the magnetization curve would go to saturation in infinitesimal fields no matter in what direction H might be applied. If l is finite, the stable positions of the magnets are parallel to the body-diagonals of the cube which is the unit of the crystal structure, and

this becomes therefore the direction of easy magnetization, a situation which is correct for nickel but decidedly not so for iron. The best correspondence between the action of the model and of iron itself is obtained if the model is made by placing a small circular current of electricity, instead of a magnet with finite length, at each lattice point of the space array. In the latter case we can explain the direction of easy magnetization in iron and the variation of magnetic energy with direction in the crystal.

In considering Ewing's model it is appropriate to estimate the energy of thermal agitation and to compare it with the magnetic potential energy as calculated from the model. Substituting in Eq. (2) the same values of μ_A and r as were used above, we obtain 10^{-16} erg per atom for the magnetic potential energy in zero field. This is to be compared with the rotational energy of a single molecule at room temperature, 2×10^{-14} erg per atom as given by the kinetic theory. Thus the energy of thermal agitation is 200 times as great as the calculated magnetic energy. Even at liquid air temperatures the thermal agitation would prevent the atomic magnets from forming stable configurations. Without some additional force the model Ewing used would behave as a paramagnetic rather than a ferromagnetic solid.

In a real material, however, it is now well established that there are very powerful forces, not contemplated when Ewing made his model and proposed his theory, which maintain parallel the dipole moments of neighboring atoms. These are the electrostatic forces of exchange (see p. 24) which Heisenberg suggested are powerful enough to align the elementary magnets against the disordering forces of thermal agitation, forces much larger than those of magnetic origin. Theory accounts only for the order of magnitude of these forces. Our best estimate of the corresponding energy of magnetization is obtained by assuming that it is equal to the energy of thermal agitation at the Curie point, $\frac{1}{2}k\theta$. For iron ($\theta = 1043$ °K) this gives 7×10^{-14} erg per atom.

The Weiss Theory

In order to understand how atomic forces give rise to ferromagnetism it is desirable to review briefly Weiss's theory [5] of ferromagnetism, which introduces a so-called "molecular field" that presently will be identified with the nature of these forces. This theory is an extension of Langevin's theory of a paramagnetic gas. The original Langevin theory culminated in a formula relating the magnetization, I, to the field-strength, H, and the temperature, T; this is the hyperbolic co-

[5] P. Weiss, *Jour. de physique* (4) *6*, 661–690 (1907). P. Weiss and G. Föex, "Le Magnetisme," Colin, Paris (1926).

tangent law,

$$\frac{I}{I_0} = \text{ctnh}\,\frac{\mu_A H}{kT} - \frac{kT}{\mu_A H}.$$

In deriving this the assumptions are made that the elementary magnets, each of moment μ_A, are subject to thermal agitation and momentarily may have any orientation with respect to the direction of the field, and that they are too far apart to influence each other. Quantum theory alters the second of those assumptions by stating that in such an ensemble of elementary magnets (atoms) there will be only a limited number of possible orientations, in the simplest case only two, one parallel and the other antiparallel to the direction of the field. In this case the equation corresponding to Langevin's is

$$\frac{I}{I_0} = \tanh\,\frac{\mu_A H}{kT}. \tag{3}$$

These two theoretical relations are plotted for variable H and constant T (room temperature) in Fig. 8, the constants being those for

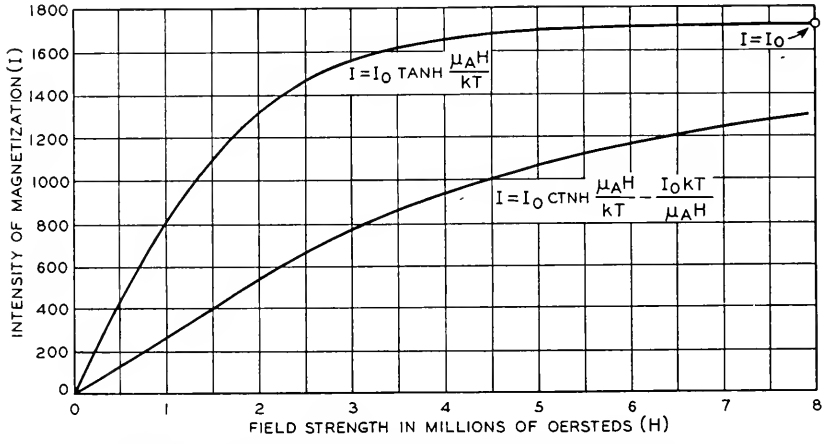


Fig. 8—With no helpful mutual action between atoms, enormous fields would be necessary to saturate a magnetic material.

iron ($I_0 = 1740$, $\mu_A = 2.04 \times 10^{-20}$ erg/gauss). It is obvious that with the highest fields so far attained in the laboratory (about 300,000 oersteds) the magnetization would attain only a small fraction of its final value I_0 if this law were obeyed, and in this range I would be sensibly proportional to the field-strength:

$$I = \frac{CH}{T},$$

where C is a constant. This relation, known as Curie's Law, is obeyed by some *paramagnetic* though not by ferromagnetic substances. It is usually written with I/H denoted by the symbol χ, representing susceptibility:

$$\chi = \frac{C}{T}.$$

Many more paramagnetic substances obey the similar "Curie-Weiss Law":

$$I = \frac{CH}{T - \theta}. \tag{4}$$

Weiss pointed out the significance of θ in this equation: it means that the material behaves magnetically as if there were an additional field, NI, aiding the true field H. This equivalence is shown mathematically by putting $\theta = NC$ in Eq. (4) with the result

$$I = \frac{C(H + NI)}{T}.$$

The quantity represented by NI is called the "*molecular field*" and that by N the "*molecular field constant.*" It is interpreted by supposing that the elementary magnet does have an influence on its neighbors, contrary to the assumptions of the simple Langevin theory.

The significance of the molecular field for ferromagnetism is now apparent if we replace the H by $H + NI$ in the more general Eq. (3) and examine the resulting equation:

$$\frac{I}{I_0} = \tanh \frac{\mu_A(H + NI)}{kT}. \tag{5}$$

This equation is perhaps the most important in the theory of ferromagnetism. It indicates that even in zero field there is still a magnetization of considerable magnitude, provided the temperature is not too high. Putting $H = 0$ and

$$\theta = \mu_A N I_0/k,$$

Eq. (5) reduces to

$$\frac{I}{I_0} = \tanh \frac{I/I_0}{T/\theta}. \tag{6}$$

This purports to specify the magnetization at zero applied field by a function that is the same for all materials, when the magnetization is expressed as a fraction of its value at absolute zero and the temperature as a fraction of the Curie temperature on the absolute scale. This magnetization *vs.* temperature relation, plotted as the solid line of Fig.

9, means that *at all temperatures below θ the intensity of magnetization has a definite value even when no field is applied.*

How is it then that a piece of iron can apparently be unmagnetized at room temperature? The answer, given by Weiss, is that below the Curie point all parts of the iron are magnetized to saturation but that different parts are magnetized in different directions so that the overall

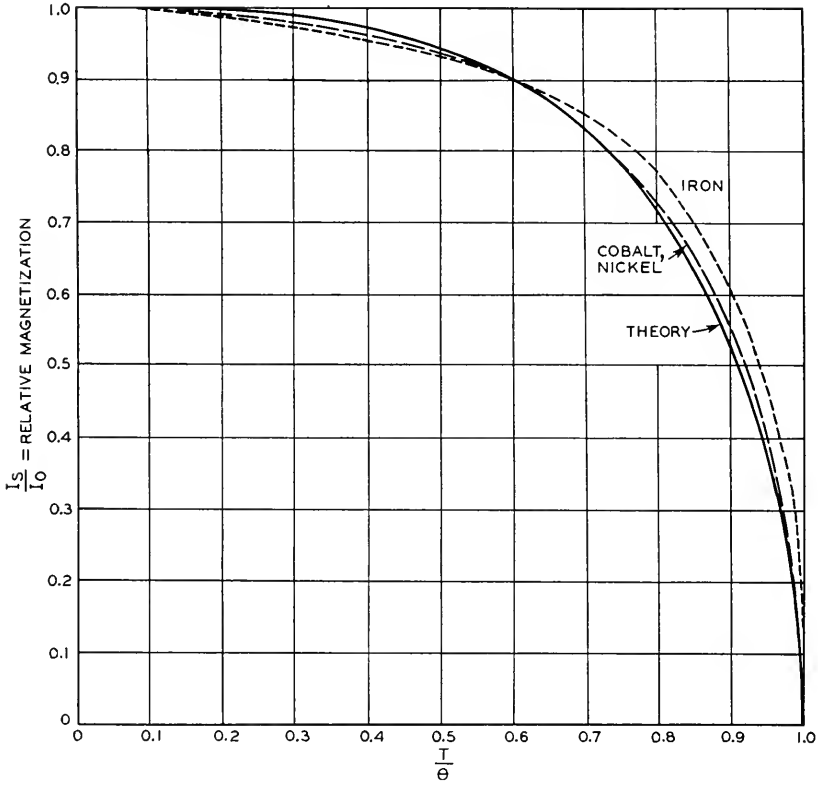


Fig. 9—Dependence on the temperature of the saturation magnetization of iron, cobalt and nickel, as compared with theory.

effect is zero. This is the concept of the domain, already discussed. According to this conception the I of Eq. (5) is that of a domain and is determined experimentally by measuring the magnetization of a specimen when all domains are parallel, *i.e.*, at (technical) saturation ($I = I_s$). Eq. (6) should then be written

$$\frac{I_s}{I_0} = \tanh \frac{I_s/I_0}{T/\theta}.$$

It is a problem of theoretical physics to determine the nature of the molecular field. Before discussing what progress has been made in doing this it will be necessary to review some of our knowledge of the structure of the atoms with which we are concerned.

Atomic Structure of Ferromagnetic Materials

The structure of an isolated iron atom has already been shown in Fig. 1. The twenty-six electrons are divided into four principal "shells," each shell a more or less well defined region in which the electrons move in their "orbits." The first (innermost) shell contains two electrons, the next shell eight, the next sixteen, and the last two. As the periodic system of the elements is built up from the lightest element, hydrogen, the formation of the innermost shell begins first, and when completed the numbers of electrons in the first four shells are two, eight, eighteen, and thirty-two, but the maximum number in each shell is not always reached before the next shell begins to be formed. For example, when formation of the fourth shell begins, the third shell contains only eight electrons instead of eighteen; it is the subsequent building up of this third shell that is intimately connected with ferromagnetism. In this shell some electrons will be spinning in one direction and others in the opposite, and these two senses of the spins may be conveniently referred to as positive and negative. The numbers on the circles show how many electrons with + and − spins are present in each shell in iron and it will be noticed that all except the third shell contain as many electrons spinning in one direction as in the opposite. The magnetic moments of the electrons in each of these shells mutually compensate one another so that the shell is magnetically neutral and does not have a permanent magnetic moment. In the third shell, however, there are five electrons with a positive spin and one with a negative so that four electron spins are unbalanced or uncompensated and there is a resultant polarization of the atom as a whole. The existence of a permanent magnetic moment for each atom obviously satisfies one of the requirements for ferromagnetism.

In the free atom the orbital motions of the electrons also contribute to the magnetic moment. When the iron atom becomes part of metallic iron the electron orbits become too firmly fixed in the solid structure to be influenced appreciably by a magnetic field. The corresponding moments do not change when the intensity of magnetization changes—this is shown by the gyromagnetic experiments discussed later—and it is supposed that the orbital moments of the electrons in various atoms neutralize one another.

In the solid structure neighboring atoms influence the motion and distribution of electrons, particularly in the third part of the third shell (3*d* shell) and the first part of the fourth shell (4*s* shell). In Fig. 10 the difference between a free atom and one that is part of a metal is illustrated. Each of the ten places for electrons in the 3*d* shell is represented by an area which is shaded if that place is occupied. The distribution corresponds in (*a*) to an isolated atom of nickel, in (*b*) to a nickel atom in a metal; in the latter situation there is *on the average* 0.6 electron per atom in the 4*s* shell (these electrons are loosely bound and are the free electrons responsible for electric conduction) and a vacancy or hole of 0.6 electron per atom in the 3*d*-shell.[6] In the 4*s* shell the number of electrons with + and with − spin are almost exactly equal, but in the 3*d* shell all of the spaces for + spin are filled. The difference between the numbers of + and − spins is equal to the net magnetic moment per atom. Experimentally the difference in the number of + spins and − spins in an atom is determined from the saturation intensity of magnetization at absolute zero. When this difference is one the atom has a moment of one Bohr magneton,

$$\mu_B = 9.2 \times 10^{-21} \text{ erg/gauss}$$

consequently the number of Bohr magnetons can be calculated from the atomic weight, A, and the density, d:

$$\text{Bohr magnetons/atom} = \beta = \frac{I_0 A}{\mu_B d}.$$

In Fig. 10 (*f*) the diagram for nickel is repeated, this time with the tops of the unfilled positions on the same level to bring out an analogy with the filling of vessels with water. Diagrams for manganese, iron, cobalt, nickel and copper are shown in parts (*c*) to (*g*). In each case the 18 electrons in closed shells are not shown. In iron the situation is somewhat different from that in nickel, neither the 3*d*+ nor the 3*d*− shell is filled. This follows from the relative constancy of the number of electrons in 4*s*, from the excess of holes in 3*d*+ over those in 3*d*− ($\beta = 2.2$), and from the total number, 26, of extra-nuclear electrons.

The distribution in space of electrons belonging to the 3*d* and 4*s* shells is known approximately [7] and is depicted in Fig. 11. In (*a*) the ordinate shows the number of electrons there are at various distances from the nucleus. The 3*d* shell is thus seen to be a rather dense ring

[6] E. C. Stoner, *Phil. Mag.*, *15*, 1018–1034 (1933); N. F. Mott, *Proc. Phys. Soc.*, *47*, 571–588 (1935); L. Pauling, *Phys. Rev.*, *54*, 899–904 (1938).

[7] Calculations were based on the equation given by J. C. Slater, *Phys. Rev.*, *36*, 57–64 (1930).

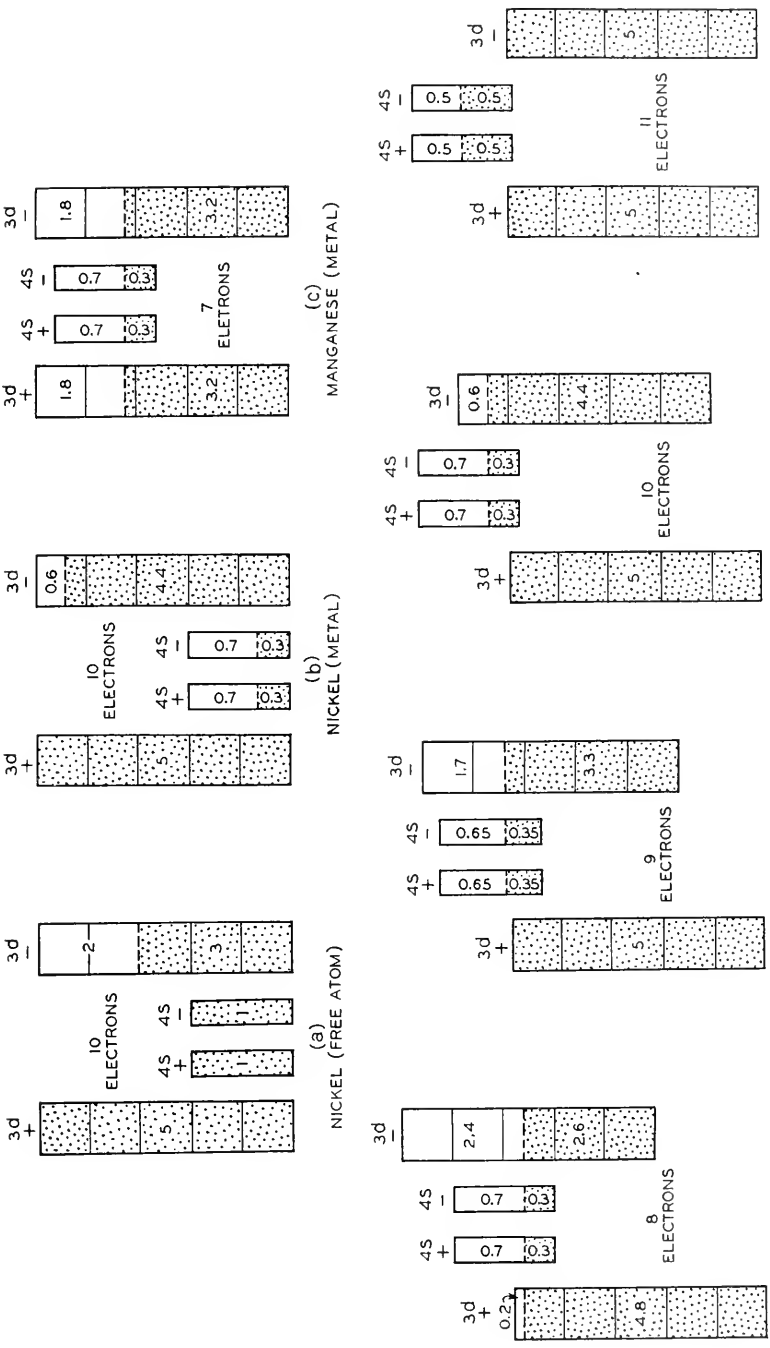


Fig. 10—The distribution of electrons among the possible electron positions in a free atom of nickel, and in manganese, iron, cobalt, nickel and copper atoms that form part of a metal.

of electrons, as contrasted with the 4*s* shell which extends farther from the nucleus, so far that in the solid the shells of neighboring atoms overlap considerably. In (*b*) the number of electrons having energy between E and $E + dE$ is plotted against the energy E; this representation is similar to that of Fig. 10 but now the squares and rectangles are replaced by the more appropriate curved surfaces. If (*b*) is turned 90° relative to (*a*) the two pairs of curves bear some resemblance to each other. This is so because the energy of binding is generally less at greater distances from the nucleus. The $3d+$ level is represented as lower in energy than the $3d-$ since one of these bands is preferred.

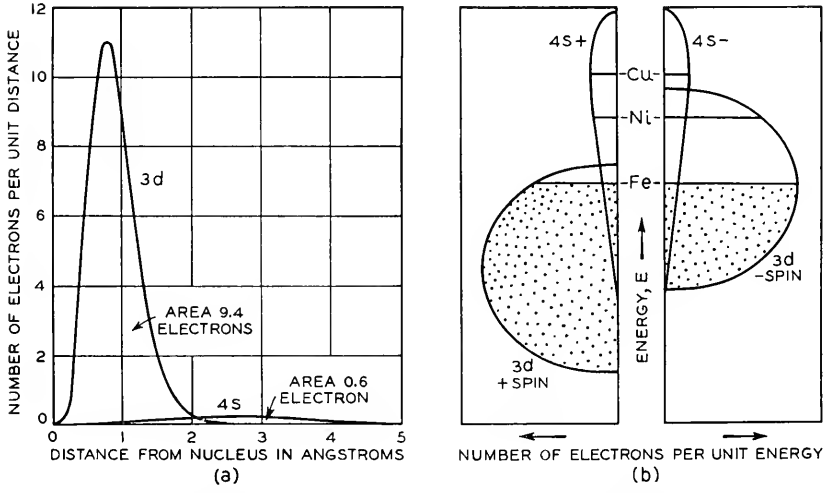


Fig. 11—The filling of electron positions in iron, and some elements near it in the periodic table. Electron positions for closed shells, containing 18 electrons, are not shown.

The area enclosed by each 3*d* curve corresponds to 5 electrons while that enclosed by the 4*s* corresponds to 2.

The line "Fe" in Fig. 11(*b*) represents the limit to which the 3*d* and 4*s* shells are filled in iron; neither $3d+$ nor $3d-$ is completely full. The lowest energy levels are filled first, and the picture is drawn so that the analogy with the filling of connected vessels with water is apparent. In cobalt and nickel the extra one and two electrons completely fill $3d+$ but not $3d-$, as indicated by the line "Ni" for nickel. Since the range of energy in the 3*d* "bands" is much greater than in the 4*s* bands the additional electrons do not alter greatly the number in 4*s*, and from the saturation intensity of nickel we estimate this number as 0.6. In copper the additional electron is sufficient to fill both 3*d* shells with one electron to spare, and this electron must go into the

4*s* shell which then becomes half full as shown by the line "Cu" as well as by (*g*) of Fig. 10. The diagram does not show changes in the relative levels of the 3*d*+ and 3*d*− bands that occur in going from one element to another; when both 3*d* bands are filled, as in copper, these levels are the same. The numbers of electrons and "holes" in metals near iron in the periodic table are given in Table I. A more

TABLE I

NUMBER OF ELECTRONS AND VACANCIES (HOLES) IN VARIOUS SHELLS IN METAL ATOMS NEAR IRON IN THE PERIODIC TABLE

Element	Number of electrons in following shells				Total	Holes in		Excess holes in 3*d*− over 3*d*+
	3*d*+	3*d*−	4*s*+	4*s*−		3*d*+	3*d*−	
Cr	2.7	2.7	0.3	0.3	6	2.3	2.3	0
Mn	3.2	3.2	0.3	0.3	7	1.8	1.8	0
Fe	4.8	2.6	0.3	0.3	8	0.2	2.4	2.22
Co	5	3.3	0.35	0.35	9	0	1.7	1.70
Ni	5	4.4	0.3	0.3	10	0	0.6	0.61
Cu	5	5	0.5	0.5	11	0	0	0

accurate determination of the form of the 3*d* and 4*s* bands for copper is given in Fig. 12, due to Slater.[8]

An especially simple and interesting illustration of the atom-model described is afforded by the alloys of nickel and copper. The substitution of one copper for one nickel atom in the lattice is equivalent to adding one electron to the alloy. This electron seeks the place of lowest energy in the alloy and finds it in the 3*d*-shell of a nickel atom rather than in the copper atom to which it originally belonged. This lowers the magnetic saturation of the alloy by one Bohr unit, since the added electron in the 3*d*− band just neutralizes the moment of one in the 3*d*+ band. Addition of more copper to nickel decreases the average moment until the empty spaces in the 3*d*− band are just full; this occurs when 60 per cent of the atoms are copper, and then the magnetic saturation at 0° K will be just zero. This is the explanation of the experimental results [9] shown in Fig. 13. There are shown also the saturation moments for other alloys of nickel; it is evident that zinc with two 4*s* electrons fills up the 3*d* band twice as fast as copper, aluminum three times as fast, silicon and tin four times and antimony five, in good accord with theory. In each of these cases the added

[8] J. C. Slater, *Phys. Rev., 49,* 537–545 (1936).

[9] V. Marian, *Ann. de Physique* (11), *7,* 459–527 (1937). Some of the data for the other alloys shown in Fig. 12 are taken from C. Sadron, *Ann. de Physique, 17,* 371–452 (1932). The interpretation of these results is due to E. C. Stoner, ref. 6.

atoms have filled up $3d$ bands, losing their more loosely bound $4s$ electrons when there are available places of lower energy. The data for palladium indicate that this element has the same number of outer

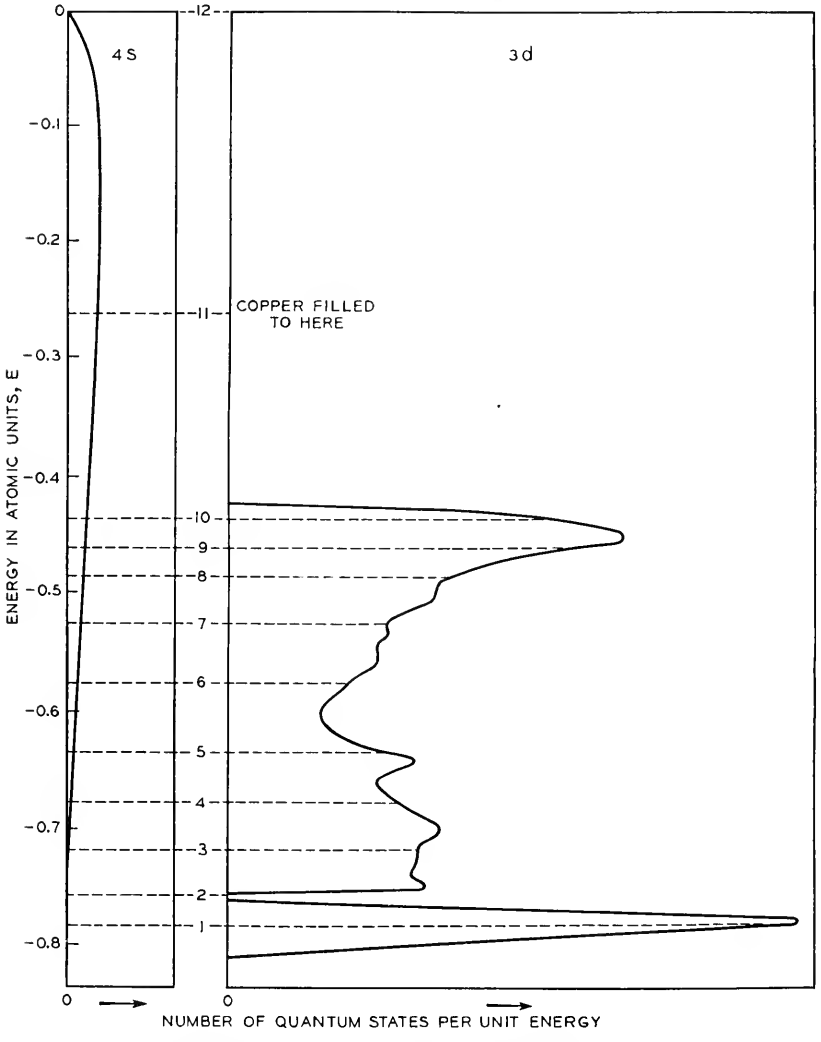


Fig. 12—Energy levels in the $3d$ and $4s$ shells in copper, according to Slater. Similar levels are believed to exist in nickel and cobalt with the levels filled to "10" and "9" respectively.

electrons as nickel; this might be expected since palladium lies directly below nickel in the periodic table. When the similar but heavier platinum is added to nickel, the decrease in average atomic moment

indicates that some of the outer electrons of platinum go into the 3*d* band of nickel, but that they do not fill this level as rapidly as the outer electrons of copper do when this element is added.

Electron shells that are completely filled behave more like hard elastic spheres than those which are only partially filled. In solid copper with

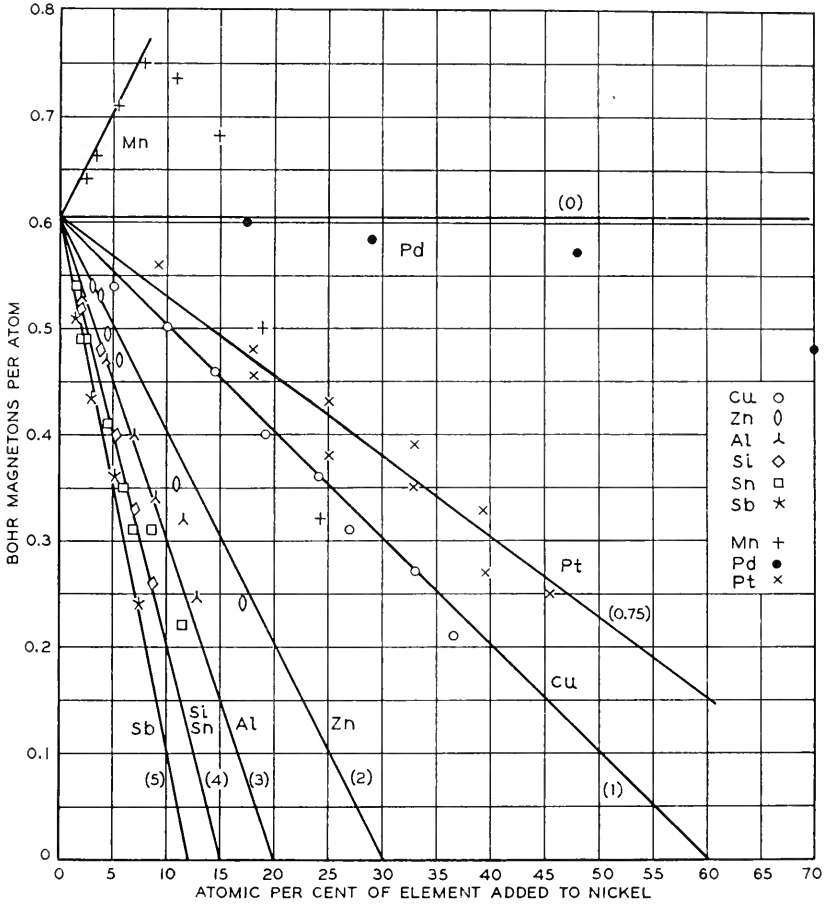


Fig. 13—The saturation magnetization of nickel decreases upon the addition of other elements having 1, 2, 3, . . . electrons in the outermost shell.

a complete 3*d* shell and a 4*s* shell just begun, the 4*s* electrons "overlap" those of neighboring atoms so much that their connection with any one atom is lost; the 3*d* shells on the other hand have very little overlap with neighboring atoms. In the ferromagnetic metals the 3*d* shells are incomplete and the overlap is greater than in copper; this affects the interaction responsible for the Weiss molecular field, now to be

discussed. But copper would not be ferromagnetic even if the interaction were large, because the completed shell means that the saturation magnetization is zero; in reality copper is diamagnetic.

A more detailed discussion of the atomic structure of metals, particularly of the band picture of the ferromagnetic metals, is given in a recent article in this journal by W. Shockley.[10]

Interpretation of the Molecular Field

It was shown by Heisenberg [11] that the molecular field can be explained in terms of the quantum mechanical forces of exchange acting between electrons in neighboring atoms. Imagine two atoms some distance apart, each atom having a magnetic moment of one Bohr magneton due to the spin moment of one electron. A force of interaction has been shown to exist between them, in addition to the better-known electrostatic and (much weaker) magnetic forces. It is known that, as one would expect, such forces are negligible when the atoms are two or three times as far apart as they are in crystals. It is supposed also, on the basis of calculations by Bethe,[12] that as two atoms are brought near to each other from a distance these forces cause the electron spins in the two atoms to become parallel (positive interaction). As the atoms are brought nearer together the spin-moments are held parallel more firmly until at a certain distance the force diminishes and then becomes zero, and with still closer approach the spins set themselves antiparallel with relatively strong forces (negative interaction). In the curve of Fig. 14 the energies corresponding to these forces are shown as a function of the distances between atoms.

Bethe's curve was drawn originally for atoms with definite shell radii and varying internuclei distances. It may equally well be used for a series of elements if we take account of the different radii of the shell in which the magnetic moment resides. The criterion of overlapping or interaction for the metals of the iron group is the radius, R, of the atom (half the internuclear distance in the crystal) divided by the radius, r, of the $3d$ shell. In Fig. 14 this ratio R/r has been used as abscissa and the elements iron, cobalt and nickel have been given appropriate positions on the curve. The recently discovered ferromagnetism of gadolinium [13] is apparently associated with a large R/r and small interaction, as compared to nickel. It is placed on the curve accordingly. Slater [7] has shown that the ratio R/r is larger in the

[10] W. Shockley, *Bell System Technical Journal*, *18*, 645–723 (1939).
[11] W. Heisenberg, *Z. f. Physik*, *49*, 619–636 (1928).
[12] H. Bethe, *Handbuch der Physik*, *24*, pt. 2, 595–598 (1933).
[13] G. Urbain, P. Weiss, and F. Trombe, *Compt. Rend.*, *200*, 2132–2134 (1935).

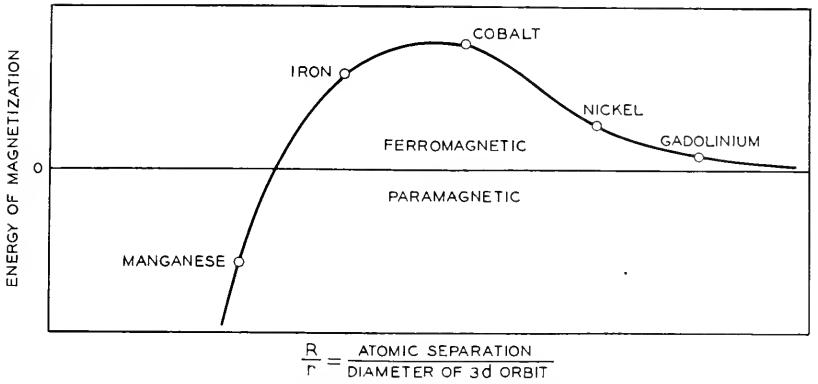


Fig. 14—Bethe's curve relating the energy of magnetization to the distance between atom-centers, with a fixed diameter of the unfilled inner shell that has the magnetic moment.

ferromagnetic elements than in other elements having incomplete inner shells, and that the point at which the curve crosses from the non-ferromagnetic to the ferromagnetic region is near $R/r = 1.5$. Values of $2R$, $2r$ and R/r, as calculated by Slater for some of the elements with incomplete inner shells, are given in Table II.

TABLE II

INTERNUCLEAR DISTANCES ($2R$) AND DIAMETERS ($2r$) OF INCOMPLETE INNER SHELLS OF SOME ATOMS, IN ANGSTROMS

	Atom $2R$	Inner Shell $2r$	Ratio R/r	Incomplete Inner Shell	Curie Temperature θ, °K.
Mn	2.52	1.71	1.47	3*d*	
Fe	2.50	1.58	1.63	3*d*	1040
Co	2.51	1.38	1.82	3*d*	1400
Ni	2.50	1.27	1.97	3*d*	630
Cu-Mn	2.58	1.44	1.79	3*d*	600
Mo	2.72	2.94	0.92	4*d*	
Ru	2.64	2.33	1.13	4*d*	
Rh	2.70	2.11	1.28	4*d*	
Pd	2.73	1.93	1.41	4*d*	
Gd*	3.35	1.08	3.1	4*f*	290
W	2.73	3.44	0.79	5*d*	
Os	2.71	2.72	1.02	5*d*	
Ir	2.70	2.47	1.09	5*d*	
Pt	2.77	2.25	1.23	5*d*	

* Calculated using Slater's formula.

The energy of interaction, J—the positive ordinate of Fig. 14—can be estimated from the value of the Curie temperature, θ, in a manner suggested by Stoner.[14]

Let $2J$ be the difference in the energy of interaction between two atoms when their moments are respectively parallel and antiparallel. The total energy of these two atoms is therefore

$$2E = 2E_0 \pm J$$

where E_0 is the energy of an isolated atom. The negative sign applies when the spins are parallel, the positive when they are antiparallel. Imagine a crystal in which each atom of moment μ_A is surrounded at equal distances by z other atoms of which x have their spins parallel and y antiparallel. Then turning one atom from the parallel to antiparallel position produces a change of $(y - x)$ in the number of parallel pairs and $(x - y)$ in the number of antiparallel pairs and, therefore, requires an energy

$$\epsilon = 2J(x - y). \tag{5}$$

Since in each atom the moment must be parallel or antiparallel to the field, the magnetization of the material as a whole will depend on the average value of $x - y$:

$$I/I_0 = \overline{(x - y)}/z. \tag{6}$$

According to Boltzmann's equation an atom will have the following probabilities of being parallel and antiparallel

$$P_p = 1/[1 + \exp(-\epsilon/kT)]$$

$$P_a = \exp(-\epsilon/kT)/[1 + \exp(-\epsilon/kT)].$$

Since all atoms behave in the same way on the average $\bar{x}$ and $\bar{y}$ must be zP_p and zP_a. Hence we have

$$I/I_0 = (\bar{x} - \bar{y})/z = P_p - P_a = \tanh(\epsilon/2kT)$$

or using (5) and (6)

$$\frac{I}{I_0} = \tanh\left(\frac{zJ}{kT}\frac{I}{I_0}\right).$$

Comparing this with the modified Weiss equation, Eq. (4),

$$\frac{I}{I_0} = \tanh\frac{\mu_A N I}{kT} = \tanh\frac{I/I_0}{T/\theta}$$

we have J in terms of the molecular field constant or the Curie temperature:

$$J = \mu_A N I_0/z = k\theta/z.$$

For iron, $z = 8$, $J = k\theta/8 = 1.8 \times 10^{-14}$ erg or 0.01 electron volt.

This derivation indicates that J is proportional to θ, and that the constant of proportionality depends on the number of nearest neighbors. The number of neighbors has not been taken into account in the following discussion of Fig. 14.

The interaction curve is substantiated in a qualitative manner by the observed variation of the Curie points of the iron-nickel alloys.[15]

[14] E. C. Stoner, *Phil. Mag.*, *10*, 27–48 (1930). Stoner's original work appears to have been in error by a factor of two; the modified treatment given here is due to W. Shockley and follows closely the method employed in dealing with order and disorder in alloys (see e.g. Eqs. 1.11, 1.12, 2.2 and 2.16 in the article by F. C. Nix and W. Shockley, *Rev. Mod. Phys. 10*, 1–71 (1938)).

[15] Summarized by J. S. Marsh, *Alloys of Iron and Nickel*, v. 1, pp. 45 and 142, McGraw-Hill, New York (1938).

shown in Fig. 15. The maximum in the curve near 70 per cent nickel apparently corresponds to the maximum of the interaction curve of Fig. 14. In alloys of higher nickel content the curve indicates that the Curie point should be increased if the material is compressed. The opposite should be true of the face-centered alloys having less than this amount of nickel. These contentions are borne out by the fact

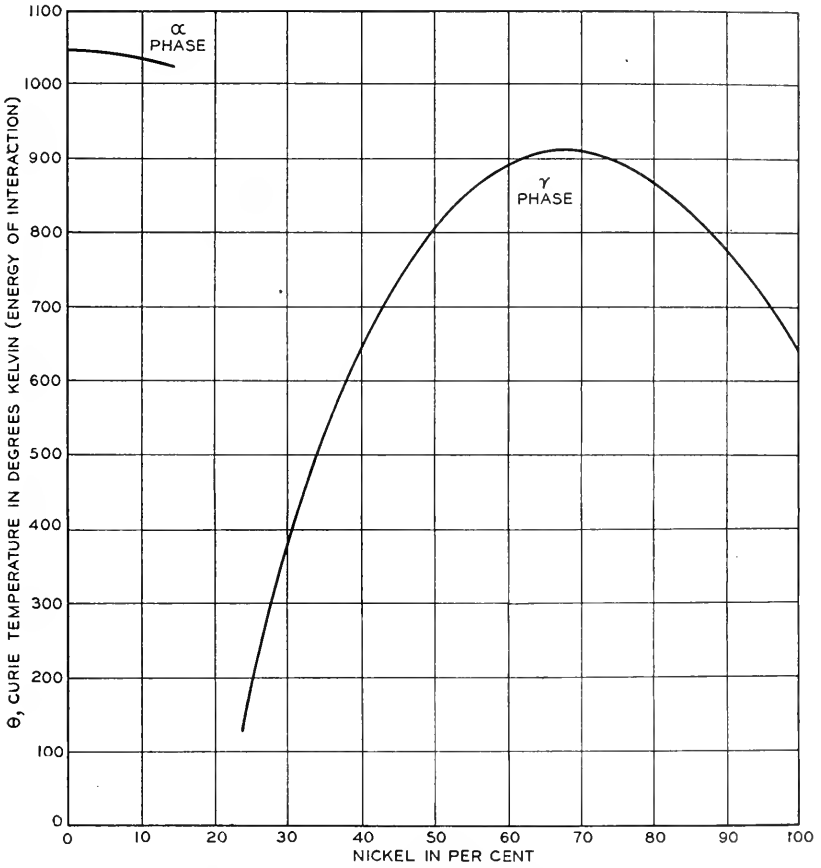


Fig. 15—The Curie temperatures for iron-nickel alloys, showing a maximum corresponding to the maximum of Bethe's curve of Fig. 14.

that under a hydrostatic pressure of 10,000 atmospheres the 30 per cent nickel alloy becomes practically non-ferromagnetic[16] at room temperature (permeability is independent of field-strength and equal to 1.7). On the other hand the effect of the pressure on the phase equilibrium is unknown so that the data might be explained also by a change of phase

[16] R. L. Steinberger, *Physics*, *4*, 153–161 (1933).

brought about by the change of pressure. More data are needed to clarify the theory.

There is an anomalous expansion of the high nickel alloys (due to loss of magnetism) as the alloy is heated through the Curie point, a contraction of the low nickel alloys, and no anomaly in the alloys having about 70 per cent nickel, as indicated by the data [15] of Fig. 16 on the expansion of these alloys in the range of temperatures including the Curie points. Bethe's curve represents the change of interaction energy with volume as a material is expanded or contracted, and it is to be expected that there will be a reciprocal effect, a change in volume

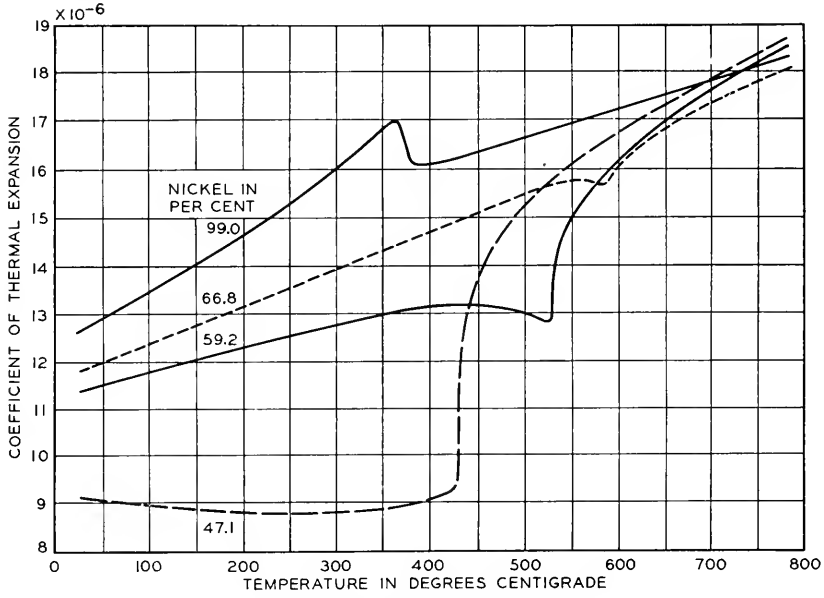


Fig. 16—The expansion coefficient of some iron-nickel alloys, showing the magnetic anomoly and its change in sign at about 70 percent nickel.

as the material passes through the Curie temperature. More careful consideration of the theory [10] shows that the effect to be expected does agree in sign with experiment. Also the disappearance of the anomalous expansion occurs as expected at the same composition as the maximum Curie temperature.

Iron lies to the left of the maximum, as indicated by its expansion curve. Calculations by Kornetski [17] indicate that the interaction energy doubles for a 2 per cent increase in lattice constant. The behavior of cobalt, nickel, and alloys of cobalt-nickel and of nickel-

[17] M. Kornetzki, *Z. f. Physik*, *98*, 289–313 (1935).

copper, indicates that all of these substances should lie to the right of the maximum. It should be expected that iron-cobalt, like iron-nickel, alloys should lie in the region including the maximum. This is not observed; instead, the Curie point continually decreases as iron or nickel is added to cobalt—in this case, however, the change of Curie point with composition is obscured by a change of phase so that no easy test of the theory is possible.

Sizes of Domains and Widths of Domain Boundaries

The quantum mechanical interaction in ferromagnetic materials tends to make the magnetic moments of neighboring atoms parallel. One infers that the whole ferromagnetic specimen should be one single large domain; nevertheless in actual fact the parallelism extends over much smaller regions only. This behavior is attributed to strains, crystal boundaries, temperature vibrations, impurities, etc. The fact that a specimen can be demagnetized so that no residual magnetization can be observed by ordinary means, indicates that the domains are not larger than microscopic in size; while the occurrence of heat effects at the Curie point shows that the magnetic unit is larger than a single atom.

A direct measure of the *domain size* is obtained from experiments on the Barkhausen effect;[18] the volume is found to be of the order of 10^{-9} cm.^{3}, so that it contains about 10^{14} atoms. The Barkhausen data give little information concerning the shape of a domain, but this has been made evident by the powder patterns of Bitter and others;[1] a typical domain is long and slender, either rod-like or plate-like with a thickness of the order of one micron (10^{-4} cm.) and a length of perhaps 10 microns. The volume thus agrees with the results of the Barkhausen effect within one or two orders of magnitude. No explanation has been given for the occurrence of domains of this particular size.

There is at present no experimental evidence regarding the nature of the *transition region* between domains, and in the schematic Fig. 3 no transition region is shown. It is believed that the boundary will not be sharp on an atomic scale, but will be spread over a region a considerable number of atoms wide. Calculation indicates that less energy is required if the electron spins change direction gradually from atom to atom as indicated in Fig. 17. The spreading of the transition region over many atoms instead of over one, is analogous to the separation of similar electric charges; the mutual forces tend to spread them over a region as large as possible and they are held together

[18] R. M. Bozorth and J. F. Dillinger, *Phys. Rev.*, *35*, 733–752 (1930).

only by some other forces such as those imposed by an electric field. The expression for the energy of interaction in a boundary layer has been derived by Bloch,[19] and found to be *inversely proportional* to the thickness of the layer,

$$\gamma_0 = \frac{k\theta}{a} \cdot \frac{1}{\delta}$$

per unit area of boundary. Here k is Boltzmann's constant, θ the Curie temperature, a the distance between atoms and δ the thickness of the layer; since the layer has no sharp limit, δ is measured between

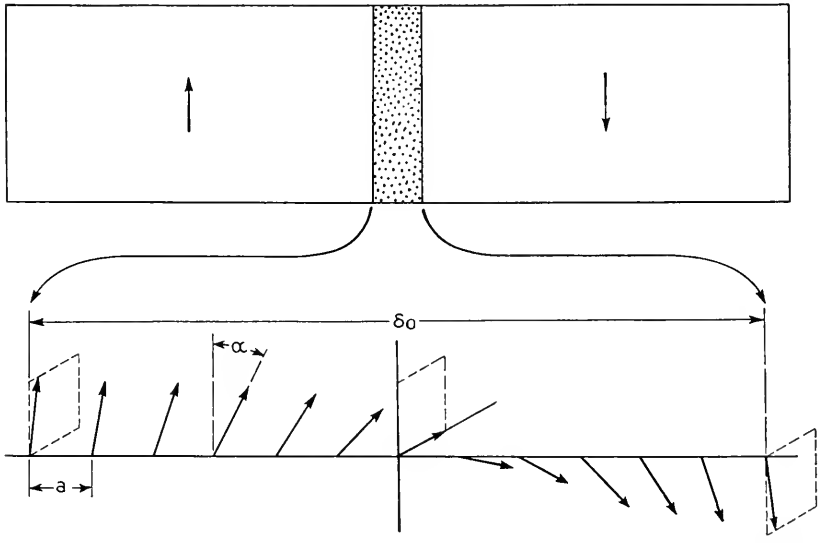


Fig. 17—The nature of the domain boundary. The transition region between two domains is believed to be about 1000 atom diameters thick.

points at which the spins are inclined at a certain small angle (α almost 0° or 180° as shown) to the spins in the middle of the domains.

The forces of interaction are opposed by forces (e.g. of crystal anisotropy or strain) which correspond to fixed values of energy *per unit volume.* This opposing energy is thus *directly proportional* to the thickness of the boundary,

$$\gamma_1 = C\delta.$$

The minimum energy occurs when

$$\frac{d}{d\delta}(\gamma_0 + \gamma_1) = 0$$

[19] F. Bloch, *Z. f. Physik*, *74*, 295–335 (1932). See also the more recent article by H. Kersten in "Probleme der Technischen Magnetizurungskurve" (R. Becker, ed.) 42–72, Springer, Berlin (1938).

or

$$\delta = \sqrt{k\theta/(aC)} = \delta_0.$$

In iron and similar materials free from any considerable strain the value of C is determined by the crystal anisotropy and is about 10^5 ergs/cm.^{3}, $\theta \approx 10^3$ °K, $a \approx 10^{-8}$ cm. and the thickness of the boundary layer comes out to be about 1000 atom diameters. This value, probably correct as to order of magnitude, indicates that the volume of the domain proper is much larger than that of the boundary or transition region.

At present it is not clear why application of an indefinitely small field will not cause continual progression of the 180° boundary in one

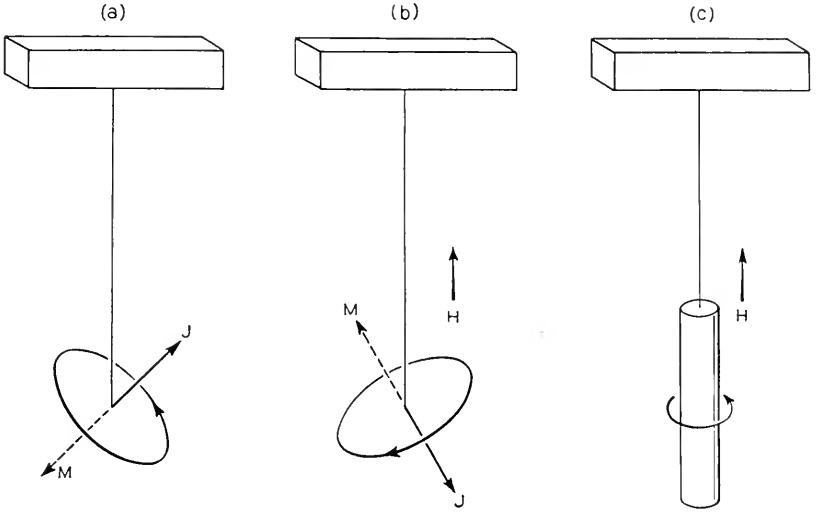


Fig. 18—The magnetic moment, M, and the moment of momentum, J, of an electron in its orbit about the nucleus. A change in one moment entails a change in the other, the (gyromagnetic) ratio remaining constant.

direction so that one domain will disappear completely. The reason for the non-occurrence of this progression except under certain circumstances is probably connected with the existence of strain gradients.

Gyromagnetic Effect

In the discussion of the structure of ferromagnetic atoms, use was made of the concept of electron spin. This section will review the evidence for the existence of this spin, its experimental determination, and its relation to magnetic phenomena.

Theory. In principle, the ratio of the moment of momentum to magnetic moment may be determined as illustrated in Fig. 18. An

electron of mass m and negative charge e revolves about its nucleus f times per second in an orbit of radius r. The magnetic moment due to the circulating current is at right angles to the plane of the orbit and is

$$M_0 = ef\pi r^2/c.$$

The moment of momentum is in the opposite direction and its magnitude is

$$J_0 = 2mf\pi r^2.$$

The ratio of the moments for this orbital motion is then

$$\rho_0 = \frac{J_0}{M_0} = \frac{2mc}{e}.$$

Imagine now that the atom is suspended in space by a fibre as shown in (*a*). If a strong magnetic field is applied the vector M representing the magnetic moment will rotate around the axis of the suspension, and J will rotate with it, as the electron precesses. As long as there is no external force or friction the angle between M and the axis will not change but only the speed of its rotation will vary. On the other hand if there is an exchange of energy with other atoms as there is in a real material subject to temperature agitation, then M approaches parallelism with H as shown in (*b*), and the components of M and J parallel to the axis change in the same ratio. Consequently the change in the magnetic moment about the axis of the suspension may be said to cause a change in the moment of momentum about the same axis. As a result of the concerted action of all of the atoms composing a rod (*c*), and the recoil of the rod as a whole, the suspension is subject to a torque equal to the (negative) time rate of change of the moments of momentum of the constituent electrons:

$$L = -dJ/dt.$$

Thus a rod suspended as shown in Fig. 17 (*c*) may be magnetized a known amount, its resulting rotation measured, and its gyromagnetic ratio M/J so determined. The same ratio may be found also by measuring the magnetic moment M caused by rotating a similar rod with a known angular acceleration; this is the inverse effect.

The existence of a magnetic moment and an angular momentum associated with an electron apart from its orbital motion in the atom, was postulated in 1925 by Goudsmit and Uhlenbeck [20] primarily to explain the structure of atomic spectra. The magnetic moment

[20] S. Goudsmit and G. E. Uhlenbeck, *Nature*, *117*, 264–265 (1926).

assigned to this spin of the electron about its own center was equal to one Bohr magneton which by definition is that of the smallest electron orbit on the Bohr theory.

$$\mu_B = \frac{eh}{4\pi mc} = 9.2 \times 10^{-21} \text{ erg/gauss.}$$

The unit of angular momentum was taken as *one-half* of that for the smallest Bohr orbit or as

$$J_s = \frac{h}{4\pi}.$$

The ratio for the spin motion, denoted by ρ_s, is

$$\rho_s = \frac{J_s}{\mu_B} = \frac{mc}{e} = \frac{\rho_0}{2},$$

and is thus twice the gyromagnetic ratio for the orbital motion of the electron. Dirac has shown that these results are consequences of relativistic quantum theory.

In general the ratio M/J is

$$\rho = \frac{mc}{e} \cdot \frac{2}{g}$$

where g is known as the Landé splitting factor. For spin moment, $g = 2$; for orbital moment, $g = 1$. When the moment of an atom is the resultant of finite spin and orbital moments, g may be found in terms of the quantum numbers, s and l, expressing the angular momenta of the spin and orbital components:

$$g = 3/2 + \frac{s(s+1) - l(l+1)}{2j(j+1)}.$$

Here s may have any of the half-integral values 0, 1/2, 1, 3/2, $\cdots$ and l any of the integral values 0, 1, 2 $\cdots$, while the number, j, representing the angular momentum of the resultant may be any positive number equal to the sum or difference of s and l. (The actual value of the resultant angular momentum is

$$J = \frac{h}{2\pi}\sqrt{j(j+1)},$$

and that of the magnetic moment is

$$M = \frac{eh}{4\pi mc} \cdot g\sqrt{j(j+1)},$$

but the components parallel to the applied field are $jh/2\pi$ and $gjeh/(4\pi mc)$, respectively.) For some values of s, l and j, e.g. 4, 2 and 2, g is greater than 2, and for some values it is less than 1.

The sign as well as the magnitude of the rotation is of importance. All experiments are consistent with the idea that the magnetic moment is due to the spinning or circulation of negative electrons rather than of positive charges.

The results to be described below show that in ferromagnetic materials generally the value of g has nearly the value two and not at all the value one, so we conclude that ferromagnetic processes are concerned primarily with the spins of the electrons and not their orbital motions. When a *change in magnetization* takes place we therefore attribute it to a change in the *direction of spin* of some of the electrons, and believe that the orientations of the orbits are disturbed but slightly. This change is illustrated in Fig. 19. In some

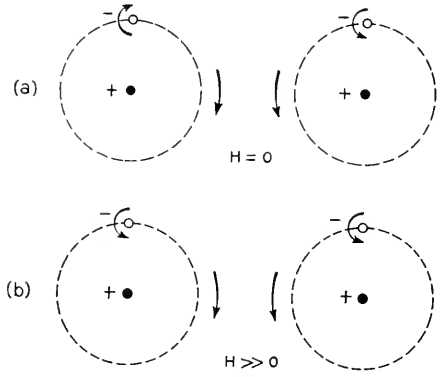


Fig. 19—In the common ferromagnetic materials a change in magnetization is effected by a change in the direction of electron spin, not in the direction of motion of the electron in its orbit.

paramagnetic materials, on the other hand, the reorientation of orbits plays an important part.

Gyromagnetic Experiments. The first gyromagnetic experiment to be performed successfully was magnetization by rotation. After an unsuccessful trial by Perry [21] in 1890, the experiment was considered independently in 1909 by Barnett [22] who in 1914 obtained the result, then inexplicable, that g was approximately twice the classical value one. Richardson,[23] in 1907, was the first to propose rotation by

[21] J. Perry, as quoted by Barnett, ref. 27.

[22] S. J. Barnett, *Science, 30,* 413 (1909); *Phys. Rev., 6,* 239–270 (1915). An accidental error in the calculation of the results was corrected in *Jour. Wash. Acad. Sci., 11,* 162 (1921). Magnetization by rotation.

[23] O. W. Richardson, *Phys. Rev., 26,* 248–253 (1908).

magnetization, and Einstein and de Haas [24] performed the experiment in 1915. It was repeated in 1918 by Stewart [25] who for the first time obtained a result consistent with Barnett's, and has been confirmed since by a number of others.

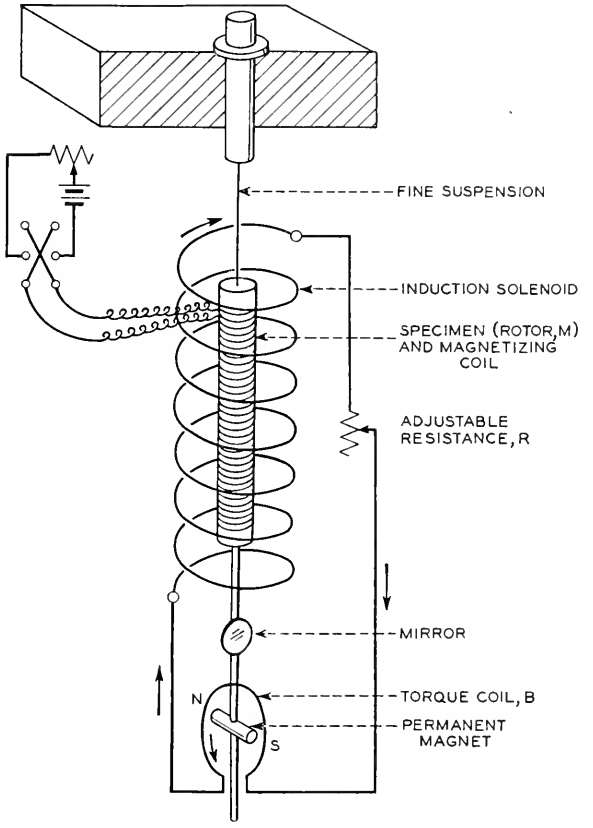


Fig. 20—Schematic diagram of the method of determining the gyromagnetic ratio.

In recent years the method most often used (rotation by magnetization) is that due to Sucksmith and Bates.[26] As modified by Barnett,[27] it is shown diagrammatically in Fig. 20. A rod of the material under

[24] A. Einstein and W. J. de Haas, *Verh. d. D. Phys. Ges.*, *17*, 152–170 (1915); *18*, 173–177 (1916); *18*, 423–443 (1916).

[25] J. Q. Stewart, *Phys. Rev.*, *11*, 100–120 (1918).

[26] W. Sucksmith and L. F. Bates, *Proc. Roy. Soc.*, *104A*, 499–511 (1923). W. Sucksmith, *Proc. Roy. Soc.*, *108A*, 638–642 (1925).

[27] S. J. Barnett, *Rev. Mod. Phys.*, *7*, 129–166 (1935). This article and the one in *Phys. Zeits.*, *35*, 203–205 (1934) give a good account of the history, methods and results to date.

investigation (the "rotor," M) is wound with a magnetizing coil and suspended by a fine quartz fibre in a second (induction) coil A. The leads from the latter are connected in series with an adjustable resistance R and a third coil B, inside of which is a small permanent magnet (moment m) mounted below the rotor and connected rigidly to it. A change in the moment of the rotor is produced by changing the current in the magnetizing coil. This causes a gyromagnetic rotation of the rotor and at the same time induces a voltage in coils A and B. R is adjusted so that the current flowing is of such strength that the field produced by it in B acts on the permanent magnet to annul the gyromagnetic torque of the rotor. The magnetizing current is alternated with a period equal to the natural period of rotation of the rotor assembly and the final deflection δ noted for various values of R. R is plotted against δ and its value, R_0, determined for zero deflection by interpolation.

Let

$$L_A = -dJ/dt$$

be the torque due to the gyromagnetic effect. The current induced in coils A and B by a change in the moment M of the rotor is

$$i = E/R = (dM/dt)(K_A/R),$$

where K_A is a constant of coil A. This current produces a torque on the magnet m in B:

$$L_B = miK_B,$$

K_B being a constant of coil B. When $R = R_0$, $L_A = -L_B$ and

$$\rho = \frac{dJ}{dM} = \frac{mK_AK_B}{R_0}.$$

The value of ρ is calculated by this formula after finding the values of the coil constants, the resistance R_0 and the moment of the permanent magnet. Barnett has taken great care to eliminate various errors caused mainly by the presence of undesirable fields such as the earth's and by asymmetry and magnetostriction of the rotor.

Experimental Values of g

The results of gyromagnetic experiments are given preferably in terms of g:

$$g = (M/J)(2mc/e),$$

and are collected in Table III. Here a g-value of two means that

TABLE III

VALUES OF g FOR SOME FERROMAGNETIC SUBSTANCES ACCORDING TO VARIOUS AUTHORS

Gyromagnetic ratio $\rho = (mc/e)(2/g)$

Substance	B[22] 1915	E & d.H[24] 1915–6	S[25] 1918	B & B[28] 1917–25	B[29] 1919	A[30] 1920	C & B[31] 1923	S & B[26] 1923–5	B[32] 1931–4	C & S[33] 1932–3	C[34] 1932–5
Iron	2.1	1	2.0	1.91	1.89	2.1	1.99	1.99	1.94	2.01	—
Cobalt	—	—	2.1	1.83	—	—	—	1.94	1.82	—	—
Nickel	—	—	—	1.96	1.75	—	1.98	2.00	1.90	—	—
Fe-Co (34% Co)	—	—	—	1.88	—	—	—	—	1.98	—	—
Fe-Ni (25% Ni)	—	—	—	1.97	—	—	—	—	1.97	—	—
Fe-Ni (75 to 80% Ni)	—	—	—	1.91	—	—	—	—	1.92	—	—
Co-Ni (54% Co)	—	—	—	1.86	—	—	—	—	1.84	—	—
Co-Cu (92% Co)	—	—	—	—	—	—	—	—	1.87	—	—
Fe-C (steel)	—	—	—	1.91	—	—	—	—	—	—	—
Mn-Al-Cu (Heusler alloy)	—	—	—	1.96	—	—	—	2.00	—	—	—
Fe_2O_3	—	—	—	—	—	—	—	—	—	—	1.96
Fe_3O_4	—	—	—	—	—	—	—	2.02	—	—	1.96
$NiFe_2O_4$	—	—	—	—	—	—	—	—	—	—	1.94
$CuFe_2O_4$	—	—	—	—	—	—	—	—	—	—	1.94
$MnFe_2O_4$	—	—	—	—	—	—	—	—	—	—	1.94
$Zn_2Fe_6O_{11}$	—	—	—	—	—	—	—	—	—	—	1.92
FeS	—	—	—	—	—	—	—	—	—	0.63	—

[22, 24, 25] Loc. cit.

[28] S. J. and L. J. H. Barnett, *Proc. Am. Acad.*, *60*, 127–216 (1925). Magnetization by rotation.

[29] E. Beck, *Ann. d. Physik*, *60*, 109–148 (1919).

[30] G. Arvidson, *Phys. Zeit.*, *21*, 88–91 (1920).

[31] A. P. Chattock and L. F. Bates, *Phil. Trans. Roy. Soc.*, *223A*, 257–288 (1922).

[32] S. J. Barnett, *Proc. Am. Acad.*, *66*, 274–348 (1931); *69*, 119–135 (1934).

[33] F. Coeterier and P. Scherrer, *Helv. Phys. Acta*, *5*, 217–223 (1932). Pyrrhotin, F. Coeterier, *Helv. Phys. Acta*, *8*, 522–564 (1935).

[34] D. P. Ray Chandhuri, *Indian J. Phys.*, *9*, 383–414 (1935).

electron spin only is operative; the ratio would be one if change in orbit orientation were the only effect. The apparent slight difference of most of the values from two, indicates that there is some small but definite change in orbit-orientation in ferromagnetic materials when they are magnetized. In the weakly ferromagnetic pyrrhotite (FeS) the experimental value 0.63 is in harmony with the theoretical value, 0.67, for a possible state of the iron atom ($s = -1/2$, $l = 2$, $j = 3/2$) in which orbital moment is of importance.

Gyromagnetic ratios for paramagnetic materials have been determined by Sucksmith [35] and are given in Table IV. The departures

TABLE IV

VALUES OF g FOR SOME PARAMAGNETIC SUBSTANCES (SUCKSMITH)

Substance	g-value		Substance	g-value	
	obs.	calc.		obs.	calc.
Nd_2O_3	0.78	0.76	$FeSO_4$	1.89	<2.00
Gd_2O_3	2.12	2.00	$CoCl_2$-$CoSO_4$	1.54	<2.00
Dy_2O_3	1.36	1.33	$CrCl_2$	1.95	<2.00
Eu_2O_3	>4.5	6.56	$MnCO_3$-$MnSO_4$	1.99	2.00
			Ni-Cu(56% Ni)	1.9	2.00

from the values 1 and 2 show that changes in both spin and orbital moments occur during magnetization. In the last column are added theoretical values deduced from spectroscopic data.

SUMMARY

In this paper the author has discussed some of the difficulties encountered in the interpretation of the fundamental phenomena of ferromagnetism, and some of the successes that have been attained by applying our recent knowledge of the structure of atoms in solids. The difficulties are large because the atomic forces controlling the magnetism are small compared to those that hold the atoms together in a solid. The successes have come largely as a result of the quantum theory which has explained, mainly in a qualitative way, many of the phenomena previously correlated by the empirical Weiss theory of the molecular field.

In some ways magnetic studies have aided materially in clarifying our picture of the atom; this has been brought out in a discussion of

[35] W. Sucksmith, *Proc. Roy. Soc.*, *133A*, 179–188 (1931); *135A*, 276–281 (1932); *Helv. Phys. Acta*, *8*, 205–210 (1935).

(1) the atomic magnetic moment (determined from the saturation magnetization at 0° K), which gives directly the numbers of electrons in certain shells in the atom, and (2) the gyromagnetic effect, experiments on which give results characteristic of an electron spinning about an axis passing through its center.

ACKNOWLEDGMENT

I take pleasure in acknowledging the benefit of many discussions with Dr. W. Shockley and of the criticism of the manuscript given by Dr. K. K. Darrow and Dr. R. W. King.

Contact Phenomena in Telephone Switching Circuits*

By A. M. CURTIS

The phenomena occurring at the closing and opening of contacts carrying weak currents have been investigated by means which include a study of the high-frequency transient voltages and currents. These influence the erosion in a complex manner which varies with contact materials, surface conditions and surrounding atmosphere. Three principal classes of effect have been distinguished. These are: (1) Disruptive sparkovers initiating a series of metallic arcs lasting less than a microsecond each; (2) A nitrogen gas glow discharge at about 300 volts, preceded by a brief group of disruptive sparkovers; (3) High field breakdowns due to cold point discharges which cause transient metallic closures of approaching contacts and similar transient reclosures of separating contacts.

THE operation of a telephone system depends on the proper performance of many millions of electrical contacts, a large proportion of which are in relays. The relays must be designed for a life during which they operate from as few as five thousand to as many as four hundred million times. Although the nominal currents and voltages carried by the contacts are rather low, the large number of operations may cause erosion which in a very small percentage of cases leads to failures to close or open the circuit. The difficulties caused by even very rare failures make the control of contact erosion a problem of major importance for the telephone companies.

Research and development work on contacts has of course been carried on continuously since very early in the development of the telephone system. The aim is to design contacts to have a life at least equal to that of the apparatus of which they form a part and to require a minimum of maintenance. Although this aim has in general been successfully met there have been some cases in which the contacts have worn out too rapidly.

Although it had long been realized that contact operation necessarily involved the generation of high-frequency transients, there was at first no apparatus available which would permit these transients to be studied. The Dufour oscillograph was for a long time the only instrument which covered the range of frequencies involved. It was em-

* Presented at Winter Convention of A. I. E. E., New York, N. Y., January 22–26, 1940.

ployed as early as 1926 in studies of contact sparking but it was very cumbersome in use, often introduced artificial conditions into the circuit of the contacts, and progress with its use was necessarily very slow. During the past few years rapid advances have been made in the development of glass envelope cathode ray oscillograph tubes. By employing the latest types of tubes, and combining them when necessary with wide band high-frequency amplifiers and with circuits which permit synchronization of the tube sweep circuit with the contact operation, it has been possible to make thousands of observations in the time originally taken by a single oscillogram, and to cover the entire range of currents, voltages, and frequencies involved. We now have available means which will permit the visual observation of transient voltages at frequencies as high as 400 megacycles per second, and transient currents with components reaching 20 megacycles per second. Single pulses lasting a small fraction of a microsecond, and complex transients containing components as high as 5 megacycles, can be clearly resolved and photographed while the envelopes of still higher frequencies can be recorded.

In order to study the transients at contacts operating at 50 volts and steady currents under one ampere, in common types of telephone circuits, voltages as high as 2000 and currents reaching 20 amperes must be within the range of the apparatus. A detailed description of the apparatus will not be attempted in this article, but the results of observations made with it and photographs of the more significant transient components will be presented.

Study of the currents requires an amplifier as an impedance matching device and some circuit conditions make a shielded input transformer necessary. An input impedance of from 0.4 to 2 ohms, a voltage gain of about seventy-five times, and a substantially flat characteristic of output versus input from 20 kilocycles to 20 megacycles are usually employed. Lower frequencies may be observed with other amplifiers and the range from zero to 10,000 c.p.s. is studied by means of the "Rapid Record" oscillograph.

With earlier cathode ray tubes, beam currents of 40 microamperes at 5000 volts were employed. The latest tubes give a beam current of about one milliampere at this voltage. A Leica camera with an F1.5 Xenon lens and ultra speed panchromatic film has been used in most of the photographic work. The photography is complicated by the presence in a single transient photograph of some components in which the beam speed may be a thousand times as fast as it is in others. However, beam speeds in excess of 200 kilometers a second are photographed, and a continuous sine wave of 5 megacycles frequency may be

clearly resolved on a single transit. Sweep speeds which permit resolution of much higher frequencies are employed for visual observation, where the transient component being studied can be found by frequent repetition of the contact operation. As the occurrence of a particular component varies in time of its position in the entire transient, very high sweep speeds are impractical for photography as a prohibitively large proportion of exposures would be blanks. A sweep speed of about 15 kilometers per second is about as high as is useful except in some special cases.

We may commence the discussion by setting up what appears to be a very simple circuit (Fig. 1), a pair of contacts, one of which is con-

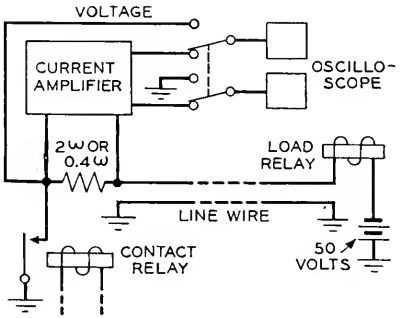


Fig. 1—Typical relay and contact circuit.

nected by a length of wire to a relay winding, which is in turn connected to one pole of a 50-volt battery. The mate contact and the other pole of the battery are grounded by very short wires. The oscillograph is arranged so that the voltage between the contacts and the current through them can be observed, great care being taken to insure that the added apparatus does not appreciably change the circuit characteristics even at very high frequencies. A low power microscope may be set up to observe the operating area of the contacts.

When the contacts close, the first thing that happens is the discharge of the relay structure (a capacity) and of the wire through the contacts. The wire may be thought of as a radio antenna, more or less open-circuited at the load relay winding terminal, and either grounded or opened at the contacts. The wire forms an oscillatory circuit of moderately heavy damping with a surge impedance of about 100 ohms. As it is charged to 50 volts, when the contacts come together an oscillation having a peak current of 0.5 ampere occurs and is over before the steady current through the relay winding has more than started to build up. The frequency of the line oscillation depends on the length

and other characteristics of the wire, but it is (in the telephone plant) rarely lower than 500,000 cycles, and on short leads it may be many megacycles. Fortunately most contacts are not much affected by closing a half ampere. Erosion and build-up will occur, but at rather slow rates, and they are usually completely obscured by effects due to the contact opening. Of course, if the contacts bounce,[1] the effect will be more complex, but we are assuming for the moment that they do not bounce. The structure of the relay itself, including the pair of springs separated at its base by an insulating sheet, is also an oscillating circuit. We have not been able to get inside of this circuit and measure the current surge but its oscillation frequency seems to be about 250 megacycles for certain telephone relays.

Now suppose that the simple circuit of our closing contacts is complicated by an additional wire connected to the contact spring terminal. This is also charged to 50 volts before the contacts close, and being a second circuit of a hundred ohms surge impedance in parallel with the original wire, the current peak discharged through the contacts will now be about one ampere. But now the contacts are likely to act differently. About a microsecond after the current reaches its peak, but before the charge in the wires has been completely dissipated, the circuit is interrupted and the discharge stops. A spark, which is visible in the microscope, suggests that the current carrying areas have been exploded and blown apart. A few microseconds later they again close and the rest of the energy is discharged, but some of the contact metal must have been destroyed.

If several "idle" wires are attached to the contact, the current surge, and the number and duration of the contact reopenings, increase, but not usually in direct proportion to the number of wires. If the idle wires are attached to the load relay winding terminal instead of to the contact, the current is smaller, as the length of single wire from relay winding to contact is effectively in series with them.

In the telephone relay circuits which we are considering, the steady state current plays little part during contact closing if the contact carrying relay is properly adjusted, as the contacts come to rest while the current is still held at a small fraction of its final value by the inductance of the load relay winding.

Under some conditions which are more likely to occur in telegraph than in telephone circuits the contact closure phenomena are somewhat different from those described above. Assume, for example, that the potential between the open contacts may be adjusted in a range be-

[1] Bounce, as distinguished from chatter, reopens the contacts after several thousandths of a second.

tween 30 and 250 volts while the final direct current is limited by circuit resistance to less than 0.5 ampere. At the low voltage, observations of the current and voltage transients indicate that the closing contacts merely discharge the line. As the voltage is raised so that the current surge peak is in the range between 0.5 ampere and 1 ampere the reopenings, due presumably to overheating of the contacting areas by the discharge, are observed. These become more frequent as the voltage and current increase, and a new type of current surge begins to appear. This is placed on the time axis ahead of the point at which the initial closures have been occurring (usually 5 to 10 microseconds earlier) and consists of one or more irregularly spaced heavily damped pulses of current lasting only a small fraction of a microsecond and evidently discharging only a minute amount of the energy stored in the system. They occur perhaps once in a hundred closures at 30 volts, nearly every closure at 100 volts, and several for every closure at 250 volts. It is believed that the transients observed indicate the formation of minute metallic bridges [2] between the approaching contacts due to a softening of the metal by a cold point discharge and its deformation by the static field, and that once formed they are exploded by the discharge of current from the relay structure and adjacent wiring. The high fields necessary for phenomena of this type are of course due to the minute distances as the contacts approach final closure. A good deal of the erosion on telegraph relay contacts operating on capacitative loads or shunted by resistance-capacity "spark-killer" circuits is probably due to these "preclosures," but they are not thought to be of much importance at the lower battery voltages of the telephone plant.

Having now described the phenomena as contacts close in a doubtless over-simplified manner, we may consider that they have been closed for a long time, the direct current and the magnetic field of the load relay are established and the contacts are to be separated. The action now becomes really complicated and much of it is as yet only surmised. Several different things may happen, and these are influenced by humidity, dirt, surface films, absorbed gases and many other factors, including the speed of contact separation, the roughness of the surfaces, and the presence or absence of a wiping motion as well as the physical properties of the contact materials.

If the steady current exceeds certain well-known values ranging between 0.4 ampere and 1 ampere, characteristic of the contact materials,

[2] "The Formation of Metallic Bridges between Separated Contacts," G. L. Pearson, *Phys. Rev.*, Sept. 1, 1939, Vol. 56, pp. 471–474.

a metallic arc is formed as the contacts separate.[3] This is maintained at an initial potential of about 15 volts, and increases to a final value usually below 30 volts. The arc may last several milliseconds, but when it breaks it is followed by a complex transient lasting possibly another millisecond. These transients may be of two general types to be described later. This case is not of much importance in the telephone plant as the steady current is ordinarily kept below the value at which prolonged arcing occurs.

Metallic arcs lasting several ten thousandths of a second, and also followed by complex transients, may occur in breaking steady currents considerably less than those ordinarily believed to cause arcing. The effect of these transient arcs on contact life has not been studied separately, but they can hardly fail to increase the erosion. Their effect is unavoidably included in the studies of contact life in the higher range of direct current values. Figure 2 shows the voltage between a pair of opening silver contacts in which the steady current (0.25 ampere) is strong enough so that a brief metallic arc (indicated by the upward deflection of the trace to a new horizontal position) precedes the final transient.

If the steady state current is low enough so that neither prolonged nor brief metallic arcs are formed at the initial contact separation, one of two general types of complex transients occurs, or both types may be mixed. These have been designated the "A" and "B" types. The "B" type transient seems to be the more normal and it is difficult, probably impossible, to set circuit and contact conditions which will never give a "B" transient. It is identified by a bright spark between the contacts, showing in a spectroscope bright lines of the vaporized metal, and consists of a series of disruptive sparkovers at gradually increasing voltages. Each sparkover is individually very complicated. The appearance of the contacts during the "A" type transient is radically different from that during the "B" transient. There will be a minute bright spark, surrounded by a violet cloud which spreads out from the immediate contact area over the negative contact and sometimes travels as far as a sixteenth of an inch from the working area.

As a result of thousands of observations of the transient currents and voltages, and many experiments, and discussions with several physicists and engineers with whom the writer is associated, a plausible explanation of the phenomena has been arrived at and will be given as at least a working hypothesis.

[3] "Minimal Arcing Current of Contacts," H. E. Ives, *Jour. Franklin Institute*, October, 1924.

The voltage wave form of an entire "B" transient, covering the time from the initial separation of the contacts to the final subsidence of the voltage charging the line wire, is shown in Fig. 3, and the a-c. components of the current in the range from 20 kilocycles to 20 megacycles are shown in Fig. 4. The low-frequency components of the current are comparatively weak. A line and load relay were chosen to give a relatively simple transient with the important components at frequencies which could be photographed. A 500-ohm Western Electric U-type relay and a line of 300 ft. of No. 22 switchboard pair were used. The mate wire of the pair was grounded at both ends. The currents and voltages were not photographed simultaneously but the types of the transients were correlated by repeated observations. A current picture will not exactly correspond to a voltage picture, as the transients produced by successive operations of a contact are never identical.

The "B" transient may be explained as follows, using as a basis the simple circuit of Fig. 1. The steady current is established and the contacts start to separate, moving apart at a speed, which is at first surprisingly slow (about an inch a second). The contacts have been deformed by the pressure between them, and as this is relaxed the current density and the temperature at the contacting areas rapidly increase until at some light pressure the area becomes so small that the current explodes it. There may be some necking out of the softened contacts before this and under some conditions there are indications of a metallic arc lasting a fraction of a microsecond, but at any rate an initial rupture occurs between hot and soft metal areas.

The wire has been at ground potential, but the battery plus the collapsing magnetic field of the load relay commence to charge it at a rate depending on the line and relay winding capacity and the relay inductance and losses. In ten or twenty microseconds, it has reached at the contacts a potential of from 50 to 200 volts. This is below the voltage at which sparkover due to ionization of the air can occur, but something usually happens which recloses the circuit. This is believed to be caused in somewhat the same manner as the "preclosures" mentioned earlier. It is probable that a cold point discharge reheats the contacts. This is followed by a collapse of the voltage to about 15 volts above zero in the direction of the previous voltage, indicating the formation of a metallic arc. This lasts a fraction of a microsecond and the voltage then drops to nearly zero, suggesting that the contact areas heated by the field current and the arc have been drawn together in solid metallic contact. The line is discharged with an oscillation of comparatively low damping (which is characteristic of the line wire)

reaching a current peak usually ranging from 0.5 to 2 amperes. The first cycle of the oscillation is distorted by the higher resistance of the path to ground caused by the arcing stage in the reclosure. After a few microseconds the contacts are opened a second time by the continued motion. Occasionally they reclose a second time but they usually stay open until the voltage has built up by the continued discharge of the load relay inductance to a value between 300 and 350 volts. Then a spark occurs at what is usually considered the minimum sparking potential between contacts in air.

Figures 5 and 6 show the voltage and current of the initial opening and reclosure of the contacts at the start of a "B" transient. The brief arc at initial opening is barely detectable in Fig. 5. Figures 7 and 8 show similar voltages and currents at an increased sweep speed. In Fig. 7 the metallic arc established during the reclosure is plainly evidenced by the collapse of the voltage to about 15 volts and its maintenance at this value for about a microsecond before it drops to zero. The effect of the arc in distorting the oscillating discharge of the current from the line wire is evident in Fig. 8. The current oscillation of Fig. 8 may be duplicated merely by charging the line wire to a suitable voltage through a high resistance and closing the contacts, the far end of the line being grounded through the load relay and a large condenser which replaces the usual battery.

It is likely that the point discharge precedes the arc on reclosure by such a short time that it cannot ordinarily be resolved. Nevertheless disturbances of the voltage and current are occasionally found which seem to indicate that a discharge path formed and was checked (possibly by melting off the point) without establishing an arc or metallic bridge. Such a disturbance of the rising voltage is indicated in Fig. 9 by a high-frequency oscillation about 5 microseconds after the first rise of the voltage trace. Figure 10, which shows the current of the second of two initial reclosures, indicates a similar phenomenon. Five microseconds after the rupture of the circuit, shown by the downward deflection of the zero line, a dim line upward records a current surge lasting a fraction of a microsecond and reaching about 3/4 ampere. This surge, however, did not result in the immediate formation of an arc which was established about 5 microseconds later.

The initial separation of the contacts does not always result in a metallic reclosure. Figure 11 shows the voltage of the early part of the "B" transient. Here the first collapse of the voltage is a sparkover from about − 300 volts which establishes an arc at about − 15 volts. This arc is broken and, as the line is not completely discharged, the voltage between the contacts rises to about + 140 volts; a second arc is

established at + 15 volts and broken in its turn. Possibly because of the continually increasing distance, the arc is not reestablished, and the voltage builds up with oscillations of a frequency characteristic of the line wire insulated at both ends until it reaches − 300 volts a second time and another spark passes. This time only one arc is formed, and the recovery of the voltage starts from the positive side of the zero axis. The current surges corresponding to the voltage collapses of Fig. 11 are shown in Fig. 12. Here the first pulse represents a sparkover which formed only one arc. As the current from the line reached about 4 amperes it was checked and the conducting arc was broken (possibly by being extended laterally into the region of cooler metal). The second pulse shows the current of a sparkover which formed two arcing periods.

These phenomena are shown in more detail in Figs. 13 and 14 which show the voltage and current of a sparkover forming only one arc, and 15 and 16 which show the wave forms when two arcs are formed. Note that the frequency of the current oscillation is that of the line grounded at one end only (the impedance of the load relay being high at this frequency) and is about half that of the voltage oscillation which is that of the line open at both ends. Oscillations of both frequencies may be found in the line at a distance from either end.

Corresponding observations may be made of the occurrence of 3, 4 and 5 arcing periods, the pattern followed being about the same. The higher the voltage at sparkover the more arcing periods; an odd number of arcing periods is followed by a recovery of the voltage from the opposite side of the zero axis from that of the voltage before sparkover, an even number by recovery from the same side of the zero axis. The arcing periods are individually complex, having superposed on them oscillations believed to be due to the relay structure and the leads to the oscillograph which are too fast to be resolved photographically by the means available. These oscillations may be observed visually by using higher sweep speeds and reach frequencies of 250 megacycles.

While the arcs ordinarily do not exceed a microsecond in duration, they are probably an important factor in determining contact erosion, as several hundred may occur at each contact opening.

As may be seen from Fig. 3 the sparkovers continue to occur, the successive voltage breakdowns corresponding to the normal sparking potential as the contact separation increases with time (with some irregularities due to residual ionization in the gap) until the separation is finally so large that the energy remaining in the load relay cannot charge the line to the breakdown voltage. At this stage the line dis-

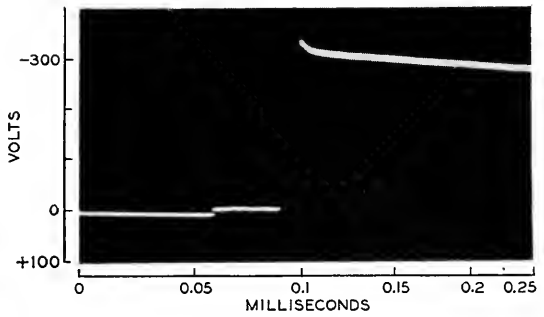


Fig. 2—"A" transient starting with metallic arc (voltage).

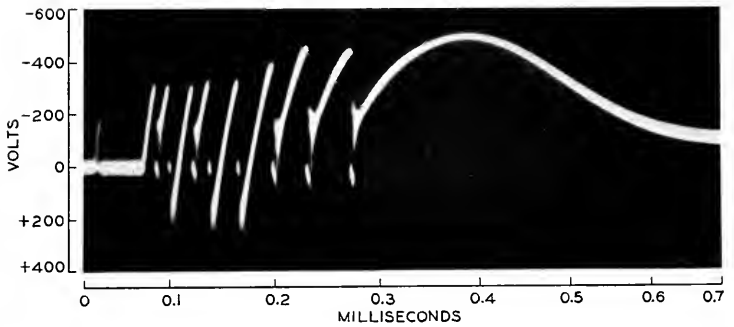


Fig. 3—Entire "B" transient (voltage).

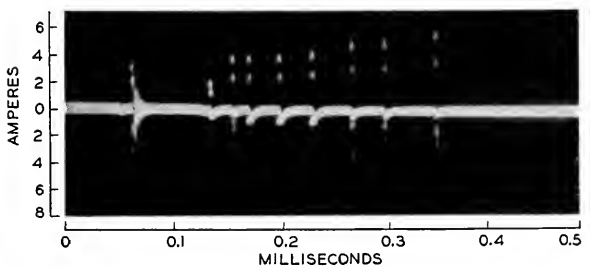


Fig. 4—Entire "B" transient (current).

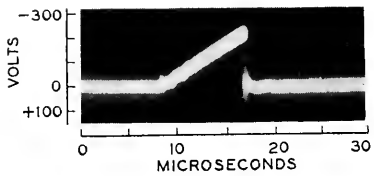


Fig. 5—Initial opening and reclosure—"B" transient (voltage).

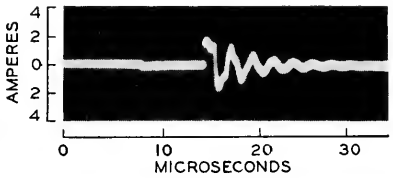


Fig. 6—Initial opening and reclosure—"B" transient (current).

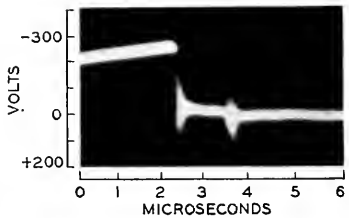


Fig. 7—Initial opening and reclosure—"B" transient (voltage) rapid sweep.

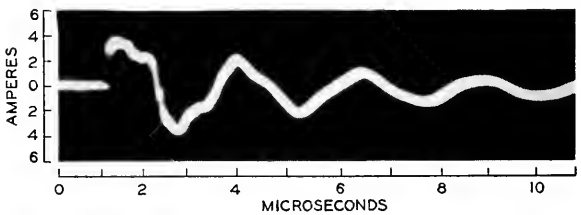


Fig. 8—Initial opening and reclosure—"B" transient (current) rapid sweep.

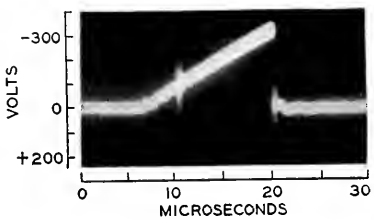


Fig. 9—Evidence of point discharge, voltage.

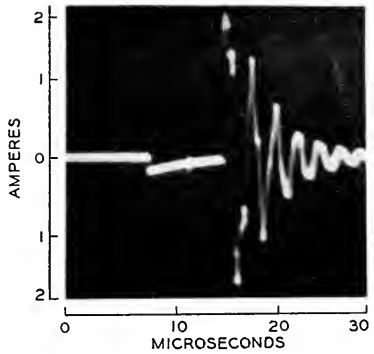


Fig. 10—Evidence of point discharge, current.

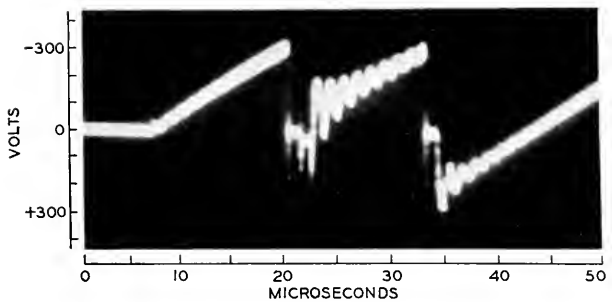


Fig. 11—Early part of "B" transient, voltage.

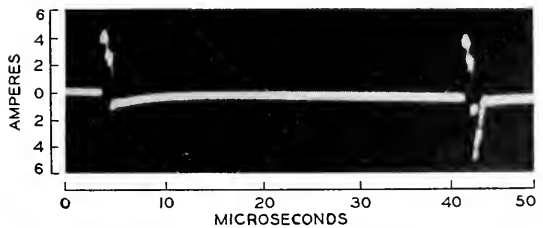


Fig. 12—Early part of "B" transient, current.

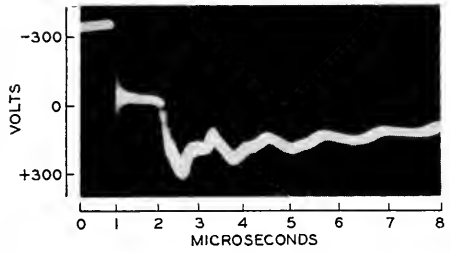


Fig. 13—Single sparkover of "B" transient, with single arc (voltage).

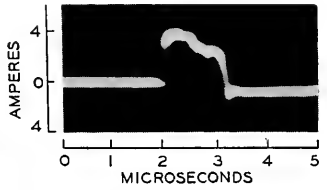


Fig. 14—Single sparkover of "B" transient, with single arc (current).

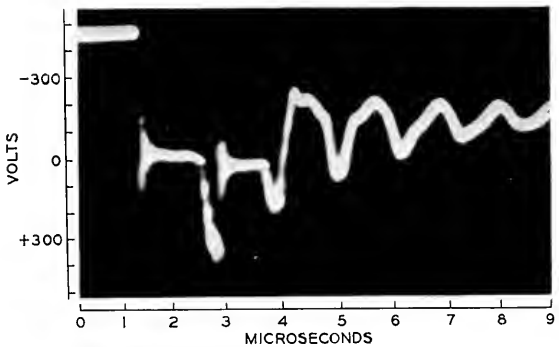


Fig. 15—Single sparkover of "B" transient, with double arc (voltage).

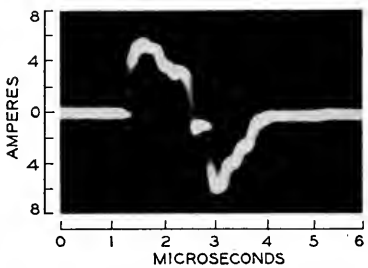


Fig. 16—Single sparkover of "B" transient, with double arc (current).

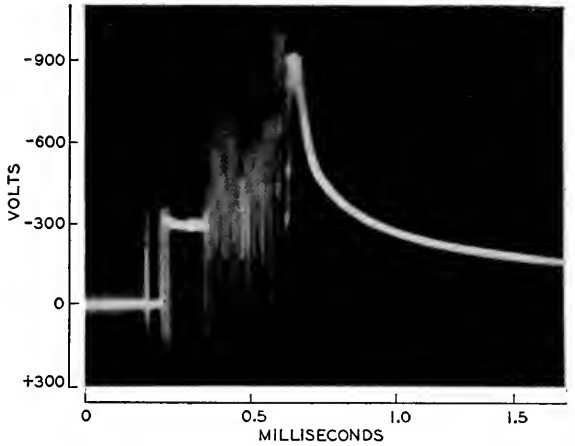


Fig. 17—Typical "mixed A and B" transient (voltage).

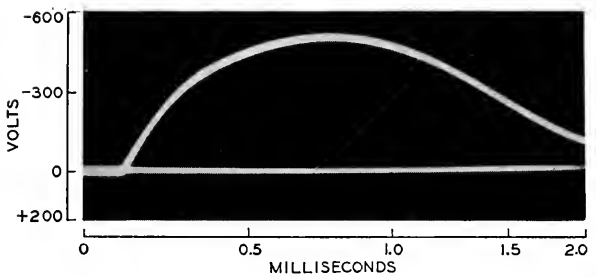


Fig. 18—Effect on voltage transient of changing wire line length—1100 ft.

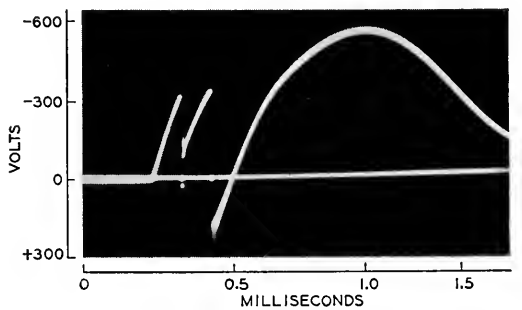


Fig. 19—Effect on voltage transient of changing wire line length—600 ft.

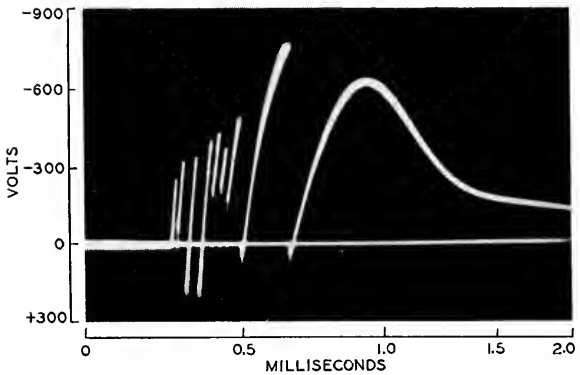


Fig. 20—Effect on voltage transient of changing wire line length—150 ft.

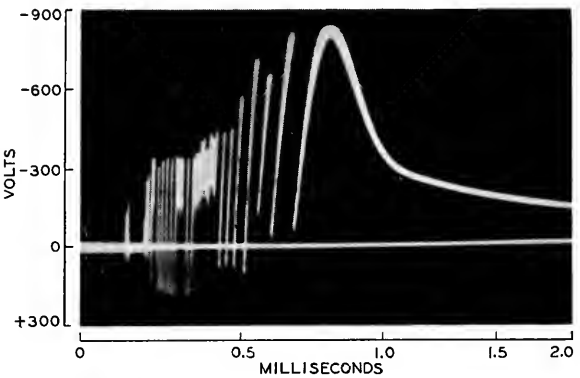


Fig. 21—Effect on voltage transient of changing wire line length—50 ft.

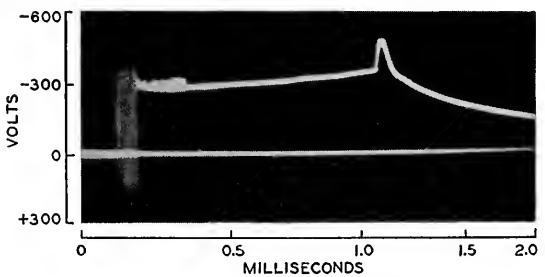


Fig. 22—Effect on voltage transient of changing wire line length—10 ft.

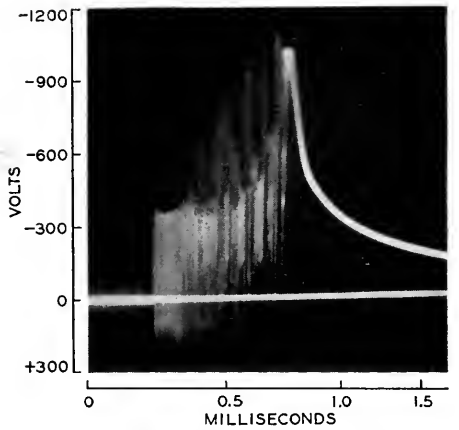


Fig. 23—Effect on voltage transient of changing wire line length—10 ft.

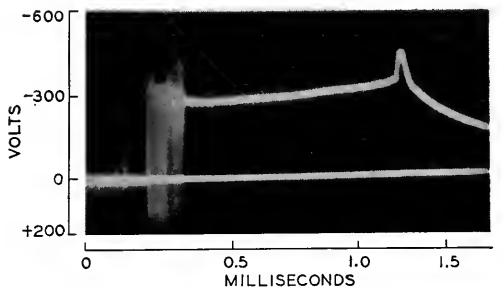


Fig. 24—Effect on voltage transient of changing wire line length—10 ft.

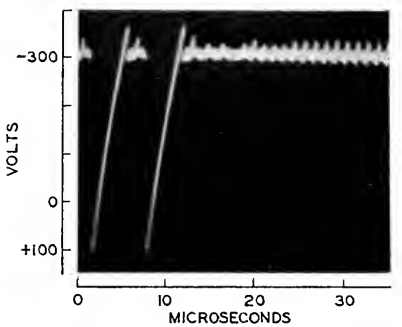


Fig. 25—Oscillation on glow discharge of "A" transient (voltage).

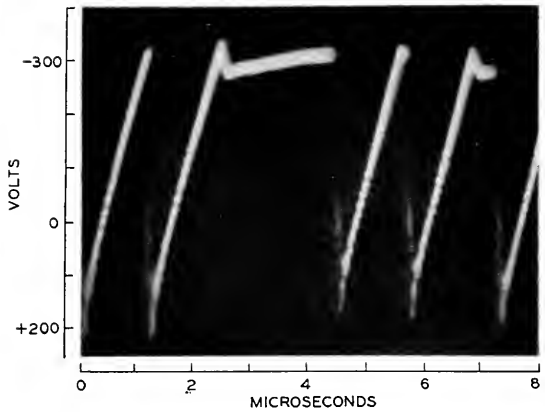


Fig. 26—Start of "A" transient (voltage).

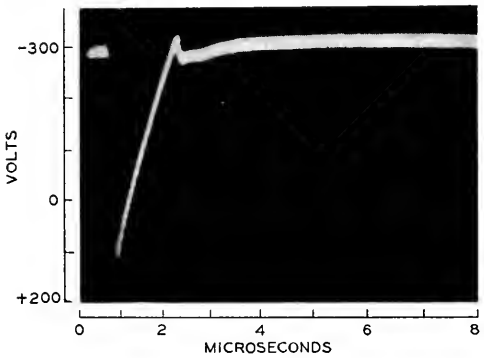


Fig. 27—Start of stable glow discharge of "A" transient (voltage).

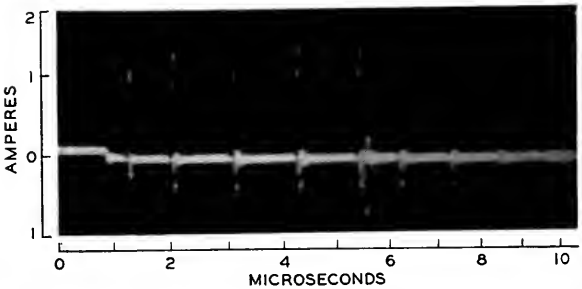


Fig. 28—Start of "A" transient (current).

charges slowly back through the load relay to the battery voltage. The peak voltage reached may be as high as 2000 volts.

The principal characteristics of the "B" discharge can be produced by a simple experiment which does not use a load relay, The transient is not dependent on a load inductance, but only on a source of voltage which will charge a wire at a sufficiently rapid (but not too rapid) rate while a pair of contacts, which initially ground the wire, are separating. If a wire about 100 ft. long is connected to a source of somewhat more than 350 volts through a resistance of from 5000 to 20,000 ohms, and is also grounded by a contact at one end, a transient is produced when the contacts open which shows the characteristics of a "B" type transient except the final dying away of the voltage to 50 volts.

It must not be understood that every spark transient is purely of either the "A" or the "B" type. It is very common for the "A" type transient to break down into the "B" type and less often the "B" transient establishes the gas glow discharge for a brief period in the middle of the sparkovers.

A "mixed" transient is shown in its entire duration in Fig. 17. Here, after a group of sparkovers, a period in which the voltage is maintained steadily at about 300 volts for about 0.0002 second intervenes, and is followed by more sparkovers from considerably higher voltages. In order to produce this transient the length of the line wire was reduced to 10 ft., at which length and with a 1000-ohm load relay the tendency is to produce intermittent groups of "A" or "B" transients, interspersed with the mixed type shown, when the relay is operated frequently.

The number of sparkovers in each "B" transient varies with the circuit conditions. As many as a thousand may be found with a load consisting of a number of relays in parallel on a wire of moderate length and as few as one in the limiting case.

While the occurrence of the "B" transient is favored by long line wires and high impedance relay loads, beyond a certain length which with telephone relays and wiring is from 300 to 2000 ft. (the longer lengths being associated with the lower impedance relays) no sparkovers at all occur. The voltage build-up is so slow that the sparkover potential is not reached at any time during contact opening and the contacts may be said to be protected by the line wire. The series of voltage oscillograms, Figs. 18 to 24 inclusive, shows the change from a smooth transient with no sparkovers through the "B" type with an increasing number of sparkovers to the final "A" type. The "A" type transient of Fig. 22, which has superposed on the 300-volt gas glow discharge stage a relaxation type of oscillation, the "B" transient

of Fig. 23 and the simple "A" transient of Fig. 24, were all produced under identical conditions in quick succession. The change in characteristics from Fig. 18 to Fig. 24 was produced merely by a reduction in the length of the connecting wire from 1100 ft. through three intermediate stages to 10 ft., a 1000-ohm load relay being used. This explains a puzzling effect noted with many contact materials. With a supposedly identical circuit, the erosion will be small with very short wires, increase rapidly as the wiring length increases, and then decrease again becoming very small with very long wires.

The "B" transient is more frequently observed with freshly filed contacts, at high humidities, and with a rolling or wiping motion of the contacts in opening. It is always found if the contacts are of oxidized metal or operate in an oxygen atmosphere. In fact, there seem to be good reasons for believing that its production is bound up with the presence of oxygen on or in the surface of the active contact metal.

It may be seen from the last series of oscillograms that if the circuit and the conditions of the contact surfaces are just right, the "B" transient is replaced by a much simpler and less stable type, the "A" transient. It will occur usually when the wiring is short or the load relay is of low impedance, with contacts which have been operated until the original surface has been burned off and have not stood idle more than a few minutes. It starts much as does the "B" type, but after a dozen or a hundred sparkovers from about 350 volts, which come much closer together in time than those of the "B" transient, the voltage becomes steady at about 300 volts. This condition lasts for perhaps 0.6 millisecond, then the voltage rises to about 400 or 450 volts and gradually reduces, reaching the battery voltage after several milliseconds. A typical "A" type transient is shown in Fig. 24. It is suggested as a hypothesis that, during the sparkover stage, the oxygen is being exhausted from the surfaces of the current carrying areas of the contacts by burning the metal and that when this has been completed, a nitrogen gas glow discharge is formed and maintained during the rest of the contact opening, if the supply of energy from the load inductance through the line is rapid enough to prevent the voltage from dropping below about 280 volts.

The glow discharge phase of the "A" transient is unstable. If transient voltages induced by the operation of relays in other circuits reach the contact gap during the time that a sustained glow discharge is attempting to form, its formation is interfered with and a mixed or "B" type transient results. Occasionally, as illustrated in Fig. 22, the glow discharge of the "A" transient has superposed on it a saw-

toothed oscillation of from 10 to 50 volts peak-to-peak. Part of an "A" type transient showing this peculiarity is illustrated by the oscillogram of Fig. 25. This appears to be a relaxation oscillation such as is commonly produced in ionized gas tubes, riding on the normal 300-volt axis of the gas glow discharge. The conditions which lead to the occurrence of this oscillation at atmospheric pressure have not been identified, but it is found to be quite stable in some cases where contacts have been sealed in a mixture of air and gases at about half atmospheric pressure.

A typical "A" transient is shown in detail in Figs. 26, 27, and 28. The circuit consisted of a 250-ohm relay connected to the contacts by 10 ft. of wire, the battery being 50 volts as usual. Figure 26 shows the voltage of the early part of the transient during which rapid sparkovers are interspersed with two brief periods during which a gas glow discharge was established but not maintained. Figure 27 shows the final sparkover before the establishment of the glow discharge at about 300 volts. A group of the current pulses corresponding to the initial part of the sparkover stage is shown in Fig. 28. These are complicated by the line oscillations (which should be of about 30 megacycles frequency) and appear to last less than 0.1 microsecond.

It may be seen that the individual sparkovers at the start of the "A" transient are somewhat different in form from those of the "B" transient. The voltage reaches 320 volts in a microsecond or so, and in some cases collapses to zero or beyond immediately. There are sometimes indications of arcing periods lasting much less than 0.1 microsecond and the voltage recovers with oscillations of the line wire but the duration of the phenomenon is too brief for very accurate analysis. But in many cases, the voltage, having reached its peak, drops to an intermediate value of 280 volts and recovers to 320 volts before it collapses. This is probably due to the temporary formation of the nitrogen glow discharge, which is finally established and maintained during the remainder of the contact opening when for some reason the sparkover does not occur. In cases where the contacts are on the verge of producing a "B" transient the voltage may rise to 500 volts and then collapse to the 300 volts of the gas glow discharge.

It is very interesting to set up a circuit which will cause the "A" transient to predominate, and start operating freshly filed contacts several times a second observing the transient voltage at contact opening on the oscilloscope. The first transient will always be of the "B" type. Usually the first few dozen will also. However, after a while one of the transients will show a flat top at about 300 volts for a very brief period and this tendency increases until finally a complete

"A" transient occurs. After this, the "A" transients become more and more common until finally the "B" transients occur perhaps once in a hundred openings. If, then, a gentle stream of oxygen is blown on the contacts, only "B" transients will occur until a few seconds after it has been turned off. Blowing the breath on the contacts has a similar but less definite effect, while a stream of dry compressed air has no effect.

If, on the contrary, the circuit conditions are selected so that "B" transients predominate, a stream of nitrogen will induce "A" transients. That is, "A" transients are not found in oxygen and "B" transients are rare in nitrogen.

If, instead of operating the contacts several times a second, they are operated at longer intervals, the tendency to produce the "A" transient is reduced. When contacts are operated in air a certain interval between operations can be found which causes all transients to be of the "B" type. This probably depends on humidity and also on circuit conditions and contact material. In one experiment, a wait of 45 seconds between operations gave all "B" transients with silver contacts, while a wait of five minutes was required with palladium contacts. This is possibly due to a different rate of film formation.

Life tests on palladium contacts show much lower erosion with "A" transients than with "B" transients. The effect of the two types of transient in terminating the life of silver contacts is not markedly different. The contours of the eroded surfaces exhibit a wide variety, and it is not easy to correlate the transient type with its effect. It is evident, however, that areas of the contacts which have never been in the direct current path may be severely eroded.

When we consider that the "B" transients produce oscillations in the line wires reaching several hundred volts and often fifteen amperes, it is not to be wondered at that clicks will be produced in circuits in the immediate neighborhood of unprotected relay contacts. The "A" transients produce much weaker currents than the "B" transients and many contacts on successive operations will produce "A", "B", or mixed types. This explains the common observation that relay clicks vary over a wide range of amplitudes. The arrangement of telephone circuits in which the cabled wiring always contains a large number of grounded conductors, and is often enclosed in a lead shield, prevents any appreciable free radiation of the spark transient oscillations.

With the foregoing information available the contact erosion process at opening contacts appears briefly to be as follows. At very minute separations high field strengths exist even for moderate voltages. The resulting cold point discharge is often followed by a metallic arc

which softens a tiny point on the contact which is pulled out and fused into metallic contact under the action of the high fields. After rupture by increasing separation or increasing current density, the process may repeat or, as is more likely, the separation is too great for another metallic bridge to form. The high field discharge then sets the stage for the next type of conduction or breakdown. This may be either a series of sparkovers interspersed with metallic arcs of extremely short duration or a gas glow discharge, initially intermittent and then more or less stable. Factors predisposing toward one or the other type of discharge are known thus far only in a most general fashion and much remains to be done before the relation between contact erosion and the transient currents and voltages can be predicted accurately. There is ample evidence that molten metal may be expelled from the immediate contact area at high velocity and may be deposited at distances of at least 0.1 inch. It also appears that both the ionized nitrogen cloud of the "A" transient and the disruptive sparks of the "B" transient may corrode the contacts and their supports at locations and distances which never enter directly into the rupture of the current path.

We have seen that the line wire contributes to the current surges through contacts due to its properties as an oscillatory circuit, charged repeatedly by the energy stored in the magnetic field of the relay. The surges and the resultant erosion may be reduced in several ways. If a radio frequency choke coil is connected between the contact and the line wire, the discharges of the latter are much reduced, and the "A" type gas glow transient favored. A group of many current surges of 15 amperes peak may in most cases be reduced to one or two of 0.15 ampere or less, and a radical reduction in erosion secured. Unfortunately choke coils are expensive and inconvenient. The usual line wire may be terminated in approximately its surge impedance by shunting both ends to ground with a resistance of about 100 ohms in series with a condenser of the order of 0.01 mf. This heavily damps the line oscillations and greatly reduces the number and severity of the current surges. It is also expensive. Instead of the copper line wire, a material such as iron or permalloy plated copper having a high surge impedance and large high frequency a-c. losses may be used. This seems more practical, but brings up new problems in design, handling, and soldering.

The most effective means of reducing erosion is of course the well known "spark-killer" (consisting of a condenser and resistance in series, shunted across the contact or load), which can be designed to hold the voltage below the sparkover point at least until the contacts have separated a safe distance.

When the conventional spark-killer is used it is generally assumed that what sparking then occurs is due to the discharge of the condenser when the contacts close, provided that the "spark-killer" prevents the voltage at contact opening from reaching 350 volts. Unfortunately the "reclosure" effect described earlier appears unless the initial rise of voltage as the contacts separate is held down to a value considerably below the sparking potential by a suitable choice of the resistance in series with the "spark-killer" condenser. If the rate of increase of the initial voltage in relation to the speed of separation of the contacts exceeds a figure which seems to depend on the contact material and the condition of its surfaces, the high field point discharge comes into play and causes the separating contacts to reclose metallically while they are still at a minute separation and moving apart very slowly. In "reclosing" the line wire and condenser are discharged, the current explodes the minute metallic bridge, producing a visible spark, and the circuit is thus reopened. This may occur a dozen times in some cases before the contacts finally stay separated. The higher the voltage which the spark-killer permits the more likely are the reclosures to take place, and the larger the number of reclosures at each contact opening. However, reclosures are usually not very common in cases where the voltage of the wave front is held below 50 volts. In the majority of cases in the telephone plant it is possible to do this without incurring much of a penalty due to erosion of the contacts on closing by the discharge of the spark-killer condenser.

This discussion is not more than sufficient to serve as an introduction to the problems of contact sparking as revealed by the improved observing technique used in this study. Only the simplest cases have been considered, and the telephone plant is far from being simple. Many relays have multiple windings or metal sleeves, and multiple connections to the contacts are very common. As these complications considerably modify the contact spark wave form and erosion, each contact with associated circuits presents its own problem. The solution of these problems involves the careful study of circuit characteristics of a type which are ordinarily left to the radio engineer, as well as of the mechanical, chemical, and metallurgical properties of the contact materials.

The writer wishes to acknowledge the collaboration of Mr. E. T. Burton in the observation and explanation of the phenomena and the assistance of Mr. I. E. Cole in the development of the testing apparatus; of Mr. Glass, who developed the cathode ray tubes; and that of many engineers and physicists in our organization, in particular Messrs. Mathes, Hogg, Goucher, and Pearson, in the formulation of some of the hypotheses expressed.

Effect of the Quadrature Component in Single Sideband Transmission

By H. NYQUIST and K. W. PFLEGER

A PREVIOUS article [1] gives an analysis of single sideband transmission. Since that article was written this subject, particularly in its application to picture transmission and television, has assumed considerable importance. For this reason it now seems desirable to amplify the previous theoretical treatment and to indicate certain experimental results which have been obtained in the meantime. The present article gives experimental evidence that, for a given bandwidth, single sideband transmission is distinctly superior to double sideband in picture transmission.[2] It also gives a theoretical discussion which indicates that this is not inconsistent with the observed fact that oscillograms with single sideband transmission show considerable distortion.

As described in the previous article distortion to be considered in single sideband transmission as compared with double sideband transmission arises in three ways.

1. There may be present a slowly varying in-phase component due principally to the inaccurate location of the carrier frequency with respect to the edge of the filter characteristic.

2. The edge of the filter characteristic where the carrier is located may be so designed that there is a net distortion due to failure of the vestigial sideband to be accurately complementary to the principal sideband.

3. There is present a quadrature component which results in considerable distortion of the envelope of the received wave under ordinary conditions.

By in-phase component is meant a component whose carrier is in phase with the steady state carrier; by quadrature component is meant a component whose carrier is in quadrature with the steady state carrier. In some of the theoretical work in the present article, idealized

[1] *Trans. A. I. E. E.*, Vol. 47, p. 617, April 1928.

[2] A paper by Goldman: "Television Detail and Selective-Sideband Transmission," *Proc. I. R. E.*, Vol. 27, pp. 725–732, Nov. 1939, dealing with the same subject, has been published since our manuscript was sent to the printer. While the two papers reach similar conclusions there is considerable difference in method between them.

transducers have been assumed such that the first two effects listed above are absent. In the physical networks which are covered by the experimental work and part of the theoretical work these effects, while not absent, are found to be unimportant. The present discussion therefore is principally concerned with the third of these effects, namely, the quadrature component.

In a recent paper Smith, Trevor and Carter [3] have studied, both mathematically and experimentally, the matter of single sideband transmission over a rather simple filter and have found that the envelope is greatly distorted when the single sideband transmission is used. They give characteristics of their filters and also the location of the carrier frequencies so that it is possible to deduce that the first two effects, listed above, are unimportant for some of the carrier frequencies used. Their filter characteristics fall easily within the usual requirements for single sideband picture transmission at a speed appropriate to the bandwidth. Substantially the sole source of distortion in their work is the presence of the quadrature component, when the carrier frequency is suitably located.

Studies have also been made of a picture transmitting system of the type described by Reynolds.[4] This system makes use of single sideband transmission which had been found in previous experiments to be practicable. These previous experiments had shown that the quadrature component was present and was of considerable magnitude, but that the impairment in the picture was rather slight. They had also shown that if sufficient current was transmitted for the darkest portion of the picture the impairment could be reduced to the point where it was practically not detectable, and that a fairly small dark current would suffice.

Computations

The present section will be devoted to the computations of in-phase and quadrature components corresponding to certain assumed idealized characteristics, and reasons will be indicated why the picture impairment should be materially less than might be expected from the appearance of oscillographic records of the signal.

Figure 1 indicates the magnitude of the transfer admittance characteristic which will be assumed. The characteristic is made up of two half-cycles of a sine wave separated by a horizontal portion. The phase shift vs. frequency characteristic is a straight line. In order to simplify subsequent sketches this constant delay has been put equal

[3] *R.C.A. Review*, Vol. 3; p. 213, October 1938.
[4] *Bell System Technical Journal*, Vol. 15; p. 549, October 1936.

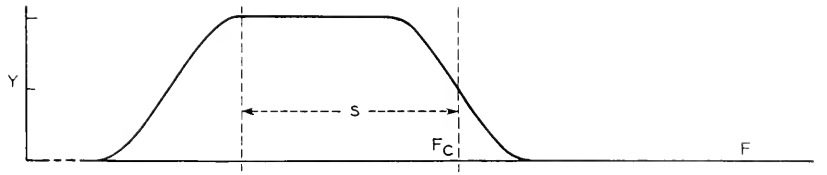


Fig. 1—Idealized transfer admittance characteristic. (Band pass system with no delay distortion; F_c is the carrier and s the fundamental dotting frequency.)

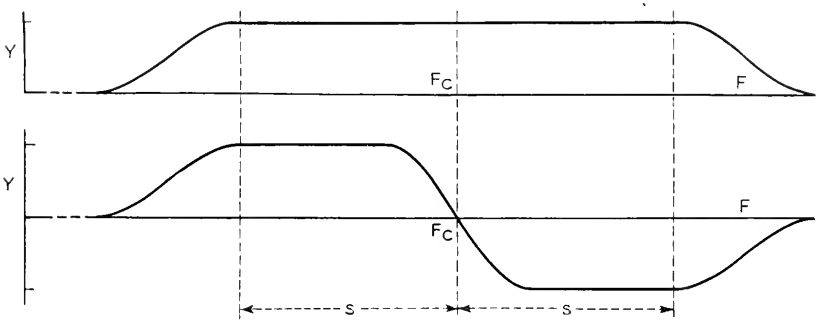


Fig. 2—Graphical analysis of transmission characteristic. (The sum of these characteristics equals that in Fig. 1. The upper gives received signals with carrier in phase, and the lower, in quadrature with the sent wave.)

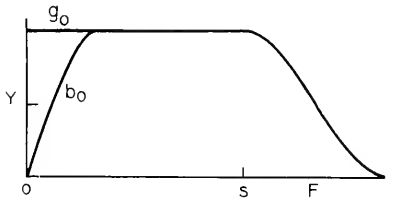


Fig. 3—Equivalent low-pass filter characteristic. (Used in computing envelopes of received signals for single sideband transmission.)

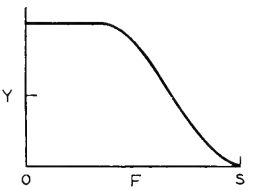


Fig. 4—Equivalent low-pass filter characteristic. (Used in computing envelopes of received signals with mid-band carrier.)

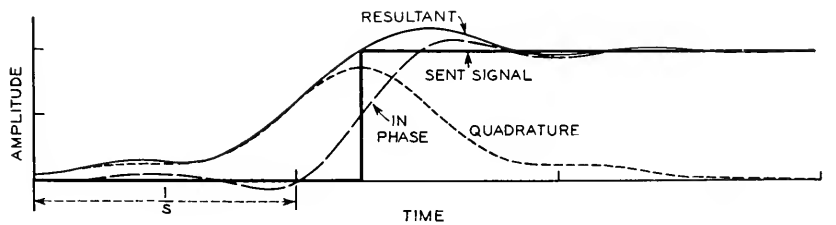


Fig. 5—Envelopes of received wave components for single transition, characteristic as shown in Fig. 1.

to zero. To take account of any constant delay it is sufficient to displace the computed curve by an amount equal to the delay. The characteristic of Fig. 1 may be separated into two components as indicated in Fig. 2, where the top one gives rise to the in-phase component and the bottom one to the quadrature component. F_c is the carrier frequency for single sideband computations, and it is assumed that F_c is great in comparison with the bandwidth. The characteristic of Fig. 1 does not differ greatly from those used in the experimental work. The quadrature component with the assumed characteristic is somewhat more pronounced than with the experimental ones. Figure 3 shows the equivalent low pass characteristics. Curve g_0 gives rise to the in-phase component and curve b_0 to the quadrature component. Figure 4 shows the low-pass characteristic which is equivalent to the original characteristic for double sideband computations with the carrier located in the middle.

Figure 5 gives the computed envelope for a single transition when this transducer is used on a single sideband basis. The figure shows the rectangular sent wave, the envelope of the in-phase component, the envelope of the quadrature component, and the envelope of the resultant wave. Figure 6 shows the corresponding received wave for the double sideband case. There is no quadrature component and the in-phase component and the resultant are identical. Figures 7 and 8 show the single sideband envelopes for a unit dot and a unit space, respectively. Figure 9 shows two dots in succession. Figure 10 shows the same case as Fig. 9 with the exception that dark current 14 db below the maximum current has been added. Figures 11 and 12 correspond to Figs. 9 and 10, the difference being that the dots are shorter. Figure 13 shows a succession of five dots. Figure 14 shows two dots as transmitted on a double sideband basis. In all the figures but 11 and 12 the fundamental dotting frequency is s as indicated in Figs. 1, 3 and 4. In Figs. 11 and 12 the dotting frequency is $4s/3$.

In comparing these figures a number of things will be apparent. In the first place, there is in the single sideband case a considerable broadening of all the marks due to the presence of the quadrature component. A second effect to be noted is that this broadening does not cause the dots to run together nearly as much as might be expected. This is particularly striking in Fig. 11 where the running together of the two dots is only slightly greater than it would be with the in-phase component alone. The reason for this is that when the dots tend to run together the contributions from successive dots to the quadrature component tend to cancel each other instead of adding to each other as is the case with in-phase components. The broadening of Fig. 11

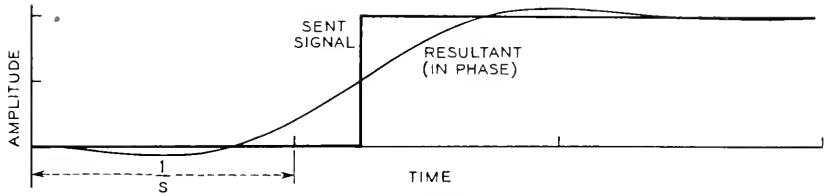


Fig. 6—Transmission of single reversal, carrier at mid-band.

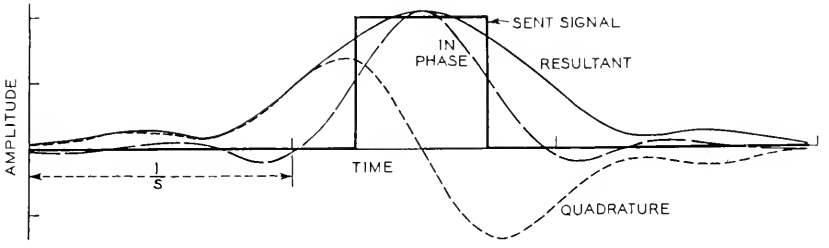


Fig. 7—Received signal for single dot, characteristic as in Fig. 1.

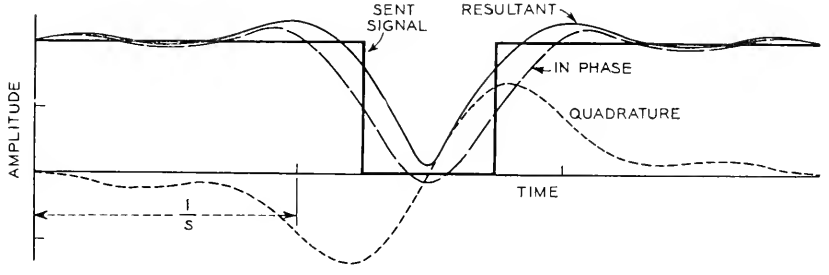


Fig. 8—Received signal for single space, characteristic as in Fig. 1.

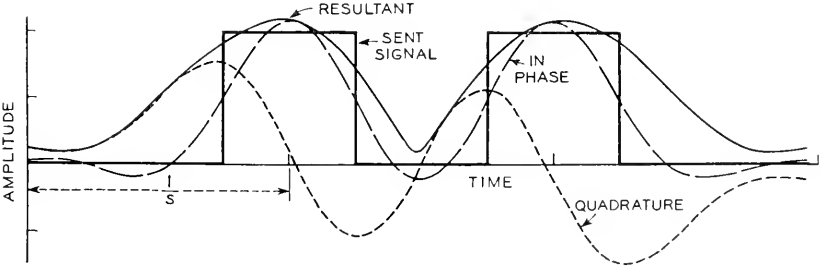


Fig. 9—Received signal for two dots in succession, characteristic as in Fig. 1.

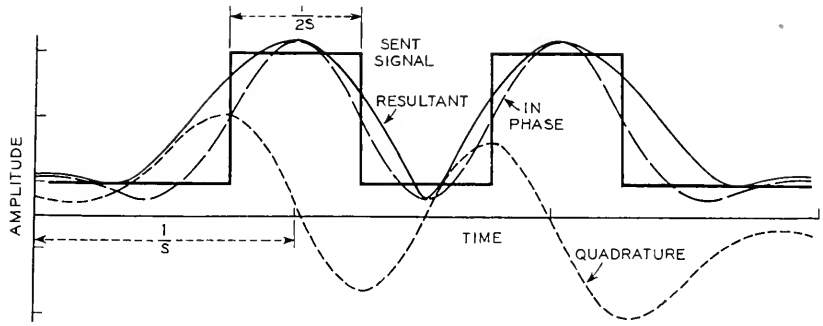


Fig. 10—Effect of transmitting dark current 14 db below maximum. (Compare with Fig. 9.)

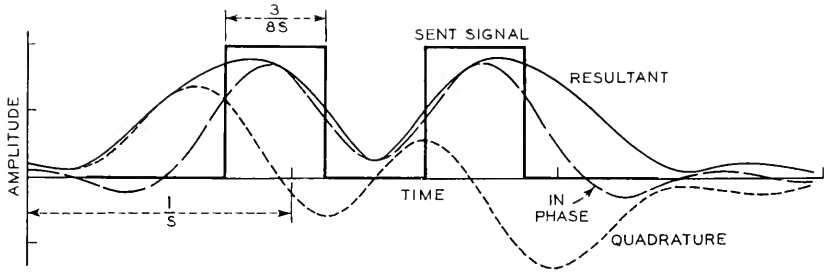


Fig. 11—Effect of shortening dots. (Compare with Fig. 9.)

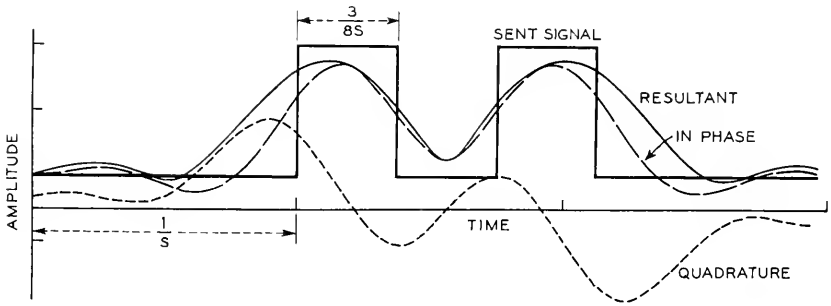


Fig. 12—Effect of adding dark current with shortened dots. (Compare with Figs. 11 and 10.)

and similar figures is principally on the outside of dots rather than on the inside. This tendency of the quadrature component to disappear when very short marks are employed, accounts for the observed fact that fine details are separated with single sideband methods as well as with double sideband methods using twice the bandwidth. Thirdly, the figures illustrate the effect of having finite dark current. This

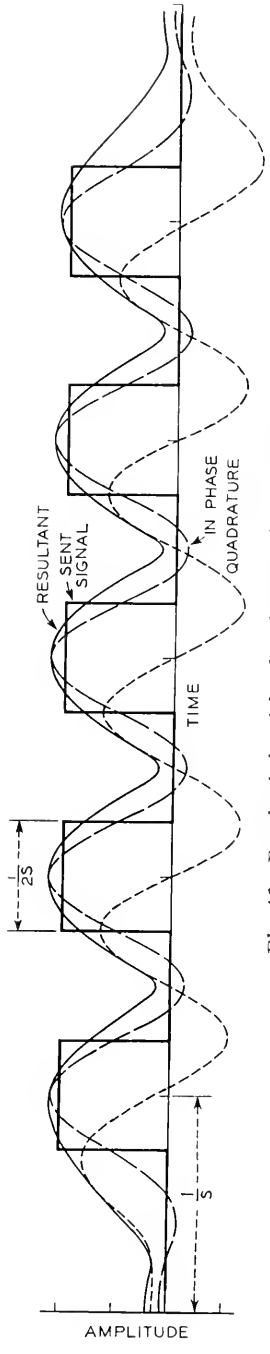


Fig. 13—Received signal for five dots, carrier as in Fig. 1.

effect is discussed below. Observations of transmitted pictures have shown that with the dark current of the magnitude indicated, it is practically impossible to detect the impairment from the quadrature component, although distortion is still evident on the computed curves. Figure 14 shows the relatively greater tendency for the double sideband dots to run together than the single sideband ones, for the same total bandwidth. The contributions from the two dots are, of course, in phase and therefore tend to add in the intervening space. It has been pointed out that with single sideband transmission the corresponding contributions to the quadrature component tend to cancel each other under these conditions.

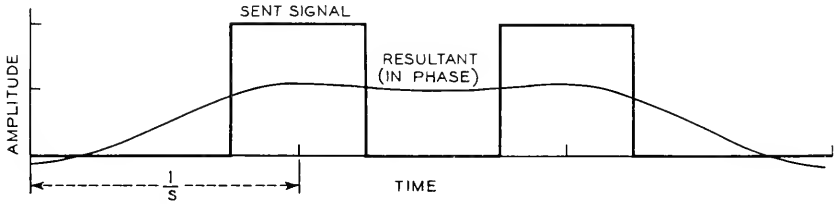


Fig. 14—Received signal for two dots, mid-band carrier.

Discussion

In estimating the effect of the quadrature component it is instructive to compare the in-phase component and the resultant in, say, Fig. 13. It will be evident that if the latter wave were used, for instance, for telegraph transmission there would be a considerable bias due to the quadrature component whereas the in-phase component shows practically no bias. Such a resultant wave would show a decided impairment unless steps were taken to counteract this bias.

If, however, the same figure is considered from the standpoint of picture transmission it will be clear that the difference is not nearly so striking. An obvious difference between a picture obtained with the in-phase component and one obtained with the resultant is that there is a tendency for a background of light gray to be present in the latter. Secondly, there is less contrast between the blacks and the whites. Both of these effects tend to be eliminated in photographic processes which follow the reception. Moreover, when they are not thus eliminated, they are not readily seen on examining the picture.

The presence of dark current increases the magnitude of the in-phase component as compared to the quadrature component. Since the resultant is equal to the r.m.s. value of these two components, it follows that increasing one component as compared to the other causes the

larger component to approach the resultant. Consequently, adding the dark current causes the resultant to become more like the in-phase component, thus reducing distortion due to the quadrature component.

In half-tone pictures many of the transitions are in small steps. The quadrature component for small steps is frequently small compared to the total in-phase component. By reasoning similar to that in the previous paragraph, it follows that distortion due to quadrature component at small steps, is apt to be negligible. The quadrature effect in half-tones is also reduced by the fact that some of the changes are gradual.

The aperture effect has not been mentioned explicitly above. The aperture effect may be considered as being equivalent to a certain frequency characteristic and it may be assumed that the filter characteristics shown, include it. Incidentally, it is found that the aperture does not greatly affect the relationship between the in-phase and quadrature components.

While it may be expected that the quadrature component should have similar effects in picture transmission and in television, it is perhaps desirable to point out that there are important points of difference such as the presence of motion in the television images and the difference in response characteristics of a television screen and a photographic surface. It is not therefore an inevitable conclusion that television images will be as little affected as picture transmission images by the quadrature component.

Experimental

The conclusions are confirmed by certain experimental transmissions which were made over a picture machine employing a single sideband system as described by Reynolds.[5] The system makes use of 100 lines to the inch and has a total bandwidth of about 1000 cycles. The speed of the spot of light over the picture is about 20 inches per second. Two specimens of printing of different sizes were transmitted. A portion of each specimen one centimeter wide, after transmission, is shown in Fig. 15 enlarged to about five times its original size in order to avoid interference between the half-tone pattern and the picture pattern. Figure 15 should be viewed at about five times the normal reading distance. Group (a) was transmitted on a single sideband basis with the dark current reduced practically to zero. Group (b) was similarly transmitted, excepting that the dark current was 14 db below the maximum current. Group (c) shows a double sideband

[5] Loc. cit.

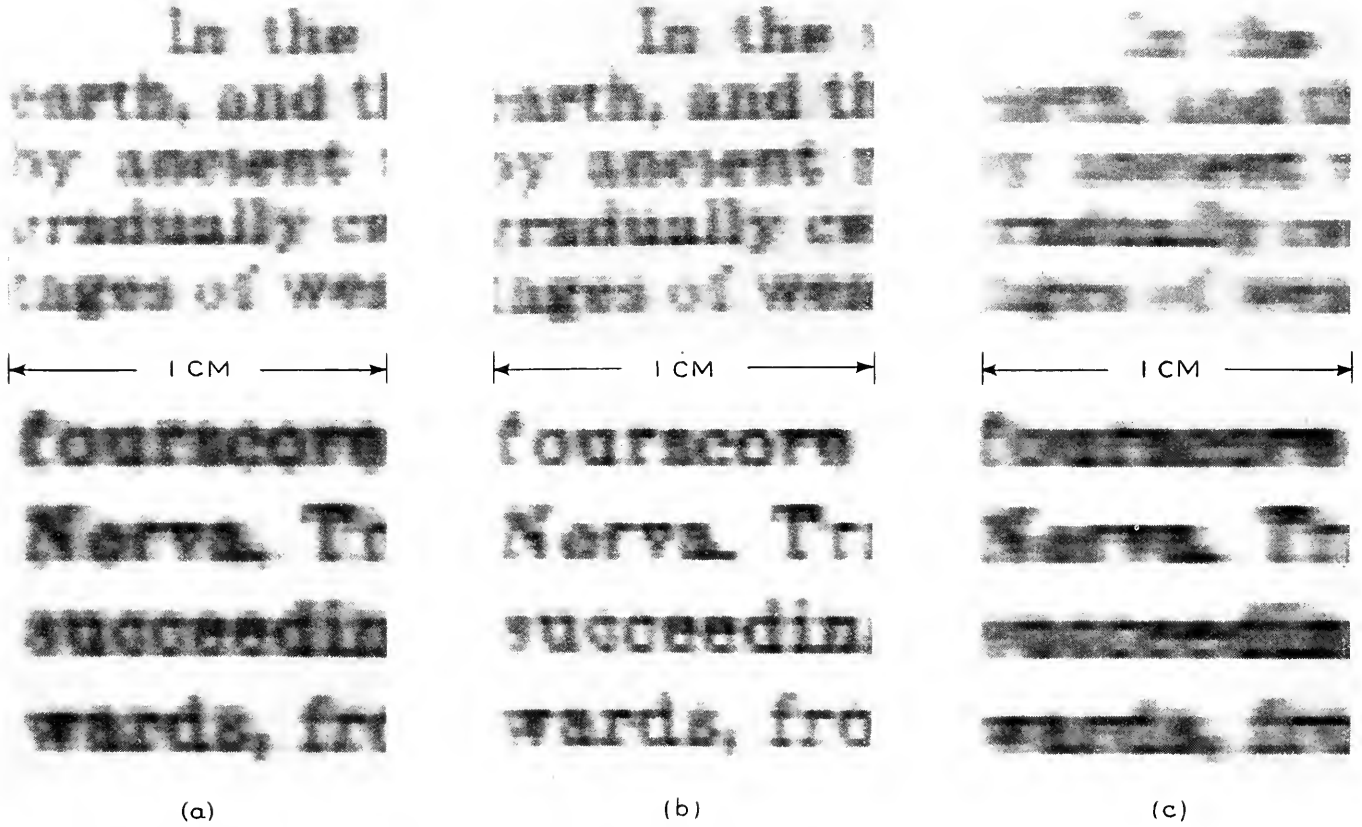


Fig. 15—Enlargements of transmitted printing. (*a*) Carrier at edge of band, dark current negligible. (*b*) Carrier at edge of band, dark current 14 db below maximum. (*c*) Carrier at center of band, dark current negligible. [The same transducer was used for (*a*), (*b*) and (*c*).]

transmission over the same transducer, the carrier being located at the center of the characteristic, the dark current being practically zero. It will be observed that the single sideband transmission gives materially more detail than the double sideband transmission, thus indicating that the presence of the quadrature component is not nearly so serious as a halving of the frequency range. It might perhaps be thought that this unfavorable showing of the double sideband transmission is due to the presence of some special distortion which might be expected in a filter designed for single sideband transmission when used in a manner not intended. On examining the characteristics no such distortion is found.

Acknowledgment

We have had the helpful cooperation of Dr. P. Mertz in this investigation.

Low Temperature Coefficient Quartz Crystals

By W. P. MASON

In this paper a review and amplification are given of the types and characteristics of existing low temperature coefficient crystals. The principal types are the coupled frequency crystals, the long bar crystals, and the AT, BT, CT and DT shear vibrating crystals. The theoretical frequencies for the AT and BT crystals agree well with those calculated from the Christofel formula for the velocity of propagation in an aeolotropic medium. For a finite plate other frequencies appear which are caused by couplings to the flexure and low-frequency shear modes. It is shown that harmonics of the high-frequency shear mode can be excited and will have low temperature coefficients. They can be made to stabilize the frequency of ultra-short-wave oscillators. The properties of the low-frequency CT and DT shear vibrating crystals are described. Overtone vibrations of the shear mode of approximately twice the frequency, having zero temperature coefficients, have been found and these have been labeled the ET and FT cuts.

It is shown that if two or more rotations of the cut are made with respect to the crystallographic axes, a line of zero temperature coefficient high-frequency crystals will be obtained. For the low-frequency shear crystals a surface of zero temperature coefficient crystals should result.

In the last section the variation of frequency with temperature of low coefficient crystals is discussed, and the variation of a new cut, labelled the GT, is described. This cut has zero first and second derivatives of the frequency by the temperature, and as a result has a very constant frequency over a wide temperature range. It has been applied to very constant frequency oscillators and frequency standards and has given a constancy of frequency considerably in excess of that obtained by other low coefficient crystals.

I. Introduction

DURING the past several years a number of crystal plates have been found which have the property that at a specified temperature their frequency will not change with a small change in temperature. These crystals have proved very useful in stabilizing the frequencies of oscillators used in frequency standards, broadcasting stations, radio communication transmitters, airplane transmitters, and for other purposes. In order to bring out their properties and spheres of usefulness a review and amplification of them are given in this paper.

The first types of zero temperature coefficient crystals were the so-called coupled types which obtained their low coefficient by virtue of the interaction between two modes of motion. The first crystal of this type was the "doughnut" crystal invented by W. A. Marrison,[1] which was used in the Bell System frequency standard. In this crystal the principal vibration is a shear and this is coupled to a flexure motion in the ring. The low coefficient is obtained from the fact that the shear has a positive temperature coefficient, while the flexure has a negative coefficient, and due to the coupling there is one region for which the temperature coefficient goes through zero. The next crystal of the coupled type was a Y cut crystal of specified dimensions invented by R. A. Heising.[2] In this crystal a high-frequency shear with a positive temperature coefficient was coupled to a harmonic of a low-frequency flexure, and a zero coefficient resulted at one temperature due to the coupling. Outside of their use in a frequency standard, such coupled types of crystals have not been applied much for commercial purposes on account of the difficulty of adjusting them, the difficulty of mounting them, and the prevalence of spurious frequencies near the desired frequency.

The next low-temperature coefficient crystals were crystals of the long bar type. It has been known for a long time that the temperature coefficient of an X cut crystal with its length lying along the Y or mechanical axis was very low provided the width of the crystal lying along the optic axis is very small compared to the length. This is illustrated by Fig. 1 taken from a former paper [3] which shows that for a crystal whose width is less than 0.15 of its length the temperature coefficient is about 2 parts per million per degree centigrade. Furthermore, it was found by the writer in 1930 [3] that if the thickness of the crystal laying along the X or electrical axis was increased the temperature coefficient was decreased and in fact for certain ratios of axes the coefficient approached zero. For a bar of square cross section the zero coefficient occurs when the ratio of width to length is approximately 0.272. This apparently is also the method for obtaining a low-temperature coefficient used in the Hilger resonator. The second harmonic of this vibration has been used in the frequency standards of the Physikalisch-Technische Reichsanstaldt.[4] In their standards

[1] "A High Precision Standard of Frequency," W. A. Marrison, *Proc. I. R. E.*, April 3, 1929.

[2] This crystal is described by F. R. Lack in "Observation on Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates," *Proc. I. R. E.*, July 1929, Vol. 17, pp. 1123–1141, and Patent No. 1,958,620 issued May 15, 1934.

[3] "Electrical Wave Filters Employing Quartz Crystals as Elements," *B. S. T. J.*, July 1934, Pages 411 and 412.

[4] A. Scheibe and V. Adelsberger, *Ann. d. Phys.* 18, 1, 1933.

the length is cut along the X axis and the vibration is excited by fields applied along the length of the bar. Since a rotation about the optic or Z axis does not change the properties of the elastic constants involved in this vibration, this bar should have a zero temperature coefficient at about the same ratio of axes as that given above. The zero angle of orientation is, however, not the most favorable angle of orientation for the fundamental vibration of a long bar, for if the length of the crystal lies at an angle of $+ 5°$ with respect to the Y or mechanical axis, the coefficient of a long bar is nearly zero.[5] These

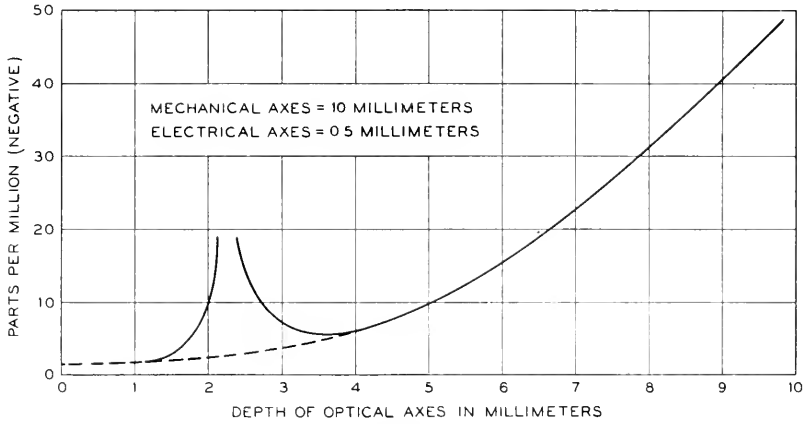


Fig. 1—Temperature coefficient of a perpendicularly cut crystal for varying ratios of width to length.

long bar type crystals have been used to a small extent to control oscillators and to stabilize the pass bands of filters. Their small use is attributable to the fact that they vibrate at low frequencies and are difficult to excite in an oscillator circuit.

The AT and BT high-frequency shear crystals and the CT and DT low-frequency shear crystals are other low temperature coefficient crystals and they are discussed in detail in section II. These crystals are cut with their planes at specified angles with respect to the crystallographic axes and all of them involve a single rotation about an axis which is parallel or approximately parallel to one of the crystallographic axes. It is shown in section III that such crystals are not the only zero coefficient crystals of these types that can be obtained, for if we allow three rotations about the crystallographic axes a whole surface of zero temperature coefficient crystals can be found. These crystals

[5] Matsumara and Kansaki, "On the Temperature Coefficient of Frequency of Y Waves in X Cut Quartz Plates," *Reports of Radio Researches and Works in Japan*, March 1932.

are more difficult to cut than the standard crystals and are more subject to couplings to other modes of motion and hence most of them are probably of more theoretical interest than of practical value.

All of the zero coefficient crystals described above are zero coefficient at a specified temperature only and for temperatures on either side of the specified temperature the frequency usually increases or decreases in a parabolic curve with temperature. This merely expresses the fact that the frequency-temperature curve is not exactly linear, but must be expressed more generally in a series of powers of the temperature. Then for all the crystals considered above, the first derivative of the frequency by the temperature is zero at the specified temperature T_0. The next term of importance is the square term and hence most crystals have a frequency which varies as the square term of the temperature about the zero coefficient temperature T_0. A crystal cut, labelled the *GT* crystal, has recently been found for which both the first and second derivatives of the frequency by the temperature are zero. As a result this "*GT*" crystal has a very constant frequency over a very wide temperature range, and in fact does not vary by more than one part in a million for a temperature range of 100° centigrade. For a temperature range of $\pm$ 15° C. it can be adjusted so that it does not vary by more than one part in ten million. This crystal has been applied to portable and fixed frequency standards and has given a constancy of frequency considerably in excess of any other piezo-electric crystal used under the same conditions. It has also been applied in quartz crystal filters to give pass bands which do not vary appreciably with temperature.

II. Standard Zero Temperature Coefficient Crystals

AT and BT Zero Temperature Coefficient Crystals

Crystals which employ the characteristics of a single shear mode of vibration to obtain a zero temperature coefficient are the *AT* and *BT* cut crystals.[6] These crystals vibrate in shear and their frequencies are determined principally by the thickness of the quartz plate. Their mode of vibration is similar to the ordinary *Y* cut, and they obtain their zero coefficient from the fact that the temperature coefficient of the shear mode changes from positive to negative as the angle of cut is rotated about the *X* axis by positive or negative angles from the position of the *Y* cut crystal. Figure 2 shows the method of cutting these plates from the natural crystal. Figure 3 shows the temperature coefficient of these crystals plotted against the angle of

[6] "Some Improvements in Quartz Crystal Circuit Elements," F. R. Lack, G. W. Willard, and I. E. Fair, *B. S. T. J.*, July 1934, pp. 453–463.

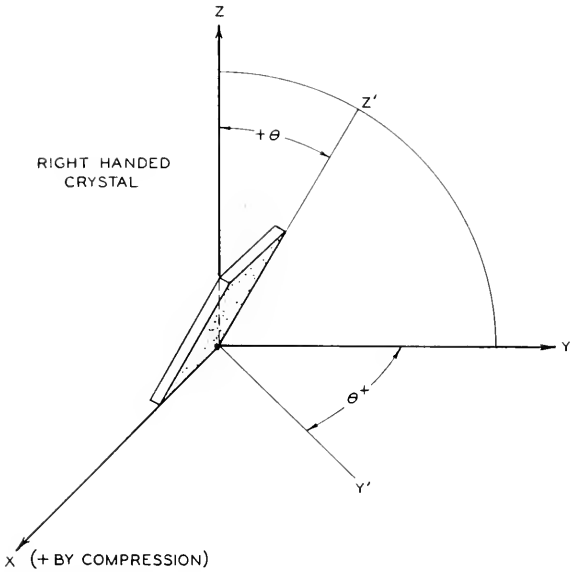


Fig. 2—Diagram illustrating angles used in expressing orientation of AT and BT plates within the natural crystal.

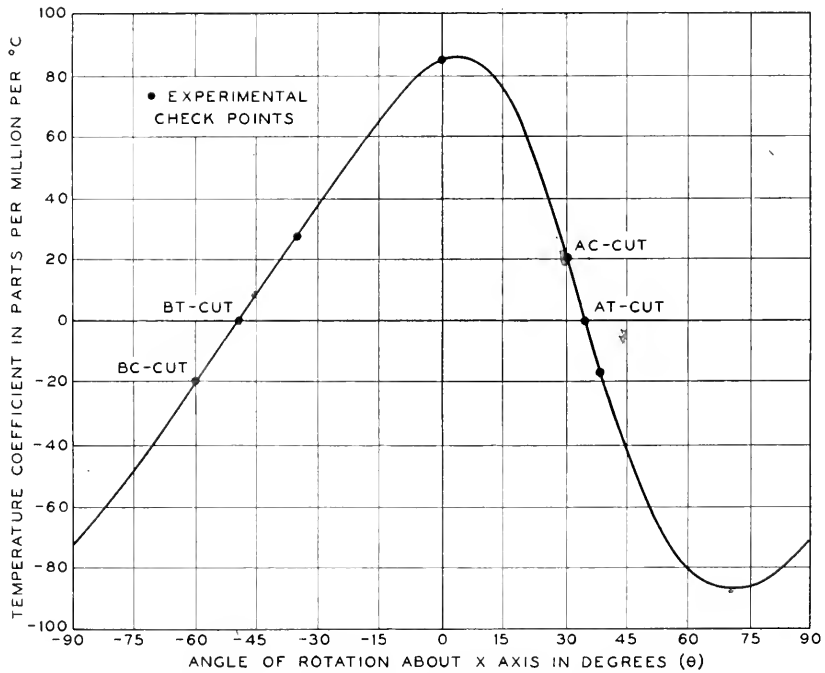


Fig. 3—Temperature coefficient for thin plates plotted as a function of the angle of cut.

cut while Fig. 4 shows the frequency constant of the crystal; i.e., the kilocycles for one millimeter thickness plotted as a function of the angular cut. The AT crystal which occurs at an orientation of $+ 35° - 20'$ has a frequency constant of 1662 kilocycles for one millimeter thickness while the BT cut which occurs at $- 49°$ has a frequency constant of 2465 kilocycles for one millimeter thickness.

The frequency curve of Fig. 4 agrees very closely with the frequency calculated from the elastic constants used in the formula for the velocity of propagation of an aeolotropic medium given by E. B. Christofel.[7] Christofel showed that for any direction of propagation in an elastic solid, there were three different waves whose velocity of propagation could be obtained from the determinant

$$\begin{vmatrix} \lambda_{11} - \rho c^2, & \lambda_{12}, & \lambda_{13} \\ \lambda_{12}, & \lambda_{22} - \rho c^2, & \lambda_{23} \\ \lambda_{13}, & \lambda_{23}, & \lambda_{33} - \rho c^2 \end{vmatrix} = 0. \tag{1}$$

In this equation ρ is the density, c the velocity of propagation, and λ's are related to the elastic constants of the crystal by the formulae

$$\begin{aligned}
\lambda_{11} &= c_{11}l^2 + c_{66}m^2 + c_{55}n^2 + 2c_{56}mn + 2c_{15}nl + 2c_{16}lm, \\
\lambda_{12} &= c_{16}l^2 + c_{26}m^2 + c_{45}n^2 + (c_{46} + c_{25})mn \\
&\qquad + (c_{14} + c_{56})nl + (c_{12} + c_{66})lm, \\
\lambda_{13} &= c_{15}l^2 + c_{46}m^2 + c_{35}n^2 + (c_{45} + c_{36})mn \\
&\qquad + (c_{13} + c_{55})nl + (c_{14} + c_{56})lm, \\
\lambda_{23} &= c_{56}l^2 + c_{24}m^2 + c_{34}n^2 + (c_{44} + c_{23})mn \\
&\qquad + (c_{36} + c_{45})nl + (c_{25} + c_{16})lm, \\
\lambda_{22} &= c_{66}l^2 + c_{22}m^2 + c_{44}n^2 + 2c_{24}mn + 2c_{46}nl + 2c_{26}lm, \\
\lambda_{33} &= c_{55}l^2 + c_{44}m^2 + c_{33}n^2 + 2c_{34}mn + 2c_{35}nl + 2c_{45}lm,
\end{aligned} \tag{2}$$

where l, m, and n are respectively the direction cosines between the direction of propagation and the x, y, and z axes. For quartz

$$c_{22} = c_{11};\quad c_{24} = -c_{14};\quad c_{55} = c_{44};\quad c_{56} = c_{14};\quad c_{66} = (c_{11} - c_{12})/2$$

and

$$c_{15} = c_{16} = c_{25} = c_{26} = c_{34} = c_{35} = c_{36} = c_{45} = c_{46} = 0. \tag{3}$$

For a rotation about the x axis for which a positive angle is measured in a counter clockwise rotation for a left handed crystal and a clockwise direction for a right handed crystal when an electrically positive face

[7] See Love's "Theory of Elasticity," page 298, fourth edition.

(determined by a compression) is up, the values of l, m, and n are

$$l = 0; \qquad m = \cos\theta; \qquad n = -\sin\theta. \tag{4}$$

In this definition, a right handed crystal is taken as one which causes the plane of polarization of light traveling along the Z or optic axis to

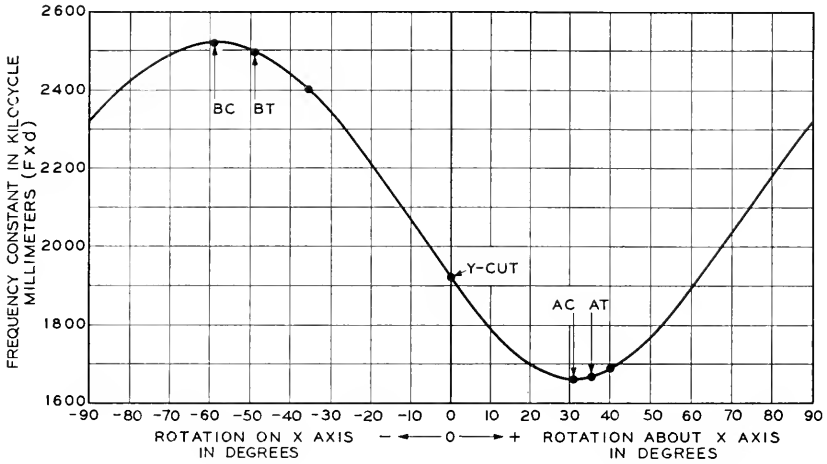


Fig. 4—Frequency constant for thin plates plotted against angle of cut.

rotate in the sense of a right handed screw. Substituting the values of (3) and (4) in (2) we find

$$\begin{aligned}
\lambda_{11} &= c_{66}\cos^2\theta + c_{41}\sin^2\theta - 2c_{14}\sin\theta\cos\theta = c_{66}', \\
\lambda_{23} &= -c_{14}\cos^2\theta - (c_{44} + c_{23})\sin\theta\cos\theta, \\
\lambda_{22} &= c_{22}\cos^2\theta + c_{44}\sin^2\theta + 2c_{14}\sin\theta\cos\theta, \\
\lambda_{33} &= c_{44}\cos^2\theta + c_{33}\sin^2\theta, \\
\lambda_{12} &= \lambda_{13} = 0.
\end{aligned} \tag{5}$$

With these values of λ, the three solutions of equation (1) are

$$c_1 = \sqrt{\frac{\lambda_{11}}{\rho}};$$

$$c_{2,3} = \sqrt{\frac{1}{2}\left[\frac{\lambda_{22}}{\rho} + \frac{\lambda_{33}}{\rho} \pm \sqrt{\left(\frac{\lambda_{22}}{\rho} - \frac{\lambda_{33}}{\rho}\right)^2 + 4K^2\frac{\lambda_{22}}{\rho}\frac{\lambda_{23}}{\rho}}\right]}, \tag{6}$$

where $K^2 = \lambda_{23}^2/\lambda_{22}\lambda_{33}$.

The frequency of any plate with its edges free to move will be

$$f = \frac{c}{2t}(2n + 1) \qquad n = 0, 1, 2, \cdots, \tag{7}$$

where t is the thickness of the plate. Hence for the A type vibration which corresponds to the first velocity c_1, the frequency will be

$$f_1 = \frac{1}{2t}\sqrt{\frac{c_{66}\cos^2\theta + c_{44}\sin^2\theta - 2c_{14}\sin\theta\cos\theta}{\rho}} = \frac{1}{2t}\sqrt{\frac{c_{66}'}{\rho}}. \quad (8)$$

The solid curve of Fig. 4 shows a plot of this equation while the measured values are shown by dots.

The frequencies of the other two modes of motion are given by

$$f_{2,3}^2 = \tfrac{1}{2}[f_A^2 + f_B^2 \pm \sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}],$$

where

$$f_A = \frac{1}{2t}\sqrt{\frac{\lambda_{22}}{\rho}}; \qquad f_B = \frac{1}{2t}\sqrt{\frac{\lambda_{33}}{\rho}}; \qquad K = \frac{\lambda_{23}}{\sqrt{\lambda_{22}\lambda_{33}}}. \quad (9)$$

This formula is the same as that for the frequencies given by two coupled modes [8] and hence can be interpreted as a mode of vibration, determined by λ_{33}, and a mode of vibration, determined by the constant λ_{22}, coupled together through the coupling compliance λ_{23}. For an isotropic medium one of these modes would be a pure shear and the other a longitudinal mode, but in a crystalline medium the motions are not strictly along or perpendicular to the direction of motion The A type vibration which is an x_y' shear vibration is not coupled to the other two since the coupling elasticities λ_{12} and λ_{13} are equal to zero. For a more general rotation, however, they will not necessarily be equal to zero and hence the general solution of equation (1) will represent two shear like vibrations, the x_y' and y_z', and a nearly longitudinal y_y' vibration all mutually coupled together.

The Christofel formula is only valid for a plate of thickness t which extends to infinity in all other directions and hence this solution does not show the coupled frequencies due to the contour dimensions which occur in a finite plate. In general there are two types of vibration which couple strongly to the A type vibration, the low-frequency shear modes and the flexure modes in which bending occurs in the xy' plane. As pointed out by Lack, Willard and Fair,[9] both the AT and the BT occur near angles of cut for which the coupling to the z_x' low frequency shear mode vanishes. Hence one would expect that these crystals would have fewer subsidiary resonances and this expectation is verified by experiment. A practical result is that the

[8] "Electrical Wave Filters Employing Quartz Crystals as Elements," W. P. Mason, July 1934, page 444, *B. S. T. J.*

[9] "Some Improvements in Quartz Crystal Circuit Elements," *B. S. T. J.*, July 1934. The problem of couplings is discussed in more detail in the U. S. Patent 2,173,589, Sept. 19, 1939 issued to R. A. Sykes and the writer. In this patent the AC cut and the $-18.5°$ X cut crystals are described.

AT and BT crystals can control considerably more powerful oscillators without danger of the crystals breaking than can the X or Y cut crystals. The frequency spectrum of an AT cut plate ground down from a large ratio of dimensions to a smaller one is shown on Fig. 5.

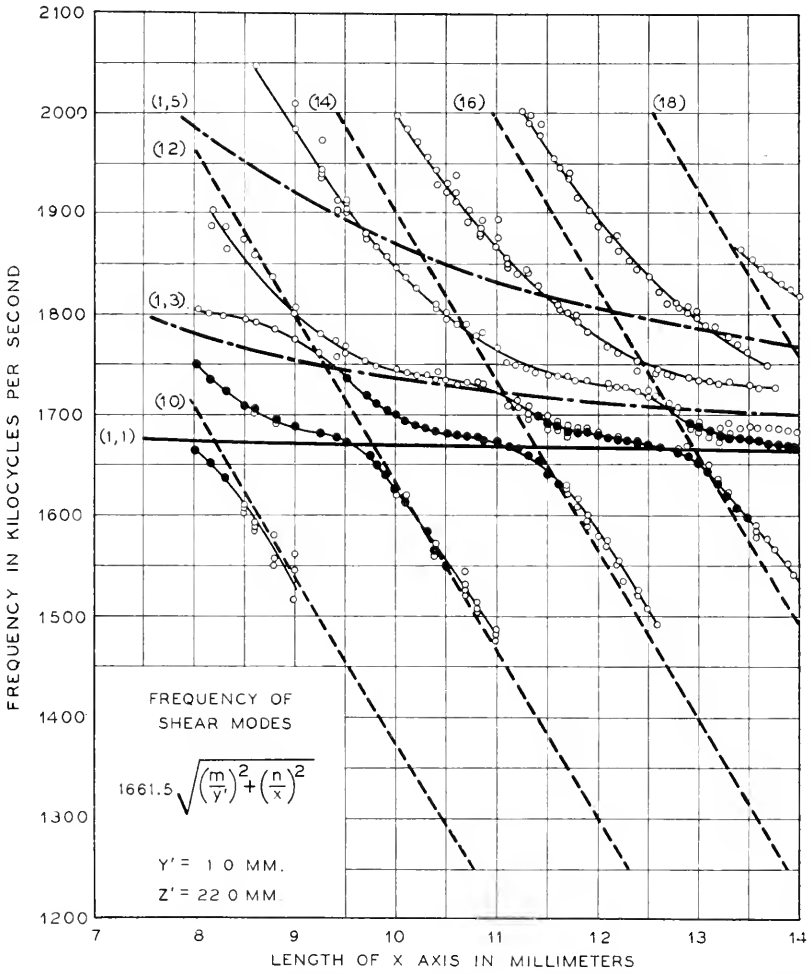


Fig. 5—Frequency spectrum (dots) for an AT plate as a function of ratio of length to thickness. The dashed lines represent calculated flexural vibrations. The whole line is the principal shear mode. The dot dash lines are other shear modes.

Most of the prominent frequencies can be identified as shear frequencies of the type discussed in a previous paper [10] and harmonics of flexure

[10] "Electrical Wave Filters Employing Quartz Crystals as Elements," W. P. Mason, *B. S. T. J.*, July 1934, page 446. The verification was made by R. A. Sykes who kindly supplied Fig. 5.

vibrations. This figure shows clearly that the strongest flexures entering are controlled by the length of the X axis rather than the Z' axis.

As shown by equations (6), (7) and (8) the AT and BT cut crystals have odd harmonic vibrations which are controlled by the same elastic constants as the fundamental vibrations. Since they are controlled by the same elastic constants, the harmonic vibrations have the same temperature coefficients as the fundamental mode and hence will have nearly zero coefficients. This property has been made use of in oscillators in controlling high-frequency vibrations with crystals whose thicknesses can be obtained commercially.

CT and DT Low-Frequency Zero Temperature Coefficient Crystals

Another set of zero temperature coefficient crystals which are particularly useful for low frequencies has recently been described by Hight and Willard.[11] They are related to the AT and BT cuts discussed above in that they use the same shearing motion to produce the low coefficient. This relation is illustrated by Fig. 6 which shows

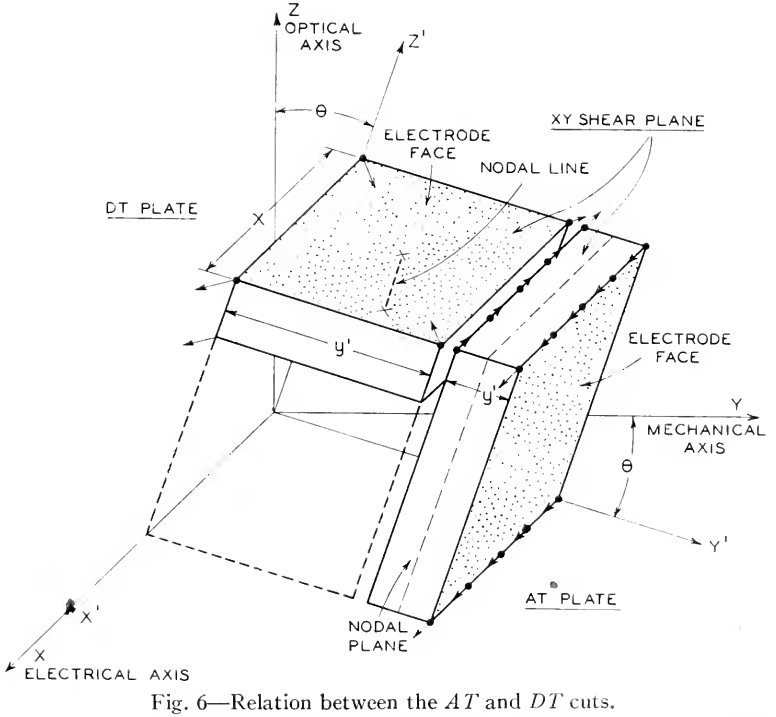


Fig. 6—Relation between the AT and DT cuts.

[11] "Presented before the Institute of Radio Engineers, March 3, 1937. Published in *I. R. E. Proc.* May, 1937, p. 549. Similar crystals are also discussed in U. S. Patents 2,111,383 and 2,111,384 issued to S. A. Bokovoy.

the approximate orientations of the AT cut and the DT cut. In the AT plate the x_y' strain is produced by a shear mode of vibration as shown by the arrows which represent instantaneous displacements. In the DT plate the x_y' strain is produced by a shear mode of vibration as shown again by the arrows. Two diagonally opposite corners move radially outward while the other two move radially inward. The relatively low frequency of the DT plate results from the relatively large frequency-determining dimensions x and y'. The temperature coefficient of frequency of these plates may be made zero, for the proper angles of cut, since it goes from a large positive value at one orientation to a large negative value for an orientation 90 degrees from the first. Actually the angle of cut of the DT plate is not exactly 90 degrees from the AT. This is due to the fact that the frequency

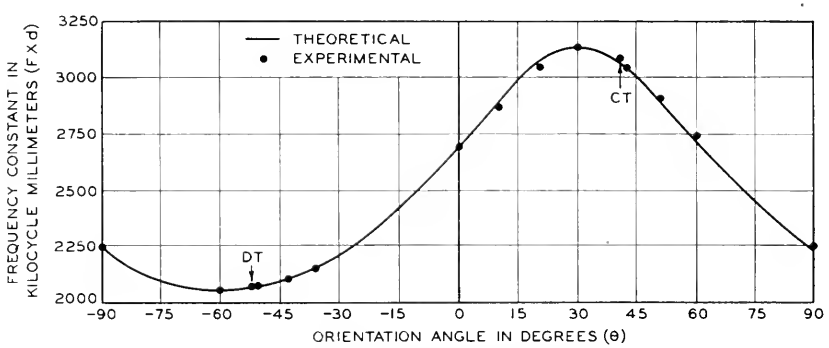


Fig. 7—Frequency constant for low-frequency shear crystal plotted against angle of cut.

for a square plate involves the s_{66}' constant rather than the c_{66}' constant which controls the frequency of a thin plate. Similarly we find that there is a crystal almost 90° from the BT which has a zero coefficient and this has been designated the CT.

Figure 6 shows that the electrode faces of the DT crystal are placed on the $z'x$ plane and hence the shear mode generated would ordinarily be called the z_x' mode even though it is similar to the x_y' shear mode in the AT crystal at right angles to it. The measured frequency constant of such a series of square plates is shown on Fig. 7. In the absence of a complete theoretical solution [12] taking account of all the elastic couplings for a square plate vibrating in shear, an empirical

[12] An approximate solution neglecting coupling was given in a former paper "Electrical Wave Filters Employing Quartz Crystals as Elements," page 446. This solution is not complete enough, however, to allow calculations of temperature coefficients with very great accuracy.

formula was developed for the frequency which is

$$f = \frac{1.25}{2d}\sqrt{\frac{1}{\rho s_{55}'}}, \tag{10}$$

where $d = x = z'$ if the plate is square and $d = (x + z')/2$ if only nearly square. The elastic constant s_{55}' depends on the orientation angle θ according to the equation.

$$s_{55}' = s_{44}\cos^2\theta + s_{66}\sin^2\theta + 4s_{14}\sin\theta\cos\theta. \tag{11}$$

Figure 7 shows the measured values of frequency and the values calculated from equations (10) and (11). Agreement is obtained within 2 per cent.

From equations (10) and (11) the temperature coefficient of frequency of a shear vibrating plate should be for a square crystal

$$T_f = -(1/2)\left[T_x + T_z' + T_\rho + \frac{s_{44}T_{s_{44}}\cos^2\theta + s_{66}T_{s_{66}}\sin^2\theta + 4s_{14}T_{s_{14}}\sin\theta\cos\theta}{s_{44}\cos^2\theta + s_{66}\sin^2\theta + 4s_{14}\sin\theta\cos\theta}\right]. \tag{12}$$

The temperature coefficient of length along the optic axis is about 7.8 parts per million (per degree centigrade) while that perpendicular to the optic axis is 14.3 parts per million. For any other direction

$$T_l = 7.8 + 6.5\cos^2\theta, \tag{13}$$

where θ is the angle between the length and the optic axis. Hence

$$T_x = 14.3; \qquad T_z' = 7.8 + 6.5\cos^2\theta,$$

and

$$T_\rho = -36.4 \text{ per degree C.} \tag{14}$$

The temperature coefficients of the six elastic constants were evaluated in a former paper.[13] Since then they have been slightly revised so that the best values now are

$T_{s_{11}} = +12,$		$T_{c_{11}} = -54.0,$
$T_{s_{12}} = -1{,}265,$	this	$T_{c_{12}} = -2{,}350,$
$T_{s_{13}} = -238,$	results in	$T_{c_{13}} = -687,$
$T_{s_{14}} = +123,$		$T_{c_{14}} = +96,$
$T_{s_{33}} = +213,$		$T_{c_{33}} = -251,$
$T_{s_{44}} = +189,$		$T_{c_{44}} = -160,$
$T_{s_{66}} = -133.5,$		$T_{c_{66}} = +161.$

(15)

[13] "Electric Wave Filters Employing Quartz Crystals as Elements," W. P. Mason, *B. S. T. J.*, 13, p. 446, July 1934.

Using these values in equation (12) the expected temperature coefficients for the low-frequency vibration are as shown on Fig. 8 The measured points are shown on the curve. The zero temperature coefficients occur at the angles + 38° and − 53°. These crystals have been designated the *CT* and *DT* low-frequency shear crystals. These types of crystals are useful for stabilizing low-frequency oscillators ranging from 50 *KC* to 500 *KC*.

Just as the *AT* and *BT* crystals have harmonics which can be used to control oscillator frequencies, so also do over-tones of the low-frequency shear crystals exist. They do not bear, however, the simple

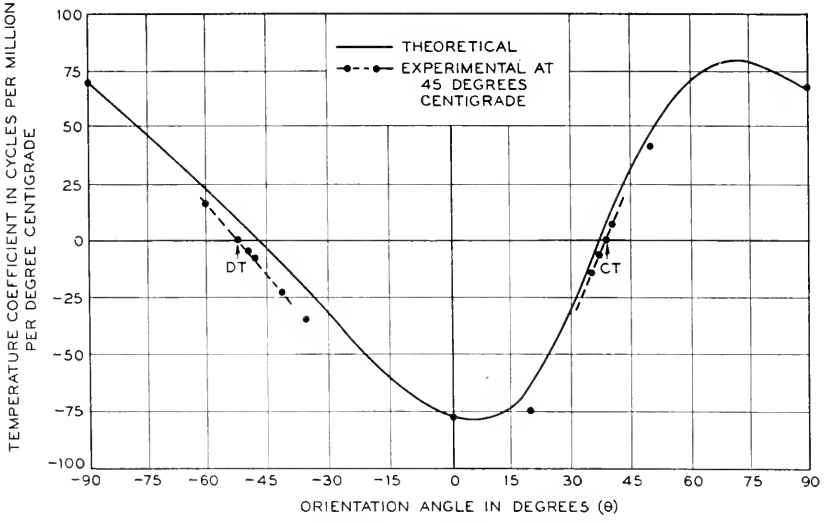


Fig. 8—Calculated and measured temperature coefficients for low-frequency shear crystals.

relation to the fundamental that the high-frequency harmonics do. S. C. Hight has found a mode of motion, which is probably related to the second flexural vibration, of nearly twice the frequency of the low-frequency shear mode and which has zero temperature coefficients at angles of + 66°-30′ and − 57°. These crystals have been designated respectively as the *ET* and *FT* crystal cuts. Figure 9 shows a plot of this frequency versus orientation. It will be observed that the frequency constant of this mode of motion is about twice that for the low-frequency shear mode and hence these crystals can be obtained in reasonable sizes for twice the frequencies that the *CT* and *DT* crystals can be obtained.

Practically all the work done has been on square or nearly square plates. Some time ago Bechmann [14] and Koga [15] published work done on crystals which departed from the square shape for which zero coefficients were obtained at somewhat different angles and different frequencies than those given for the CT and DT crystals. This is due to the fact that when the crystal shape departs from the square, the frequency approaches more nearly the resonant frequency of the crystal vibrating in its second flexure mode and the increased coupling changes the angle for which the coefficient becomes zero. The square crystal is the one which has fewer secondary frequencies and is therefore more desirable.

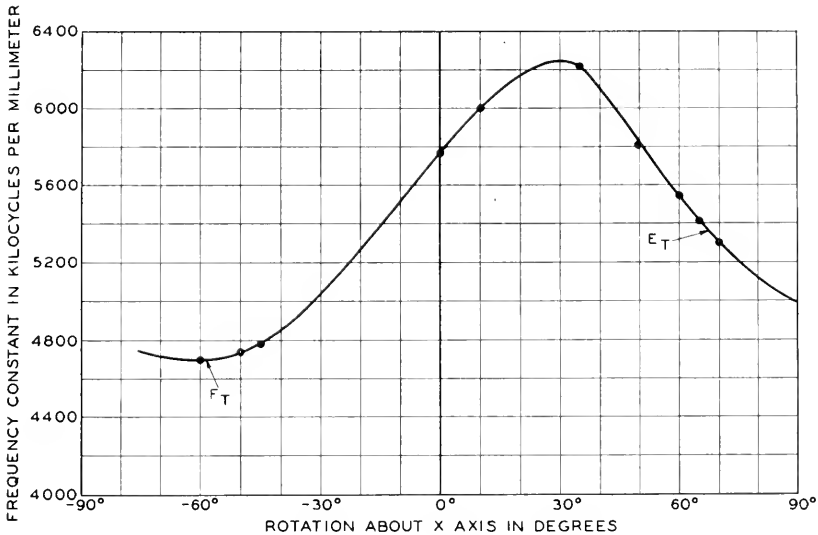


Fig. 9—Frequency constant for E and F type vibrations.

III. Zero Temperature Coefficient Crystals for More General Orientations

Shortly after the discovery of the AT and BT crystals it was realized that zero temperature coefficient crystals could be obtained at a variety of angles provided two rotations of the crystal with respect to the crystallographic axes were used. This would allow the direction of the shearing axis to point in any direction with respect to the crystallographic axes. Using the c_{66}' constant as the elastic constant determining the frequency, it was found that there was a whole series of zero

[14] R. Bechmann, *Hochfrequenstechnik u Elektroakustic*, Vol. 44, No. 5, p. 145.
[15] I. Koga, *Report of Radio Research in Japan*, Vol. IV, No. 2, 1934. See also Patents 2,111,383 and 2,111,384 issued to S. A. Bokovoy.

temperature coefficient crystals whose plot as a function of the two rotations would be a line in which the AT and BT cuts would be in the region of two points on the line. A few of these crystals whose angles were in the region of the AT crystal were measured and were found to have zero coefficients but also had a much more complicated frequency spectrum than the AT or BT crystals when cut to have their major faces more nearly parallel to the x axis.

Recently Bechmann [16] has made calculations and experiments in respect to double orientation crystals which have zero temperature coefficients. The calculations were made by means of the Christofel formula of equations (1) and (2). Although this gives the same result as that calculated from the constant c_{66}' for rotations around the x axis, it differs from it somewhat for more general rotations. If we expand equation (1) we obtain the cubic equation for the frequency of oscillation.

$$f^6 - f^4(f_A^2 + f_B^2 + f_C^2) + f^2[f_A^2 f_B^2(1 - K_{AB}^2) + f_A^2 f_C^2(1 - K_{AC}^2) + f_B^2 f_C^2(1 - K_{BC}^2)] - f_A^2 f_B^2 f_C^2(1 - K_{AB}^2 - K_{AC}^2 - K_{BC}^2 + 2K_{AB}K_{AC}K_{BC}) = 0, \quad (16)$$

where

$$f = \frac{c}{2t}; \qquad f_A = \frac{\sqrt{\lambda_{11}/\rho}}{2t}; \qquad f_B = \frac{\sqrt{\lambda_{22}/\rho}}{2t}; \qquad f_C = \frac{\sqrt{\lambda_{33}/\rho}}{2t};$$

$$K_{AB} = \frac{\lambda_{12}}{\sqrt{\lambda_{11}\lambda_{22}}}; \qquad K_{AC} = \frac{\lambda_{13}}{\sqrt{\lambda_{11}\lambda_{33}}}; \qquad K_{BC} = \frac{\lambda_{23}}{\sqrt{\lambda_{22}\lambda_{33}}}.$$

f_A, f_B, f_C can be interpreted as the three primary frequencies and would correspond to the three solutions of (16) if the couplings K_{AB}, etc., were zero. The three solutions of (16) then will be these three primary modes modified by the coupling between them. If we let

$$P = [(f_A^2 + f_B^2 + f_C^2) - 3[f_A^2 f_B^2(1 - K_{AB}^2) + f_A^2 f_C^2(1 - K_{AC}^2) + f_B^2 f_C^2 (1 - K_{BC}^2)]]/9, \quad (17)$$

$$Q = [2(f_A^2 + f_B^2 + f_C^2)^3 - 9(f_A^2 + f_B^2 + f_C^2) \times [f_A^2 f_B^2(1 - K_{AB}^2) + f_A^2 f_C^2(1 - K_{AC}^2) + f_B^2 f_C^2(1 - K_{BC}^2)] + 27 f_A^2 f_B^2 f_C^2(1 - K_{AB}^2 - K_{AC}^2 - K_{BC}^2 + 2K_{AB}K_{AC}K_{BC})]/54,$$

[16] "Researches on Natural Elastic Vibrations of Piezo-Electrically Excited Quartz Plates," R. Bechmann, *Zeit. f. Technisch Physik*, Vol. 16, No. 12, 1935, pp. 525–528. This multiple orientation of high-frequency shear crystals is also the basis of the V cut crystal of Bokovoy and Baldwin discussed for example in British Patent No. 457,342 issued May 27, 1936.

and set

$$\cos \psi = \frac{Q}{P^{3/2}},$$

the three solutions will be

$$f_1 = \sqrt{2\sqrt{P}\cos\frac{\psi}{3} + \frac{(f_A^2 + f_B^2 + f_C^2)}{3}},$$

$$f_{2,3} = \sqrt{-2\sqrt{P}\cos\left(\frac{\psi}{3} \pm \frac{\pi}{3}\right) + \frac{(f_A^2 + f_B^2 + f_C^2)}{3}}, \qquad (18)$$

From these equations and equation (2), the frequencies and temperature coefficients of all three modes of motion have been calculated by Bechmann. Based on these calculations the angles of zero coefficient

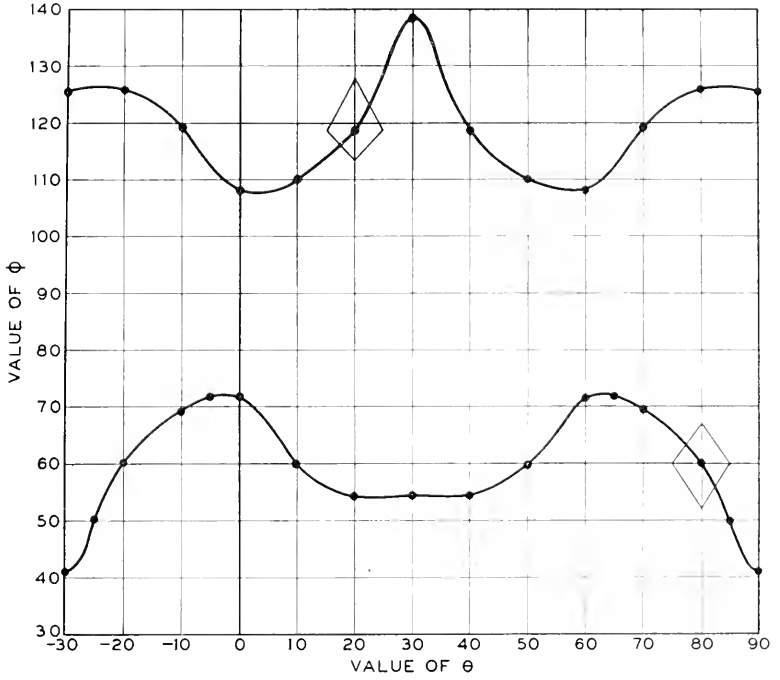


Fig. 10—Angles of cut for zero temperature coefficient high-frequency shear crystals for two rotations.

are shown on Fig. 10 for the angular placement of the direction of propagation adopted on Fig. 11.

Using the empirical formula (10) for the low-frequency shear vibration a surface of zero coefficient low-frequency shear vibrating crystals can be calculated.[17] For this crystal three angles are required to

[17] Multiple orientation low- and high-frequency shear crystals are discussed in British Patent 491,407 issued to the writer on September 1, 1938.

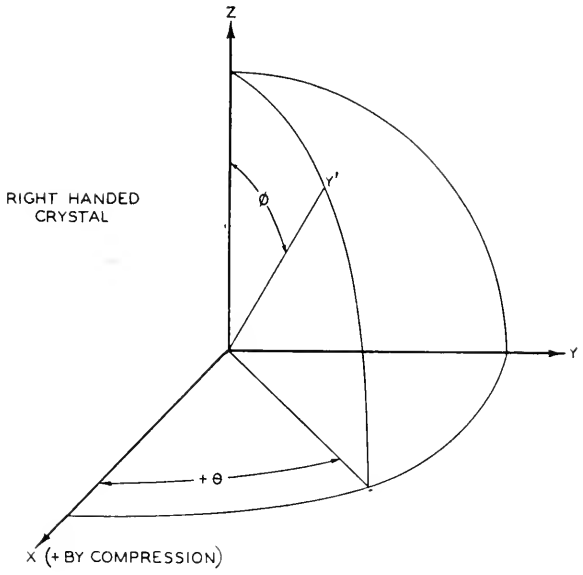


Fig. 11—Angular system for locating the axis of shear of high-frequency crystals with two rotations.

specify the position of the plate since, for a low-frequency shear crystal, rotating the plate around its shearing axis will change the s_{55}' constant and hence the frequency and temperature coefficient of the plate. If we let the position of the plate with respect to the crystalline axes be denoted by the angles, θ, φ and γ, measured as shown on Fig. 12 it can be

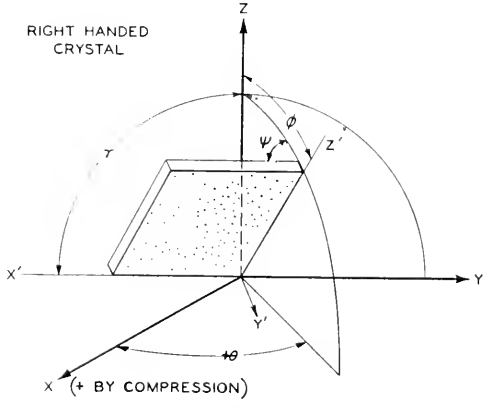


Fig. 12—Angular system for locating low-frequency shear crystals with three rotations.

shown that the s_{55}' constant is given by the equation

$$\begin{aligned} s_{55}' = {} & (s_{11} - 2s_{13} + s_{33}) \cos^2\psi \sin^2 2\varphi + s_{66} \sin^2\varphi \sin^2\psi \\ & + 4s_{14} \sin\varphi [\sin 3\theta \cos\varphi (\cos^2\psi \cos^2\varphi - \sin^2\psi) \\ & + \cos 3\theta \sin\psi \cos\psi (\cos 2\varphi + \cos^2\varphi)] \\ & + s_{44} (\cos^2\psi \cos^2 2\varphi + \sin^2\psi \cos^2\varphi), \end{aligned} \quad (19)$$

where the angle γ is given in terms of a new angle ψ and φ by the equation

$$\cos\gamma = \sin\varphi \cos\psi. \quad (20)$$

If we introduce this expression into equation (12) and introduce the numerical values of equation (15), the expression for the temperature coefficient of a low-frequency shear crystal cut at any angle becomes

$$\begin{aligned} T_f = \Bigg[& 4.5 + 2.9 (\sin^2\varphi \cos^2\psi + \cos^2\varphi) + \Bigg[\frac{-5877.5 \cos^2\psi \sin^2 2\varphi}{195 \cos^2\psi \sin^2 2\varphi} \\ & + 15790 \sin^2\varphi \sin^2\psi + 10340 \sin\varphi [\sin 3\theta \cos\varphi (\cos^2\psi \cos 2\varphi - \sin^2\psi) \\ & + 292.8 \sin^2\varphi \sin^2\psi - 172.4 \sin\varphi [\sin 3\theta \cos\varphi (\cos^2\psi \cos 2\varphi - \sin^2\psi) \\ & + \cos 3\theta \sin\psi \cos\psi (\cos 2\varphi + \cos^2\varphi)] \\ & + \cos 3\theta \sin\psi \cos\psi (\cos 2\varphi + \cos^2\varphi)] \\ & - 19{,}525 (\cos^2\psi \cos^2 2\varphi + \sin^2\psi \cos^2\varphi) \Bigg] \Bigg] \cdot \\ & + 200.5 (\cos^2\psi \cos^2 2\varphi + \sin^2\psi \cos^2\varphi) \Bigg] \end{aligned} \quad (21)$$

Figure 13 gives a contour map of the location of the angles of zero temperature coefficient. The dotted lines indicate the paths for which the piezo-electric constant is a maximum and hence for which the crystal is most easily excited.

IV. A New Crystal Cut, Labeled the GT Crystal, Which Has A Very Constant Frequency for A Wide Temperature Range

All of the zero temperature coefficient crystals so far obtained have a zero temperature coefficient only for a specified temperature, while on either side of this temperature the frequency either increases or decreases in a parabolic curve with the temperature. This is well illustrated by Fig. 14 which shows a comparison of the frequency stability of the standard zero temperature coefficient crystals over a wide temperature range. What is plotted is the number of cycles change in a million from the zero coefficient temperature. These curves show that for a 50° C. change from the zero coefficient temperature the frequency of standard zero temperature coefficient crystals may change from 30 to 140 parts per million. The curves are usually nearly para-

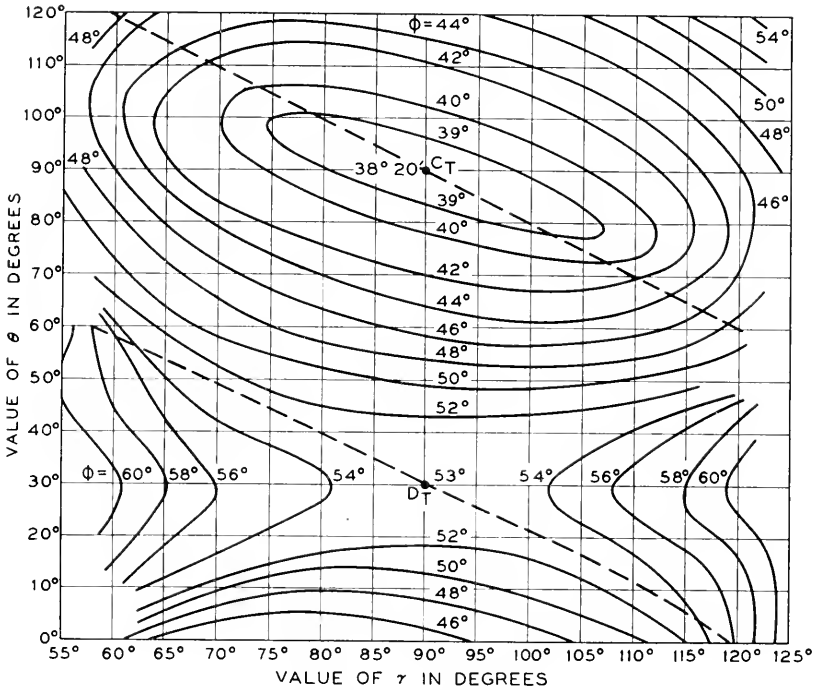


Fig. 13—Contour map of zero temperature coefficient low-frequency shear crystals with three rotations.

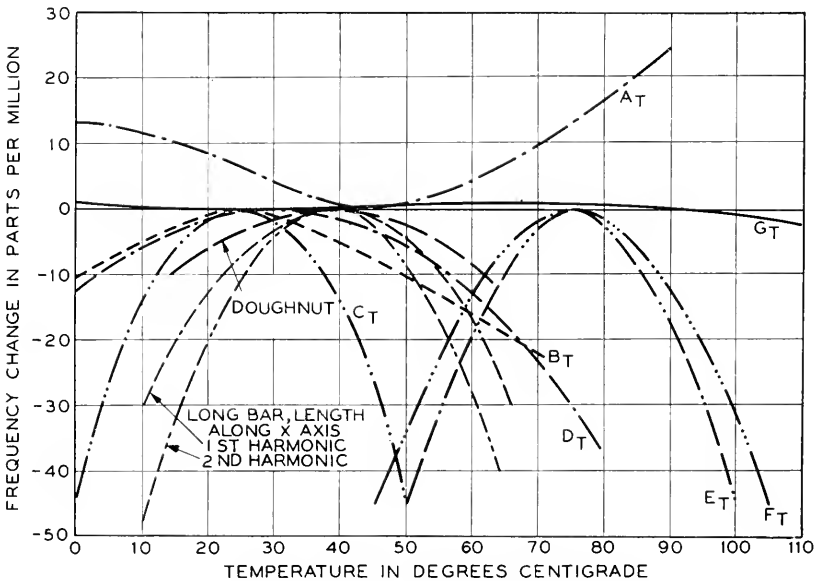


Fig. 14—Frequency temperature relations for zero temperature coefficient crystals

bolas. This is what would be expected for in general we can write the frequency as a function of temperature by the series

$$f = f_0[1 + a_1(T - T_0) + a_2(T - T_0)^2 + a_3(T - T_0)^3 + \cdots], \quad (22)$$

where T_0 is any arbitrary temperature. Differentiating f with respect to T we have

$$\frac{df}{dT} = f_0[a_1 + 2a_2(T - T_0) + 3a_3(T - T_0)^2 + \cdots]. \quad (23)$$

For a zero coefficient crystal the change in frequency will pass through zero at some temperature T_0. Hence $a_1 = 0$, and the frequency will then be

$$f = f_0[1 + a_2(T - T_0)^2 + a_3(T - T_0)^3 + \cdots]. \quad (24)$$

Since a_2 will ordinarily be much larger than succeeding terms, a parabolic curve will be obtained. If a_2 is positive the frequency will increase on either side of the zero coefficient temperature T_0 and if negative it will decrease.

Recently a new crystal cut, labeled the *GT*, has been found for which both a_1 and a_2 are zero. As a result the parabolic variation with temperature is eliminated and the frequency remains constant over a much wider range of temperature. The variation obtained is plotted on Fig. 14 by the curve labeled *GT*, and, as can be seen, the frequency does not vary over a part in a million over a 100° C. change in temperature.

This crystal, which will be described in a forthcoming paper, has found considerable use in frequency standards, in very precise oscillators, and in filters subject to large temperature variations. It has given a constancy of frequency considerably in excess of that obtained by any other crystal.

A New Standard Volume Indicator and Reference Level *

By H. A. CHINN,† D. K. GANNETT, and R. M. MORRIS ‡

In recent years it has become increasingly difficult to correlate readings of volume level made by various groups because of differences in the characteristics and calibrations of the volume indicators used. This paper describes a joint development by the Columbia Broadcasting System, National Broadcasting Company, and Bell Telephone Laboratories which resulted in agreement upon, and standardization in the respective broadcast and Bell System plants, of: a new copper-oxide rectifier type of volume indicator having prescribed dynamic and electrical characteristics; a new reference level based on the calibration of the new instrument with a single frequency power of one milliwatt; and a new terminology, the readings being described in "vu." It is hoped that other users of volume indicators will join in the adoption of these new standards.

The paper gives in considerable detail the technical data and considerations on which was based the choice of the characteristics of the new volume indicator and the other features of the new standards. Particular attention is paid to the technical data supporting the decision to make the new volume indicator approximately an r-m-s rather than a peak-reading type of instrument.

Introduction

THE student of electrical engineering, when introduced to alternating current theory, learns that there are three related values of a sine wave by which its magnitude may be expressed. These are the average value, the r-m-s (or effective) value, and the peak (or crest) value. Certain fundamental electrical measuring devices provide means for determining these values. As the student's experience broadens, he becomes familiar with complex, non-sinusoidal periodic waves and finds that these waves have the same three readily measured values. He learns how to determine from the problem under consideration whether the average, the r-m-s or the peak value of the wave is of primary importance.

* Presented at joint meeting of A. I. E. E. and I. R. E., San Francisco, California, June 1939, and at Fourteenth Annual Convention of I. R. E., New York, September 1939.

† Mr. Howard A. Chinn is Engineer-in-Charge, Audio Engineering, Columbia Broadcasting System, Inc.

‡ Mr. Robert M. Morris is Development Engineer, National Broadcasting Company, Inc.

If the student later enters the field of communication engineering, he immediately encounters waves which are both very complex and non-periodic. Examples of typical speech and music waves are shown in the oscillograms of Fig. 1. When an attempt is made to measure such waves in terms of average, r-m-s or peak values, it is found that the results can no longer be expressed in simple numerical terms, as these quantities are not constant but variable with time and, moreover, are apparently affected by the characteristics of the measuring instrument and the technique of measurement. However, the communications engineer is vitally concerned with the magnitude of waves of the sort illustrated, as he must design and operate systems in which they are amplified by vacuum tubes, transmitted over wire circuits, modulated on carriers, and otherwise handled as required by the various communication services. He needs a practical method of measuring and expressing these magnitudes in simple numerical fashion.

This need may be better appreciated by considering the communication systems employed for broadcasting. These are very complicated networks spread over large geographical areas. A typical network may include 15,000 miles of wire line and hundreds of amplifiers situated along the line and in the 50 to 100 connected broadcasting stations. Every 15 minutes during the day the component parts of such a system may be shifted and connected in different combinations in order to provide for new points of origin of the programs, and for the addition of new broadcasting stations and the removal of others from the network. In whatever combination the parts of the system are put together, it is necessary that the magnitude of the transmitted program waves, at all times and at all parts of the system, remain within the limits which the system can handle without impairment from overloading or noise. To accomplish this, some convenient method of measuring the amplitude of program waves is needed.

These considerations led to the conception of a fourth value, known as "volume," whereby the magnitude of waves encountered in electrical communications, such as telephone speech or program waves, may be readily expressed. This value is a purely empirical thing, evolved to meet a practical need. It is not definable by means of a precise mathematical formula in terms of any of the familiar electrical units of power, voltage or current. Volume is simply the reading of an instrument known as a volume indicator, which has specified dynamic and other characteristics and which is calibrated and read in a prescribed manner. Because of the rapidly changing character of the program wave, the *dynamic characteristics* of the instrument are fully as important as the value of sine wave power used for calibration. The

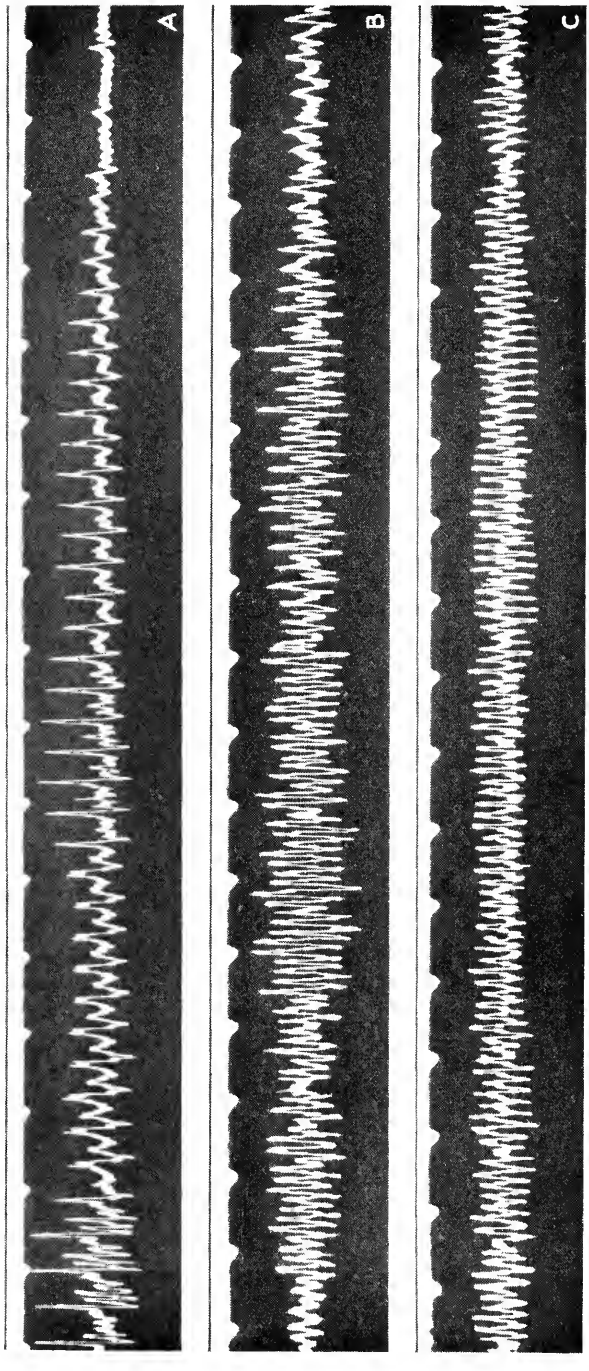


Fig. 1—Examples of program wave forms.

A—Male speech ("how many").
B—Male solo with orchestra.
C—Dance orchestra.
(The frequency of the timing impulses is 60 per second.)

readings of volume have been customarily expressed in terms of decibels with respect to some volume level chosen as the "reference" level.

In the past, because of a lack of complete understanding of the matter, there has been little uniformity in the design and use of volume indicators, although attempts have been made by some organizations toward standardization. The devices used were of the r-m-s and peak-reading types having slow, medium or high pointer speeds; half- or full-wave rectifiers; critically to lightly damped movements and reference levels based on calibrations with 10^{-9}, 1, 6, 10, $12\frac{1}{2}$ or 50 milliwatts in 500 or 600 ohms. This great array of variables led to considerable confusion and lack of understanding, especially when an attempt was made to correlate the measurements and results of one group with those of another.

To remedy this situation, the Bell Telephone Laboratories, the Columbia Broadcasting System and the National Broadcasting Company entered upon a joint development effort during January 1938, with the object of pooling their knowledge and problems, of pursuing a coordinated development program, and of arriving at a uniform practice of measuring volume levels. The outcome of this work is a new volume indicator, a new reference volume level, and new terminology for expressing measurements of volume level. The results of this development work have been discussed with, and approved by, more than twenty-four other organizations, and were presented at an open round table conference at the Annual Convention of the Institute of Radio Engineers on June 17, 1938. During May 1939, it was adopted as standard practice by the above two broadcasting companies and the Bell System, and it is hoped that they will be joined by others. It is the purpose of this paper to describe the new standards and the considerations which led to their adoption.

Early History of Volume Indicators

As a background for understanding the present development, it will be helpful to review briefly the early history of volume indicators. The particular occasion for the development of the first volume indicator was the setting up of the public address system which enabled the ceremonies attendant upon the burial of the Unknown Soldier on Armistice Day 1921, to be heard by large audiences at Arlington, New York and San Francisco.[1] It was noted in some of the preliminary tests that distortion due to overloading of an amplifier was more objectionable when heard in a loud speaker than when heard in an ordi-

[1] "Use of Public Address System with Telephone Lines," W. H. Martin and A. B. Clark, *Transactions A. I. E. E.*, February 1923.

nary telephone receiver. Consequently, to avoid overloading the telephone repeaters when they were used on the public address circuits, a device was proposed which would give visual indication on an instrument when the speech level was such as to cause the telephone repeaters to overload.

Further development of this idea led to the experimental device which was used in the Armistice Day ceremonies and which later, with no fundamental change, became the well-known 518 and 203 types of volume indicators. This device consisted of a triode vacuum tube functioning as a detector, to the output of which was connected a d.-c. milliammeter. Associated with the input was a potentiometer for adjusting the sensitivity in 2 db steps. The method of using the device was, to adjust the potentiometer so that the maximum movement of the milliammeter needle reached the mid-scale point on an average of about once every ten seconds, occasional greater deflections being disregarded. The volume level was then read from the setting of the potentiometer which was marked in decibels with respect to a reference volume level.

The reference level was chosen as that level of speech which, when transmitted into the long telephone circuits, would cause the telephone repeaters with which they were equipped to be just on the verge of overloading as evidenced by audible distortion. The gains of the telephone repeaters were normally adjusted so that the level at their outputs was 10 db higher than at the sending end of the circuit. Reference volume was therefore specifically defined as 10 db below the maximum speech level which could be satisfactorily transmitted through the particular amplifier and vacuum tube used in the telephone repeaters. This level was determined experimentally and the potentiometer steps of the volume indicator were marked accordingly. The reference volume was also approximately the volume delivered over a short loop by the then standard subset when spoken into with a fairly loud voice.

It is apparent that the volume indicator was born in response to a definite need, and it has filled an important niche in the rapidly growing radio broadcasting industry and in other communication fields. Large numbers of volume indicators similar to this early type have continued in service to the present time.

It is a frequent characteristic of a rapidly expanding art that at first standards multiply, and finally a point is reached where simplification and agreement upon a single standard becomes imperative. This has occurred in connection with volume indicators and since the development of the first one, a variety of instruments have been produced

by the various manufacturers and have come into service in the plants of the different companies. These instruments had different calibrations and characteristics with little correlation between their readings.

A further divergence occurred, regarding the philosophy of the calibration of the original type of volume indicator. One view recognized no correlation between the point at which the galvanometer was normally read on peaks (the 30 division point on the scale, Fig. 12) and the power of six milliwatts used for calibration. When calibrating the instrument on six milliwatts of sine wave energy in 500 ohms, the galvanometer would read 22 divisions with the associated sensitivity switch on step zero. There was not intended to be any correlation between this calibrating power and reference volume. Nevertheless, many people were led by this technique of calibration to refer to the volume indicator as a 6-milliwatt instrument. This idea was furthered by the fact that the vacuum tube to whose speech-carrying capacity the reference volume was originally referred, has a nominal full load capacity on sine waves of 60 milliwatts. The reference volume being defined as 10 db below the maximum output of this tube, it was natural to try to relate this reference volume to the corresponding figure of 6 milliwatts for sine waves.

The second view was based on the experimental fact that when the potentiometer controlling the sensitivity was set at "0 db," a sine wave potential of 2.5 volts (r-m-s) applied to the volume indicator caused a deflection to mid-scale (scale reading of 30 divisions). This was equivalent to 12.5 milliwatts in a 500-ohm circuit, and the supporters of this view therefore referred to the volume indicator as a 12.5-milliwatt instrument.

Thus the same volume indicator, having the same sensitivity and giving the same readings of volume level, was variously referred to as a 6-milliwatt and a 12.5-milliwatt device. This increased the difficulty of coordination between the plants of the different companies which are interconnected in rendering broadcast service.

Some degree of standardization of the technique of reading volume levels had already been made within different organizations both here and abroad. The importance of the present development lies not only in the particular merits of the proposed standards, but also in the fact that they have been jointly developed and adopted by three of the larger users of volume indicators, and have been approved by many others. Thus there is good prospect that the needed standardization is about to be realized, and that all will shortly use the same instruments, the same reference levels, the same terminology, and the same nominal value of circuit impedance.

Choice of Peak vs. R-M-S Types

General

The first important decision to be made and one which would affect the entire character of the development was whether the new volume indicator should be of the r-m-s or of the peak-reading type. These two types of instrument represent two schools of thought. The peak-reading instrument is favored for general use by many European engineers and is specified by the Federal Communications Commission for use as modulation monitors in this country. The r-m-s type has, however, been commonly employed in this country on broadcast program networks and for general telephone use. In view of the importance of the decision and the difference of opinion that has existed, the data on which the choice was made are given below in considerable detail.

In accord with common practice, the terms "r-m-s" and "peak-reading" are used rather loosely throughout this paper. The essential features of an r-m-s instrument are some kind of rectifier or detector and a d.-c. milliammeter. The latter is not especially fast, generally requiring tenths of a second to reach substantially full deflection. Obviously, if a sufficiently slow wave is applied, say one whose frequency is one or two cycles per second, the instrument can follow it and the true peaks of the wave will be indicated, but when much higher frequency waves are applied, such as the complex speech or program waves, the instrument is too slow to indicate the instantaneous peaks but averages or integrates whole syllables or words. As shown by tests and practical experience, it is of secondary importance whether the detector actually has an r-m-s (or square law) characteristic, or has a linear or some intermediate characteristic.

A peak-reading instrument capable of truly indicating the sharpest peak which might occur in a high quality program wave would have to respond to impulses lasting only a very small fraction of a millisecond. Cathode-ray oscilloscopes or gas tube trigger circuits are capable of doing this, and therefore might be used as peak-reading volume indicators. However, the so-called peak-reading volume indicators used in practice, designed to give a visual indication on an instrument, are far from having the above speed although they are much faster than the r-m-s instruments. They generally respond to impulses whose duration is measurable in hundredths or thousandths of a second. They therefore truly indicate the peaks of sine-wave voltage whose frequency does not exceed, say, 50 to 100 cycles per second. They are similar to the r-m-s instruments in that they are

not fast enough to indicate the instantaneous peaks of speech or program waves but tend to average or integrate a number of peaks of the wave.

A feature of the usual peak-reading instrument which from the analytical standpoint is of secondary importance, is that it is usually given a characteristic of very slow decay as well as rapid response. This is usually accomplished by a circuit such as illustrated in Fig. 2, which shows the principle of the experimental instrument used in the tests described later. The 0.01-mf. condenser is charged through a full wave vacuum tube rectifier, the rates of charge and discharge being determined by the resistances. The d.-c. amplifier and d.-c. milliammeter indicate the charge on the condenser. The advantage of making the discharge rate of the condenser very slow is that the d.-c.

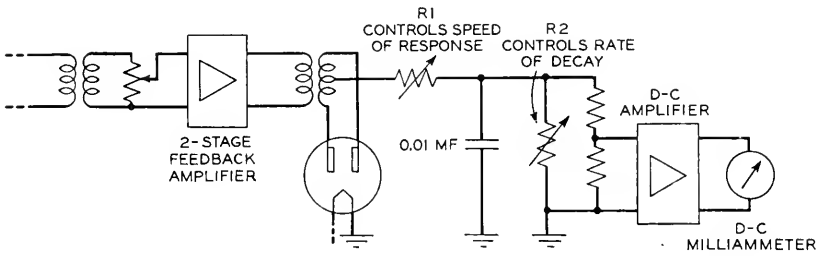


Fig. 2—Schematic diagram of experimental peak reading volume indicator.

milliammeter need not then be particularly fast and, moreover, the ease of reading the instrument is greatly increased.

From the above analysis it is seen that the r-m-s and the peak-reading instruments are essentially similar and differ principally in degree. Both indicate peaks whose durations exceed some value critical to the instrument and both average or integrate over a number of peaks the shorter, more rapid peaks encountered in speech or program waves. Either may have a linear or a square law detector, or one of some intermediate characteristic. The important difference between the two types lies in the speed of response as measured by the length of impulses to which they will fully respond, that is, in the time over which the complex wave is integrated.

A general purpose volume indicator may be called upon to serve a number of uses, such as:

(*a*) Indication of a suitable level for a speech or program wave to avoid audible distortion when transmitted through an amplifier, program circuit, radio transmitter or the like.

(*b*) Checking the transmission losses or gains in an extended program network by simultaneous measurements at a number of points on particular peaks or impulses of the program wave which is being transmitted.

(*c*) The indication of the comparative loudness with which programs will be heard when finally converted to sound.

(*d*) The indication of a satisfactory level to avoid interruption of service due to instantaneous overloads tripping protective devices in a radio transmitter, damage to sound recording systems, etc.

(*e*) Sine-wave transmission measurements.

These services are different in nature and the ideal requirements for an instrument for each may not necessarily be the same. One instrument to serve them all must, therefore, be a compromise. From the standpoint of the companies engaged in this development, items (*a*), (*b*) and (*c*) in the above list were considered to be the most important and therefore attention was first directed to the relative merits of the two types of volume indicators with respect to them.

Aural Distortion Due to Overload

Tests of volume indicators as overload indicators with aural distortion as the criterion [item (*a*)] had previously been made on a number of occasions and more tests were undertaken during the present development. The general procedure in such tests is to determine for some particular amplifier the volume level at its output at which distortion due to overloading can just be heard by a number of observers on each of a variety of programs. The volume levels thus determined are read on the various volume indicators which are being compared. The best instrument is considered to be the one whose readings are most nearly alike for all the programs when overloading can just be detected.

The sole criterion of distortion due to overloading is the judgment of observers, since it is the final reaction on listeners which is of importance. This judgment is not subject to exactness of measurement, but is in fact somewhat of a variable, even with conditions unchanged and with the most experienced observers. For significant results to be obtained, therefore, a careful technique of conducting the tests is required, many observations must be made, and statistical methods of analyzing the resultant data must be employed.

The arrangement of equipment and circuits used in these tests is shown in simplified form in Fig. 3. A source of program, which may

be a phonograph pickup, a direct microphone pickup, or a program circuit, is connected through control circuits to the amplifier which is to be overloaded, and thence through additional circuits to a loud speaker. The loud speaker employed in the tests reported here was a special high quality two unit loud speaker having a response which is substantially flat from 40 to 15,000 cycles per second.[2] Including the power amplifier used with it, the overall response of the system was substantially uniform from 40 to 11,000 cycles.

The arrangement of the circuit is such that the volume level at the output of the test amplifier may be raised or lowered while keeping the overall gain of the system constant. Two controls are provided for this purpose. One, operated by a key, transfers a 15 db loss from ahead to behind the test amplifier. This permits comparing a test

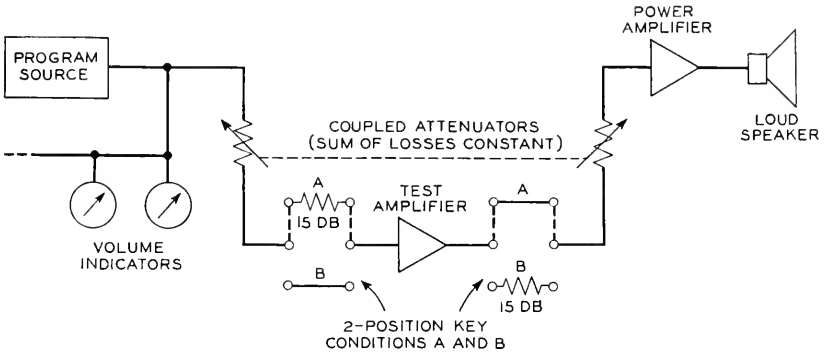


Fig. 3—Arrangements for determining volume level at which overload of amplifiers is audible.

condition with a reference condition in which the load on the amplifier is 15 db lower, while the loudness with which the program is heard remains the same for either condition. The other control, represented in Fig. 3 by the coupled attenuators, permits the load on the amplifier for the test condition to be varied, also without changing the loudness. The volume indicators to be compared are connected for convenience, to a point where the volume level is unaffected by the controls. Their readings are corrected for each test by the measured loss or gain between the point where they are situated and the output of the test amplifier, so as to express the levels which would be read at the amplifier output.

Two techniques were employed for conducting tests with this equipment. In one, the individual method, a single observer at a time

[2] "Auditory Perspective—Loud Speakers and Microphones," E. C. Wente and A. L. Thuras, *Electrical Engineering*, January 1934.

listens to the program and adjusts the volume level at the output of the amplifier by means of the coupled attenuators, until he determines the point at which distortion due to overloading is just audible, when the key is operated from the reference to the test condition. This is repeated for a number of different programs and observers until a large number of observations have been obtained. The volume levels indicated by the different volume indicators at the amplifier output are determined for each observation. These are found to have a considerable spread, due not only to the differences in the nature of the programs but also to differences in the acuity of perception of the distortion by the various observers. The method of analyzing the data is described later.

In the second technique, the group method, a group of observers simultaneously listens to a program which is repeated with the key operated alternately to the test and reference positions. The two conditions are distinguished to the observers (but not identified as to which is which) by a letter associated with each condition in an illuminated sign. The letters A, B and C are used, two being chosen at random for each test. A vote is taken as to which condition, designated by one of the two letters employed in the particular test, is preferred with respect to freedom from distortion. A number of such tests, covering the range from a level below the point where distortion can be detected by anyone to a level high enough for all to observe distortion, establishes a curve between the per cent of observers correctly choosing the reference condition as having the least distortion, and the amplifier output level as read by each volume indicator used in the tests. Similar curves are determined for a number of kinds of program material, and for purposes of comparison the overload point for each program is taken from the point on the curve for each volume indicator, where 80 per cent [3] of the observers voted correctly.

As noted, judgment tests of this sort require many observations and checks to obtain reliable results. A larger volume of data is available for the individual method, so the results from tests made by that method have been chosen to be reported here. Some tests have also been made with the group method and, while the results are less conclusive, they substantiate those recorded below.

Tests by the individual method to compare peak-reading and r-m-s volume indicators have been carried out a number of times during the past two years. In each of these tests a number of observers have taken part and a number of samples of program material of a variety of

[3] "Audible Frequency Ranges of Music, Speech and Noise," W. B. Snow, *Journal of the Acoustical Society of America*, July 1931.

types have been employed. For the majority of the tests, the sources of program were high quality recordings, convenient because of the ease and exactness with which the programs could be repeated. For some of the tests, however, actual speakers and musical instruments were employed with direct microphone pickup.

A number of the types of volume indicators in common use were represented in these tests. Since the 700A Volume Indicator was common to all of the tests, it has been chosen to represent the r-m-s type of volume indicator in the data presented below. The peak-reading type was represented by the especially constructed experimental instrument, whose fundamental circuit is shown in Fig. 2. The resistances controlling the rates of charge and discharge of the condenser were adjustable, permitting a range of characteristics to be obtained. The adjustments for which the data referred to below were obtained, correspond to a rate of charge of the condenser such that impulses of single frequency applied to the input for 0.025 second would give a reading within 2 db of the reading obtained with a sustained wave of the same amplitude. The rate of discharge of the condenser was about 19 db per second. These rates are generally similar in magnitude to those specified by the International Consultative Committee on Telephone Transmission (the C. C. I. F.) for broadcast service, and by the Federal Communications Commission for modulation monitors.

The d.-c. amplifier and d.-c. milliammeter which indicates the charge on the condenser included features, not shown in the simplified sketch, which made the response logarithmic. The instrument had a substantially uniform decibel scale covering a range of 50 db.

The data from four different series of tests, made at different times, were collected in one body, and distribution curves were plotted showing the relative frequency of occurrence among the data of the different levels at which incipient overload was detected. Curves for tests on a Western Electric 94B Amplifier, which is an amplifier designed with negative feedback and therefore having a relatively sharp cutoff, similar to a radio transmitter, are illustrated in Fig. 4. It will be noted that the curve obtained with the r-m-s volume indicator has a slightly greater spread than that for the peak-reading volume indicator. Twelve different observers took part in these tests, and 13 samples of program were employed, including male and female speech, dance music, piano, violin and brass band selections.

The data may more readily be interpreted when plotted in the form of cumulative distribution curves, obtained by integrating the above distribution curves. Cumulative curves for the data just referred to

are shown in Fig. 5. For convenience and ease of interpretation, these curves have been plotted on "probability" rather than rectangular coordinates, as probability coordinates have the property of making data whose distribution follows a normal law [4] form a straight line. It will be noted that the experimentally determined points actually fall so nearly on straight lines, that it is reasonable to assume straight lines to represent them. It is likely that with a greater volume of data, still greater conformity to the straight lines drawn, would be obtained.

In order to superpose the curves for the two volume indicators, the levels are plotted in decibels with respect to the average overload level

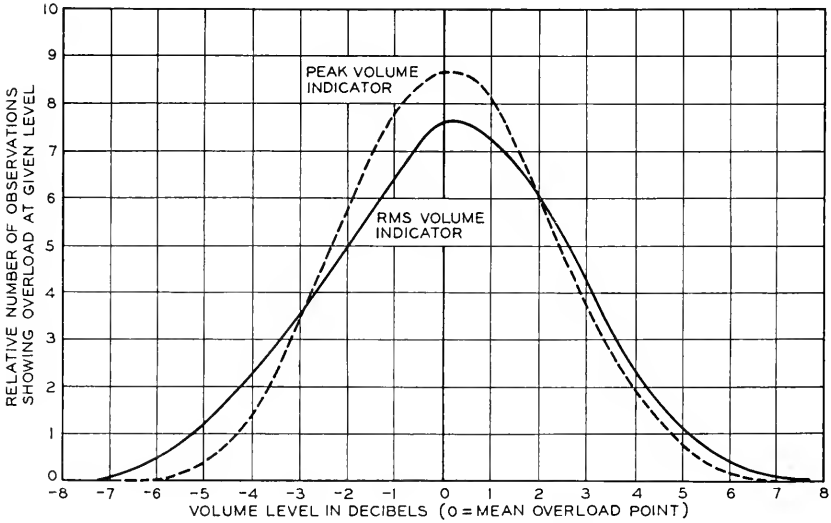


Fig. 4—Distribution of overload points.

determined from the tests. When calibrated to read alike on the same sine-wave power, the experimental peak-reading instrument (with the adjustments described above) reads on the average 7.4 decibels higher on actual programs than the r-m-s instrument used in the tests.

Now let it be imagined that the test amplifier is the one critical link in a broadcast network and that an operator is given the duty of satisfactorily adjusting the volume levels through the amplifier using either of the two volume indicators tested. If he lets the louder portions of the programs just reach the volume level marked "0 db" on the curves, it will make no difference which volume indicator he

[4] The "normal" law has the form $y = Ae^{-ax^2}$.

uses. In either case, on the average, half of the listeners will hear distortion when the program is loudest. However, this result would probably be considered too poor, so suppose the maximum level is lowered 3.5 decibels. Referring to the curves, it is seen that if the

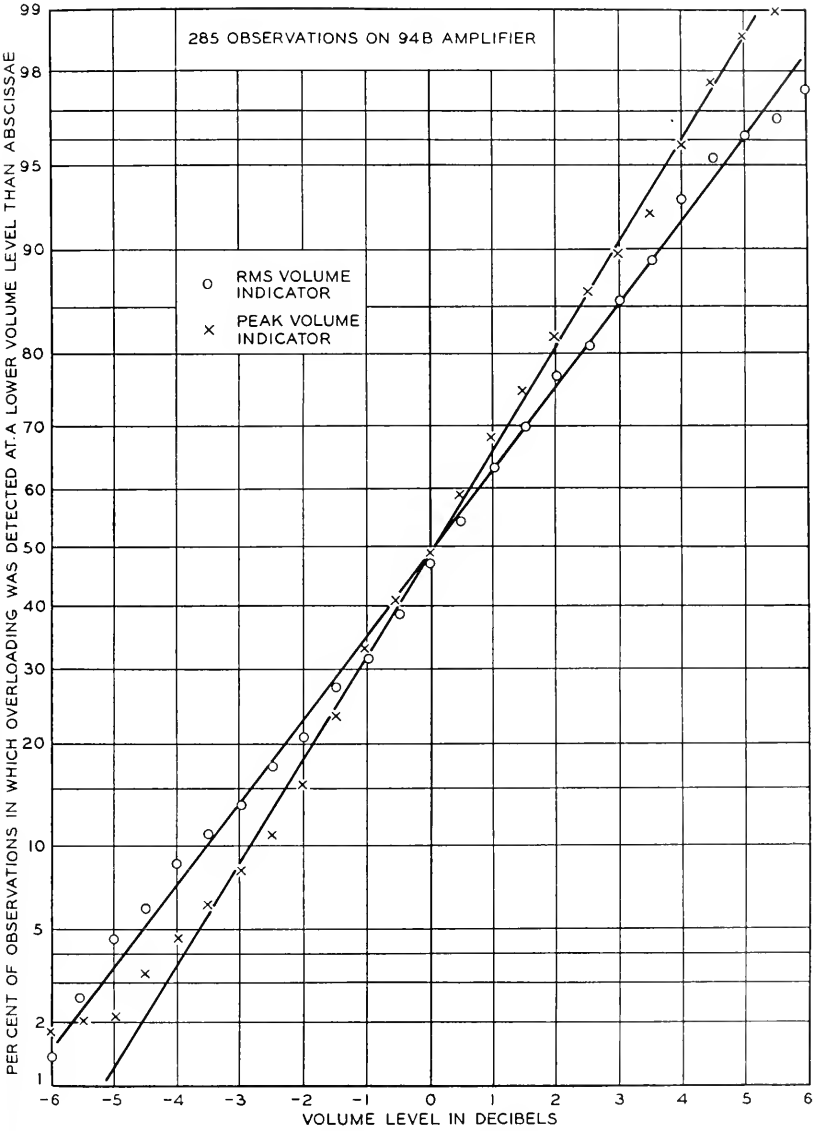


Fig. 5—Comparison of peak vs. r-m-s volume indicators as overload indicators (using W.E. 94B amplifier).

peak-reading volume indicator is used, only about 5 per cent of the listeners will now on the average hear distortion on the loudest program passages, while if the r-m-s instrument is used, about 10 per cent will hear distortion. To reduce the latter figure to 5 per cent would require lowering the maximum volume level another decibel. Thus with this criterion, the peak instrument has a slight advantage, as it would permit the transmission of a 1 decibel higher average volume level for the same likelihood of distortion being heard.

The above statements assume that the observers and programs used in the tests just described were representative of the listening public and the programs they hear. Actually, the observers were trained by experience in making many tests and were no doubt much more critical than the average listener. Moreover, the conditions under which the tests were performed, with the availability of frequent comparison with the undistorted reference condition, were more conducive to critical detection of overload than are average listening conditions. These facts, together with the inevitable inability of the control operator in practice to make his adjustments perfectly in anticipation of the coming changes in the programs, tend to make the real practical advantage of one instrument over the other considerably less than shown by the tests. A further factor reducing the importance of the small differences shown by the tests is the growing use of volume limiting amplifiers at critical points in a broadcast system, such as at the radio broadcast stations, which automatically prevent the transmission of excessive levels.

Another cumulative distribution curve is shown in Fig. 6, representing similar tests on a Western Electric 14B Program Amplifier. This is a simple push-pull triode amplifier without negative feedback and therefore having a more gradual cutoff than the 94B. (The gain versus output power level curves at 1000 cycles per second are shown in Fig. 7 for the two amplifiers.) It will be seen from Fig. 6 that the data for the two volume indicators show no significant difference and that the single curve equally well represents either set of data in the region of interest. Somewhat fewer data are represented by this curve and the agreement with the normal law is not quite so close as in the previous case.

The peak-reading instrument with the adjustment used in these tests, although having characteristics similar to those usually proposed for this type of device, is still far too slow in response to indicate the true instantaneous peaks of the program wave. The question naturally arises, therefore, whether any greater difference would be indicated if the peak-reading instrument were made sufficiently fast in response

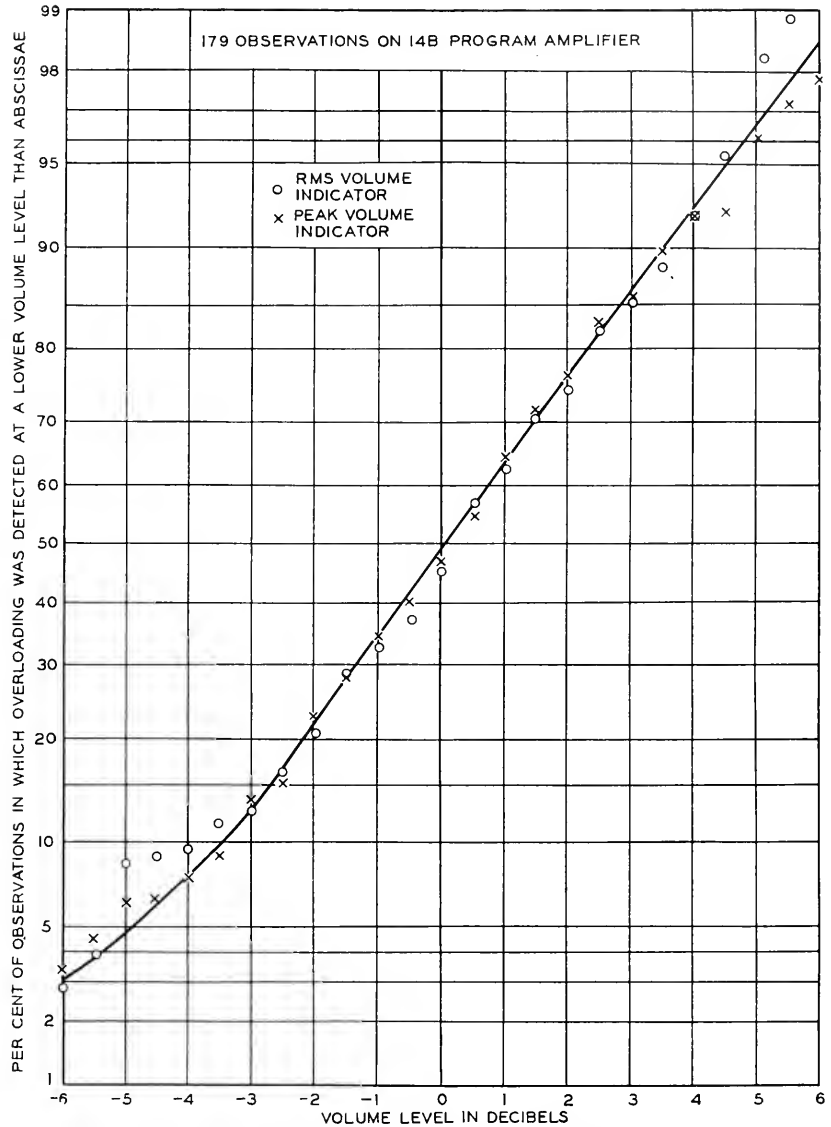


Fig. 6—Comparison of peak vs. r-m-s volume indicators as overload indicators (using W.E. 14B program amplifier).

to indicate the actual instantaneous peaks. To check this point, some tests similar to those described above were made, using a gas tube trigger circuit capable of measuring the true instantaneous peaks. The results of these tests, using the 94-B amplifier, are shown in Fig. 8.

Although a smaller number of observations are included in these data, the results show conclusively that there is no substantial difference between the experimental peak-reading volume indicator and the faster trigger tube arrangement, in their performance on actual program waves.

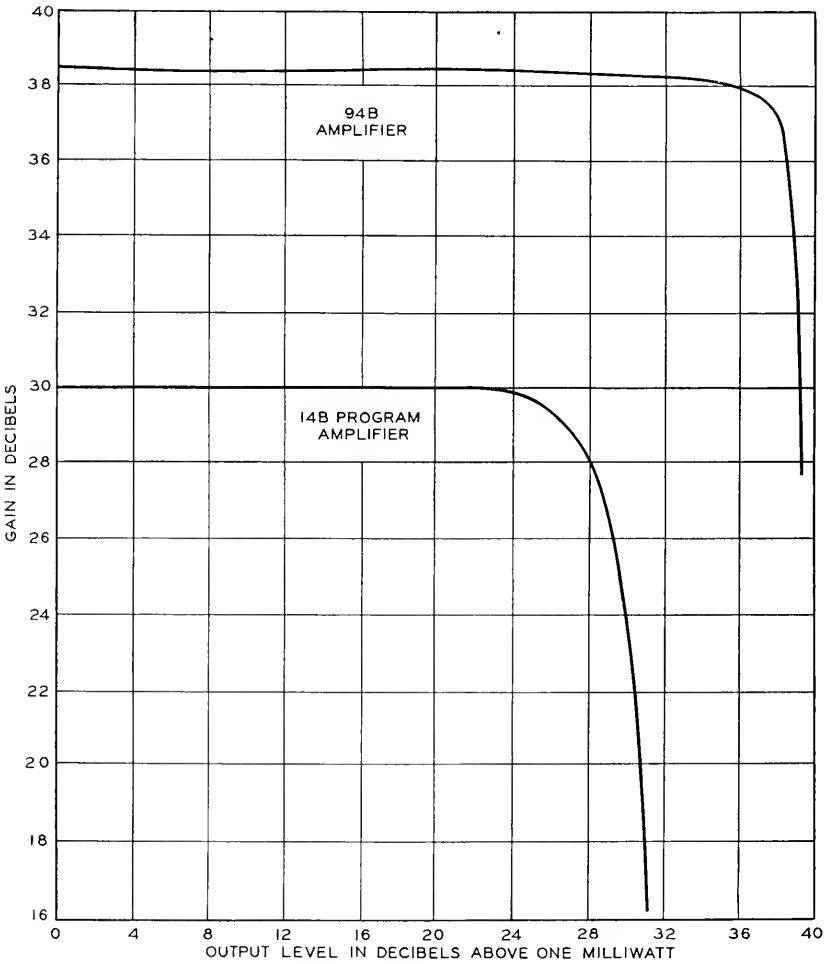


Fig. 7—Gain vs. load characteristics of amplifiers.

The data from the tests have been presented above in the form which most directly indicates the comparative performance of the two types of volume indicators. However, a breakdown of the data with respect to the types of program may be of interest and is shown in Tables I

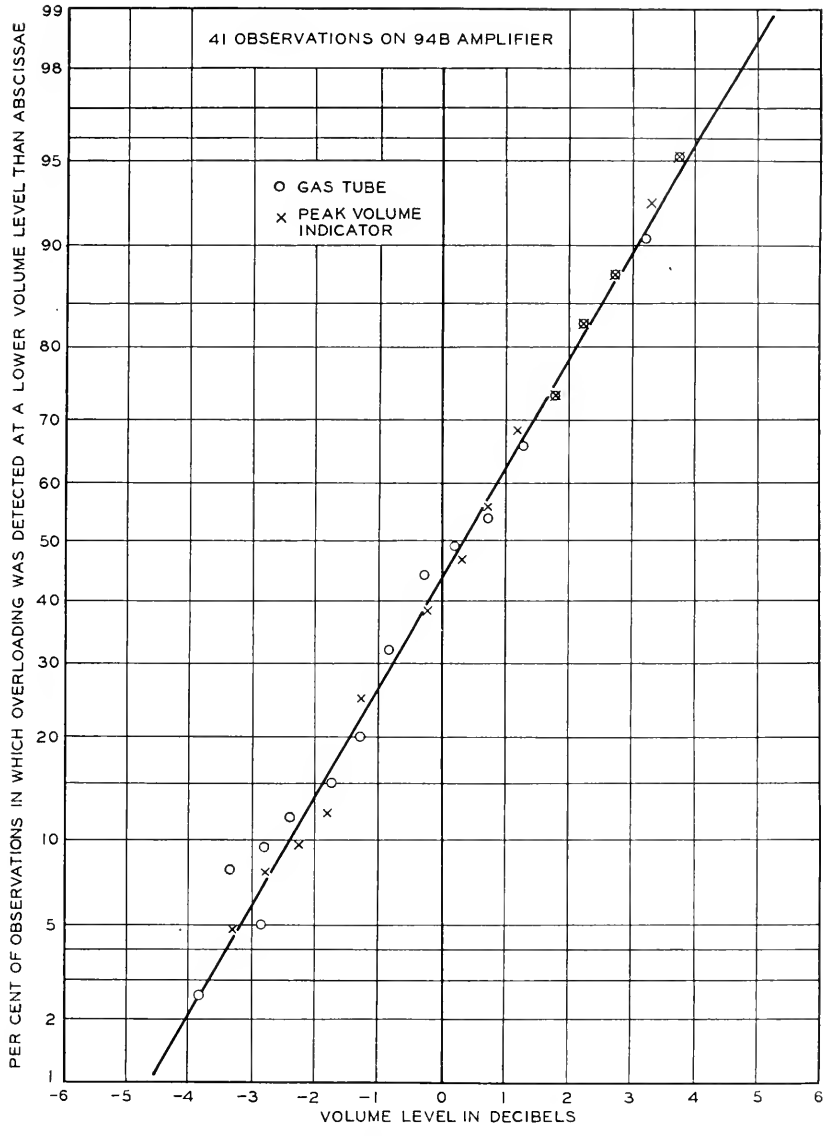


Fig. 8—Comparison of experimental peak volume indicator with gas tube trigger device as overload indicators.

and II for the data on the 94B amplifier shown previously in Figs. 4 and 5.

It will be observed in Table I that the average overload points for the different types of programs fall within a range of about 2 db for

TABLE I

AVERAGE OVERLOAD POINTS OF DIFFERENT KINDS OF PROGRAM MEASURED AT THE OUTPUT OF THE 94B AMPLIFIER

Character of Program	No. of Tests	Total No. of Observations	Average Overload Point*	
			R-M-S V.I.	Peak V.I.
Male Speech	8	81	22.1 db	31.9 db
Female Speech	8	82	22.8	30.1
Piano	5	40	24.1	30.9
Brass Band	4	25	24.1	31.0
Dance Orchestra	5	42	24.7	29.4
Violin	1	15	25.8	31.1
Average Speech	16	163	22.4	31.0
Average Music	15	122	24.5	30.5
Grand Average	31	285	23.3	30.7

* These tests antedated the new standards, and the values given are in db with respect to a reference point based on a single frequency calibration of .006 watt in 600 ohms.

TABLE II

SPREAD OF OVERLOAD POINTS WHOSE AVERAGES ARE GIVEN IN TABLE I

Character of Program	R-M-S V.I.	Peak V.I.
Male Speech	6.1 db	3.7 db
Female Speech	4.6	2.5
Piano	3.6	4.9
Brass Band	4.0	3.9
Dance Orchestra	3.7	2.4
All Types	7.3	5.9

either volume indicator. However, it will be noted that with the r-m-s instrument the average overload point for speech is about 2 db lower than for music, while there is no significant difference with the peak instrument. This undoubtedly is because speech waves have a higher "peak-factor" (ratio of peak to r-m-s values) than music.

Table II shows the spread of the overload points (difference between highest and lowest values) for the various tests on each type of program whose average is given in Table I. Most of the types of program show a significantly narrower spread for the peak than the r-m-s instrument. For comparison with values taken from Figs. 5 and 6, discussed above, these spreads should be divided by 2 to show the difference between the lowest and the mean values.

It is concluded from the tests just described that the disadvantage in using r-m-s instead of peak-reading volume indicators for controlling volumes to avoid aural distortion due to overloading, is substantially none when the overloading device does not have too sharp an overloading characteristic, and only slight when it does overload sharply.

The explanation probably lies in the physiological and psychological factors involved in the ear's appreciation of overload distortion, which permit to pass unnoticed considerable amounts of distortion on rarely occurring instantaneous peaks of very short duration.

Peak Checking

A very important use of volume indicators is that of checking the transmission losses or gains along a program network by measurements made on the transmitted program material [item (*b*) in the list given earlier]. The program circuits which make up the large program networks are in continuous use for many hours each day, and during that period are switched together in many combinations as called for by the operating schedules. It is not convenient to interrupt service for sine-wave transmission measurements; hence to check the transmission conditions during service hours, it is the custom to take simultaneous readings at two or more points in the program networks on particular impulses of whatever program wave is being transmitted, coordinating these readings by the use of an order wire. On such readings, the r-m-s type of instrument is far superior to the peak-reading type, because of phase distortion and slight non-linearity in the program circuits. These effects are undetectable to the ear, but change the wave shape of the program peaks sufficiently to cause serious errors in the readings of the peak-type instrument. On the other hand they have no noticeable effect on the r-m-s instruments.

Tests were made on this effect by taking readings on several kinds of program at the beginning and end of a program circuit extending from New York to Chicago and return (about 1900 miles). The circuit was lined up so that either volume indicator read the same at both ends of the circuit on a 1000-cycle sine wave. In all the tests, the readings obtained on program material with the r-m-s instrument at the two ends of the circuit agreed within a very few tenths of a decibel. The readings of the peak instrument, however, disagreed by the values shown in Table III, when the program material was applied to the circuit at the normal maximum operating level.

It is of interest that the errors shown by the table are affected by the frequency range of the program material transmitted, being greater for the broader band. The frequency range was controlled by the use of low-pass filters inserted between the source of program and the line before the point at which the sending end levels were read. Tests were also made of the effect of a 180-degree phase reversal at the center of the loop. This was found to increase the errors in some cases and to decrease them in others.

TABLE III

ERRORS RESULTING FROM USE OF PEAK-TYPE VOLUME INDICATOR ON A LONG PROGRAM CIRCUIT

	Upper Frequency Limit of Program	
	5000 Cycles	8000 Cycles
Male Speech	−3.5 db	−4.5 db
Female Speech	−1.5	−3.0
Dance Orchestra	−2.0	−1.5
Brass Band	−3.0	−2.0
Piano	−0.5	−1.5

The large errors indicated in the table are, of course, intolerable. The effect of the line on the reading of the peak instrument is partly due to the cumulative effects of the slight non-linearity in the many vacuum tube amplifiers and loading coils in the circuit, and partly to phase changes which alter the wave front and amplitude of the peaks. It might be thought that phase changes which destroy some peaks would tend to create others. However, a Fourier analysis of a sharp peak will show that an exact phase relationship must exist between all of the frequency components. The probability that phase shift in a line will chance to cause all of the many frequency components of a complex wave to align themselves in the relationship necessary to create a peak where none existed before, is very slight,—indeed infinitesimal compared to the probability of the occurrence of a peak in the original wave.

Loudness

Another important consideration is the correlation between volume levels and the comparative loudness of different types of programs [item (*c*) in the list given earlier]. This was tested by a method similar to the "group method," described above in connection with the tests on aural overload distortion. A group of observers was permitted to listen to alternate repetitions of a test program and a reference program, and was asked to vote upon which appeared the louder. A particular selection of male speech was used as the reference program for all of the tests and its level was kept constant. The test programs included several different types and several samples of each type of program. The samples of program were about 30 seconds in length. Each test program was presented at a number of levels covering a range from a low level where all the observers judged the reference program to be the louder to a higher value where all of them judged the test program to be the louder.

Thus, a curve was established for each type of program between the per cent of observers judging the test program to be the louder, and the level of the test program. A sample of such a curve is shown in

Fig. 9. The 50 per cent point on the curve is interpreted as indicating the level of the test program at which it appears to the average observer to have the same loudness as the reference program. The test program is then set at this "equal loudness" volume level and the levels of both test and reference programs are read with each of the types of

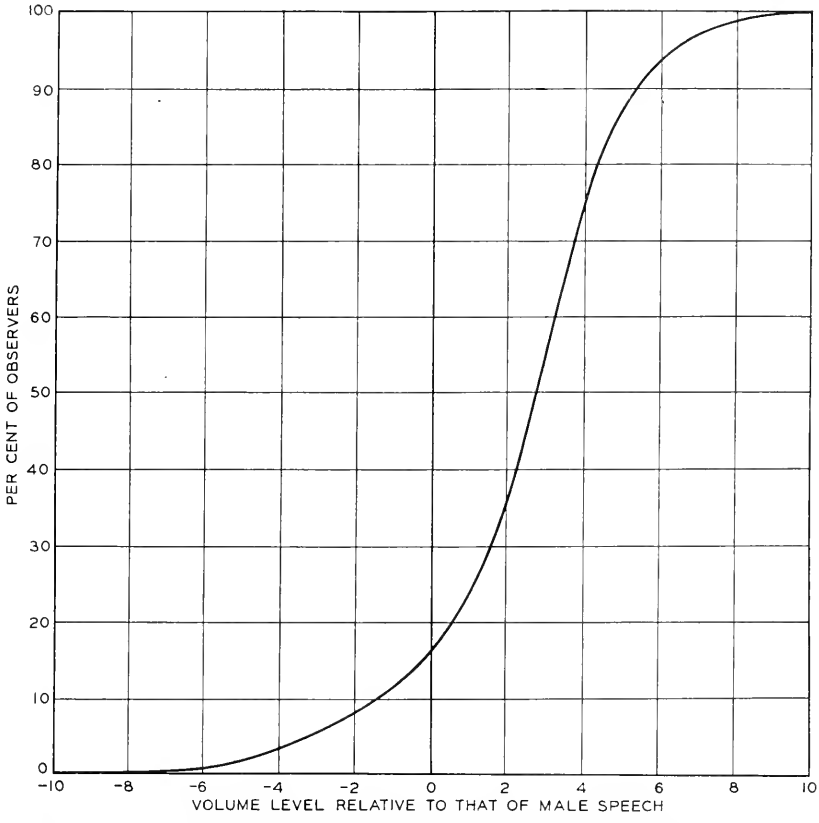


Fig. 9—Per cent of observers choosing symphony music at indicated volume levels to be louder than the male speech reference.

volume indicators of interest. In this way, the figures given in Table IV were determined.

It is evident from the figures in the table that there is no significant advantage for either type of volume indicator where loudness is the criterion.

Table IV shows that when the new volume indicator is used the musical programs must be 2 to 3 db higher than speech to sound equally loud. It is of interest to note that according to Table I this same

TABLE IV

Type of Program	Volume Indicator Readings for Same Loudness as Male Speech	
	New Volume Indicator	Peak Volume Indicator
Male Speech	0	0
Female Speech	−0.1	−2.2
Dance Orchestra	+2.8	−2.2
Symphony Orchestra	+2.7	−2.3
Male Singing	+2.0	−2.5

difference was shown to exist between the average overload point of the 94B amplifier on speech and music, when measured with the r-m-s volume indicator. This would seem to indicate that if allowance is made for this difference between speech and music in controlling the volume levels to avoid overloading, they will also then sound equally loud to the listeners.

Choice of Type

The tests of aural distortion due to overload showed so slight a disadvantage for the r-m-s instrument and the experiments on peak checking showed such a marked advantage for this type as compared with the peak instrument, that it was decided to develop the r-m-s type of instrument. Another consideration was that, with the advances in copper-oxide types of instruments, it has become possible to make r-m-s instruments of sufficient sensitivity for most purposes without the use of vacuum tubes and their attendant need of power supply, an advantage not shared by peak-reading instruments, at least at present. Thus, the r-m-s instrument has advantages of comparative low cost, ruggedness, and freedom from the need of power supply, and can, moreover, be readily made in portable forms when desired.

Dynamic and Electrical Characteristics

It will be appreciated from the earlier discussion that, for a volume indicator to be truly standard, its dynamic and electrical characteristics must be controlled and specified so that different instruments will read alike on the rapidly varying speech and program waves. Therefore, the next step in the development was to determine suitable values for these characteristics.

In deciding upon the dynamic characteristics, an important factor included in the consideration was the ease of reading the instrument and the lack of eye strain in observing it for long periods.

First, a number of existing instruments were studied, including some experimental models constructed independently for the two broadcasting companies prior to the start of this joint development. In

this, the opinions of technicians, accustomed to reading volume indicators as a part of their regularly assigned duties, were sought, as well as those of the engineers. The instruments studied included a considerable range of speeds of response and of damping. From this work, the following conclusions were reached:

a. For ease of reading and minimum of eye fatigue, the movement should not be too fast. As a result of observations under service conditions and other tests the requirement was adopted that the sudden application of a 1000-cycle sine-wave of such amplitude as to give a steady deflection at the scale point where the instrument is to be read, shall cause the pointer to read 99 per cent of the final deflection in 0.3 second.

b. The movement shall be slightly less than critically damped, so that the pointer will overswing not less than 1 per cent nor more than 1.5 per cent when the above sine-wave is suddenly applied.

This last point deserves further discussion. It was noted that on speech or program waves, instruments which were critically damped or slightly overdamped had a more "jittery" action than instruments slightly underdamped, and the strain of reading them was greater. The reason for this will be understood by reference to the theoretical curves shown in Fig. 10. These curves represent, for three different degrees of damping, the deflection versus time following the sudden application of a steady sine-wave. Curve *A* is for a movement underdamped by the amount specified above. Curve *B* is for a critically damped movement, while curve *C* is for a movement which is overdamped by the same factor that *A* is underdamped. It is assumed that the periods of the three movements are so adjusted that all reach a deflection of 99 per cent in the same time and that the sensitivities of each are the same.

It will be noted that the velocity of the pointer in curve *A* is more nearly uniform than in the other curves, and that the maximum velocity in *A* is only about half that in *C*. Because of the lower and more uniform velocity, there will be much less eye strain in watching pointer *A* as it dances about in response to program waves than either of the others. Moreover, the same curves inverted will equally well represent the motion of the pointers when the applied wave is suddenly stopped. It is evident, by inspection of the region shown near zero, that pointers *B* and *C* will start downward very rapidly whereas pointer *A* will pause for a moment and then start downward more slowly. This is of importance since it is the maximum excursions of the pointer which must be observed in reading volume levels. The

tendency to pause at the top of the swing before starting downward makes *A* easy to read, and the failure to do so explains the observed "jittery" motion of instruments such as *B* and *C*.

As a further part of this study, high speed moving pictures were taken of the available volume indicators, showing their response to suddenly applied sine-waves. The pictures were taken at 400 frames a second and included on the edge of each frame was a photograph of a clock device which indicated time in thousandths of a second. From

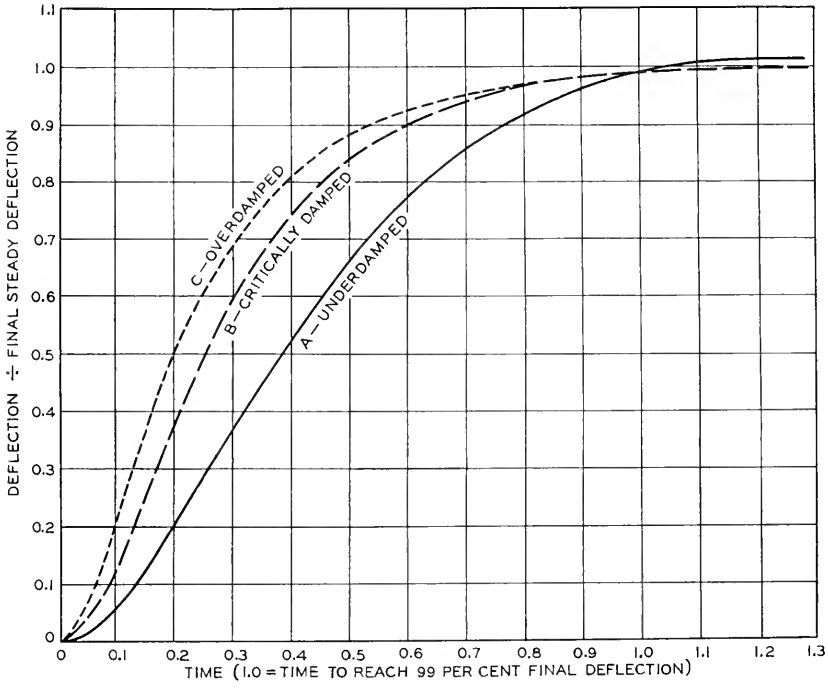


Fig. 10—Effect of damping on instrument characteristics.

measurements made on these films, the data plotted in Fig. 11 were obtained. It is interesting to observe how lightly damped are the oscillations of the 203C volume indicator, which until the advent of the new instrument has been in use in considerable numbers. The curve for the peak volume indicator on Fig. 11 must not be mistaken for the true speed of response but is merely the speed with which the instrument reads the charge on the condenser (see Fig. 2). The charge builds up quite rapidly, but the instrument follows in more leisurely fashion as shown. The instrument, as noted earlier, will actually give

a reading of 80 per cent on an impulse of sine-wave as short as .025 second.

The above characteristics were decided upon only after many tests corroborated by field trials under actual working conditions. The validity of the conclusions reached in the tests of earlier r-m-s volume indicators was checked with respect to the new instrument by further tests.

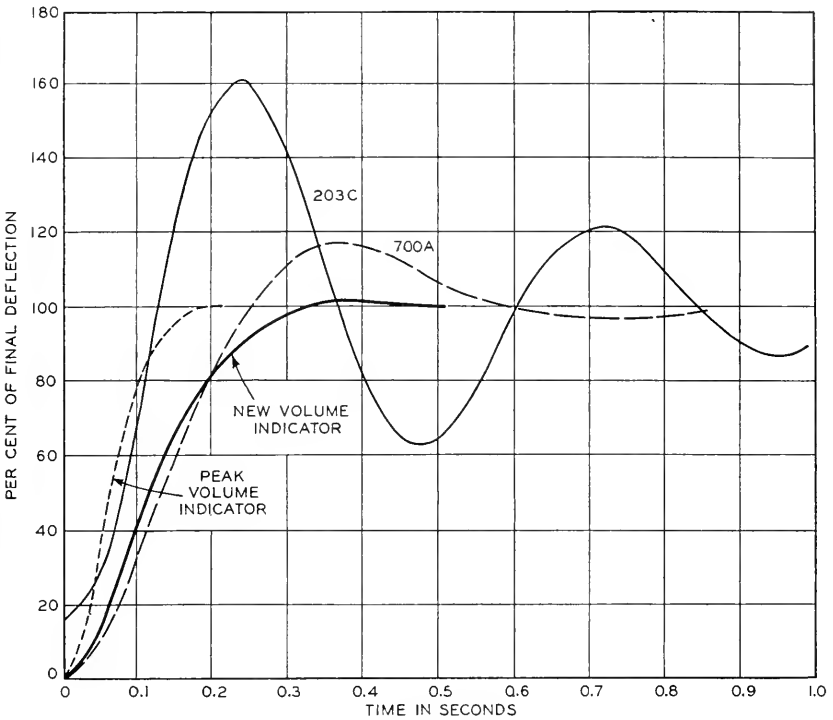


Fig. 11—Deflection of volume indicators to suddenly applied sine-wave.

The question of whether the rectifier should be half-wave or full-wave needs little discussion. The oscillogram of the speech wave shown in Fig. 1 shows a very marked lack of symmetry. Evidently if a volume indicator is to give the same reading no matter which way its input is poled, a balanced full-wave rectifier is required.

Throughout this paper, the term "r-m-s" has been used loosely to describe the general type of instrument under consideration. Some tests were made to determine how closely the new volume indicator approximates this characteristic.

The procedure was based on determining the exponent "p" in the equation

$$i = ke^p,$$

which is equivalent to the actual performance of the instrument for normal deflections. (In the equation "i" is the instantaneous current in the instrument coil and "e" is the instantaneous potential applied to the volume indicator.) Two methods were employed. One consisted of determining the ratio of the magnitudes of the sine-wave a.-c. and the d.-c. potential which when applied to the volume indicator give the same deflection. The second method consisted of determining the ratio of the single frequency potential to the potential of each of two equal amplitude, non-harmonically related frequencies which when simultaneously applied give the same deflection.

Without going into the mathematics involved, several of the new volume indicators were found to have average exponents of about 1.2, so that they had characteristics that were between a linear ($p = 1$) and a square law or "r-m-s" ($p = 2$) characteristic. Applying the second method to a Western Electric 1G Volume Indicator, which is considered to be an "r-m-s" instrument, the exponent was found to be 1.89.

Instrument Scale

Among the more important features to be considered in the development of a volume indicator is the design of its scale. In broadcast studios, volume indicators are under observation almost continuously by the control operators, and the ease and accuracy of reading, and the degree of eye strain are of major importance.

Prior to the adoption of the new standard volume indicator there was a wide variety of volume-level indicator scales in use by the electrical communications industry. This, coupled with the use of a number of different kinds of instruments, reference levels, etc., resulted in considerable confusion when volume measurements were involved.

Volume level indicators, as already explained, are used (*a*) as an aid in compressing the wide dynamic range of an original performance to that of the associated transmission medium and (*b*) for locating the upper part of the dynamic range just within the overload point of an equipment during its normal operation. For the first of these uses, a scale having a wide decibel range is preferable. For the latter purpose, a scale length of 10 db is usually adequate. Since a given instrument may be used for both applications, neither too large nor too small a range is desirable in a volume level indicator for the above purposes. A usable scale length covering 20 db appears to be a satisfactory compromise.

It is evident that the instrument scale should be easy to read in order that the peak reached by the needle under the impetus of a given impulse may be accurately determined. The instrument scale, therefore, should be as large as practical since in the case of the broadcast and motion picture applications, attention is divided between the action in the studio and the volume indicator.

The instrument scale graduations should convey a meaning, if possible, even to those not technically inclined but who are, nevertheless, concerned with the production of the program material.

Finally, the scale must be properly illuminated so that the relative light intensity on the face of the instrument is comparable to that on the sound stage. Unless this condition prevails, the eye will have difficulty in accommodating itself with sufficient rapidity to the changes in illumination as the technician glances back and forth from the studio to the volume-indicator instrument.

Existing Scales

The volume-indicator scales most commonly employed in the past are shown in Figs. 12, 13, 14 and 15. It is evident that all these scales differ from each other in one or more respects.

Fig. 12—Scale on 203C volume indicator.

The color combinations employed for the scale shown in Fig. 12 and the simplicity of its markings are outstanding virtues. The division markings and the numerals of the main scale are black on a yellow

background. The decibel divisions and associated numerals are in red and considerably less conspicuous than the main scale.

However, the 0 to 60 scale, which is used on both of the instruments shown in Figs. 12 and 13, is an arbitrary one bearing no simple relation to the electrical quantity being measured. Because of this, some of the non-technical persons concerned with program production are prone to request that a certain "effect" which they desire to transmit at a louder-than-normal level be permitted to swing the indicating

Fig. 13—Scale on 21 type volume indicator.

needle beyond the normal reference point of "30" on the scale. It is not evident to them from the instrument scale that the normal reading of "30" corresponds to maximum "undistorted" output of the system.

The scale shown in Fig. 14, on the other hand, was primarily intended for steady-state and not volume level measurement purposes. Consequently, this scale has little, if anything, to commend it for program monitoring use. Nevertheless, the simplicity and the fine electrical features of this type of instrument, together with its relatively reasonable cost, have resulted in its general application to volume

indicator service. It is evident, however, that the scale card, which contains all kinds of identification data, is entirely too confusing for quick, accurate observations as the needle swings rapidly back and forth across the scale.

The scale shown in Fig. 15 has the merit of simplicity and easy readability. It is, however, somewhat limited in the decibel range appearing on the scale.

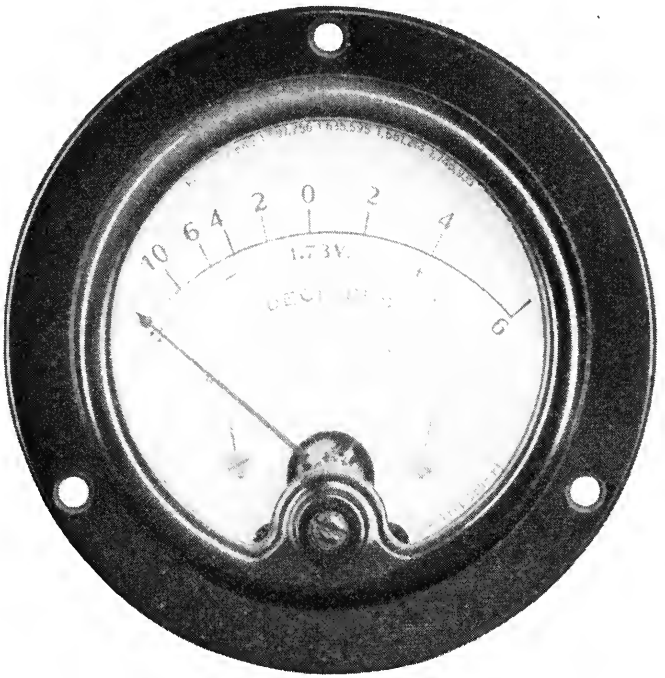


Fig. 14—Scale on type 586 power level indicator.

New Scale

Both vu [5] markings and markings proportional to voltage are incorporated in the new instrument scale. The need for the former is obvious, but the philosophy which leads to the inclusion of the latter requires an explanation.

It is evident, assuming a linear system, that the voltage scale is directly proportional to percentage modulation of a radio transmitter upon which the program is finally impressed. If the system is adjusted for complete modulation for a deflection to the 100 per cent

[5] Defined later.

mark, then subsequent indications show the degree of modulation under actual operating conditions. In the interests of best operation, it may be desirable, of course, to adjust the system for somewhat less than complete modulation when the 100 per cent indication is reached.

In any event, the indications on the voltage scale always show the *percentage utilization of the channel.* This is a decided advantage because everyone involved has a clear conception of a percentage indication. Furthermore, since the scale does not extend beyond the 100

Fig. 15—Type of scale used on 1G and 700A volume indicators.

per cent mark (except in the form of a red warning band) and since it is impossible to obtain more than 100 per cent utilization of the facilities, there is no incentive on the part of non-technical people connected with program origination to request an extra loud "effect" on special occasions.

Actually, two scales, each containing both vu and voltage markings, have been devised. One of these, known as the type A scale, Fig. 16, emphasizes the vu markings and has an inconspicuous voltage scale. The second, known as the type B, Fig. 17, reverses the emphasis on

the two scales. This arrangement permits the installation of the instrument which features the scale that is most important to the user, while retaining the alternate scale for correlation purposes.

The new scale retains the simplicity and the general color scheme of the former Fig. 12 scale. The main division markings and the associated numerals are, in each case, in black. The secondary data are

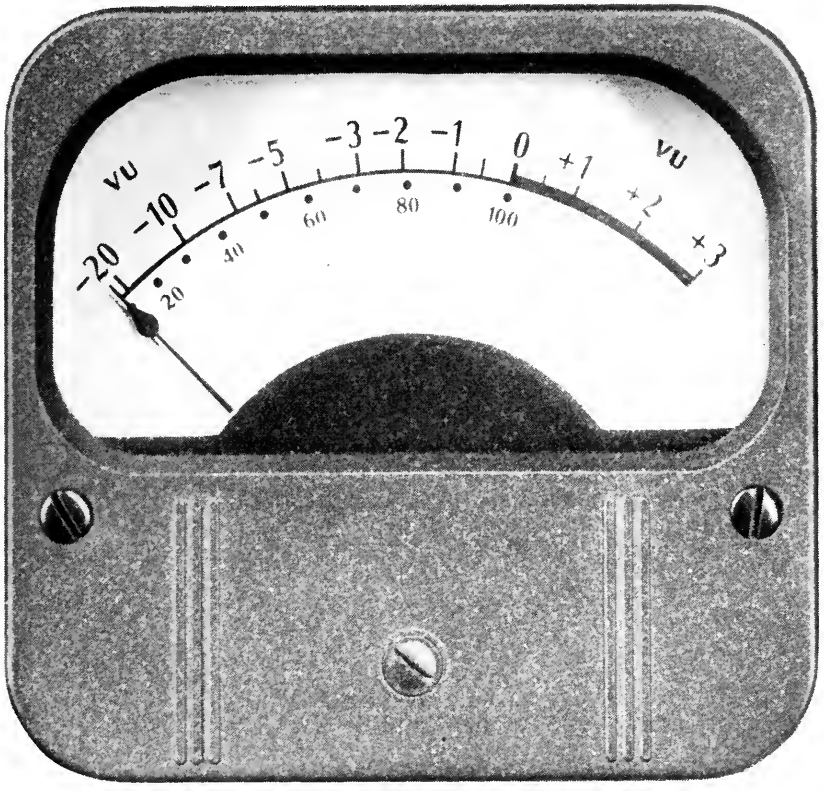


Fig. 16—New volume indicator—A scale.

smaller (and in one case are in red) and therefore less conspicuous than the others. All irrelevant markings have been omitted from the scale.

The color of the scale card, which is a rich cream, seems to be a satisfactory compromise between high contrast and reduced eye-strain and fatigue. This choice is based upon the preference of a large group of skilled observers and upon the reports of certain societies for the improvement of vision. The use of matte finished instrument

cases having fairly high reflection coefficients, such as light grey, is also desirable for ease of vision.

The location of the "reference" point is such that 71 per cent of the total scale length is utilized as compared to only 42 per cent in former instruments. This feature, combined with the use of a larger size instrument, results in a useful scale more than 2.5 times the length of former scales.

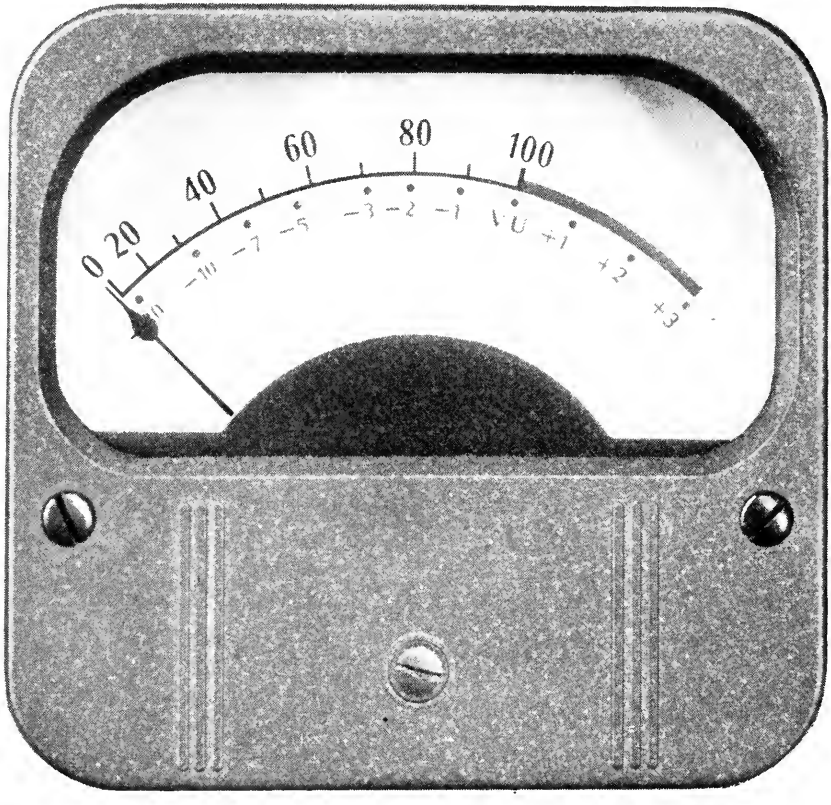


Fig. 17—New volume indicator—B scale.

Although the reference point is no longer in the traditional vertical or near-vertical position, it has been found that even those who have long been accustomed to the old arrangement, soon discover the advantages of the new scale. This is attested, in the case of the broadcasting application, by the general acceptance of this scale by the personnel of a number of stations located in various sections of the country.

A small but important feature of the new scale is the use of an arc to connect the lower extremities of the vertical black division marks. This arc affords a natural path along which the eye travels as it watches the needle flash up and down the scale. The omission of this arc would result in a number of vertical division marks, hanging in space, as obstacles to the free back-and-forth motion of the eye.

It is evident, upon comparison of Figs. 12, 13, 14 and 15 with Figs. 16 and 17, that the dynamic volume range visible on the scale is at least twice as great as on former instruments. This range, as already explained, is a good median value for general use.

Mention was made of the opinions of a group of skilled observers. This group consisted of more than 80 broadcast technicians who, in the performance of their duties, watch volume-indicator instruments almost continuously throughout the working day. The opinions of this group were obtained by submitting working models for their individual considerations. It is believed that some of the results of these observations are of interest.

1. 83 per cent preferred the cream in place of a white scale card.
2. 90 per cent preferred the "0–100" scale to the "0–60" scale.
3. 92 per cent preferred the longer scale length (3.5" vs. 2.36").
4. 97 per cent preferred the numerals placed above the arc.
5. 50 per cent preferred the spade pointer to the lance type.
6. 93 per cent agreed on the adequacy of 3 db leeway above the reference point.

New Reference Level and Terminology

Having agreed on the characteristics of the new standard volume indicator, the interests of complete standardization call for agreement, as well, upon a uniform method of use and a uniform terminology. Agreement upon a uniform method of use must include establishing the reference volume or zero volume level to which the readings are to be referred and agreeing upon the technique of reading the volume indicator.

It is important to appreciate that "reference volume" is a useful practical concept, but one which is quite arbitrary and not definable in fundamental terms. For example, it cannot be expressed in any simple way in terms of the ordinary electrical units of power, potential, or current, but is describable only in terms of the electrical and dynamic characteristics of an instrument, its sensitivity as measured by its single frequency calibration, and the technique of reading it. In other words, a correct definition of reference volume is *that level of*

program which causes a standard volume indicator, when calibrated and used in the accepted way, to read 0 vu.

It is especially cautioned that reference volume as applied to program material should not be confused with the single frequency power used to calibrate the zero volume setting of the volume indicator. If a volume indicator is calibrated so as to read zero on a sine-wave power of, say, one milliwatt in a stated impedance, a speech or program wave in the same impedance whose intensity is such as to give a reading of zero will have instantaneous peaks of power which are several times one milliwatt and an average power which is only a small fraction of a milliwatt. It is therefore erroneous to say that reference volume in this case is one milliwatt. Only in the case of sine-wave measurements does a reading of 0 vu correspond to one milliwatt.

It should be emphasized that, although it is convenient to measure the performance of amplifiers and systems by means of single frequencies, there is no exact universal relationship between the single frequency load carrying capacity indicated by such measurements and the load carrying capacity for a speech and program waves expressed in terms of volume level. This relationship depends upon a number of factors such as the rapidity of cutoff at the overload point, the frequency band width being transmitted, the quality of service to be rendered, etc.

It has already been brought out that in the past there have been a multiplicity of reference volumes differing from each other not only because of the various single frequency calibrations which have been employed, but also because of the dissimilar dynamic characteristics of the different instruments used to measure volume levels. It is also apparent that the introduction of a new volume indicator whose characteristics are not identical with any of its predecessors inherently means the introduction of a new reference volume no matter how it is calibrated. Therefore, there did not seem to be any compelling reason to make the calibration of the new instrument agree with any of the calibrations used in the past. Moreover, to many there seemed to be some advantage in setting the new reference level at a sufficiently different order of magnitude from those which had been in most common use, so that there will be little chance of confusing the new standards with any of those that went before.

After much thought and discussion, it was agreed that the new reference volume should correspond to the reading of the new volume indicator when calibrated with one milliwatt in 600 ohms across which the volume indicator is bridged. Other calibrating values considered were 10^{-16} watt, 6 milliwatts and 10 milliwatts, in 600 ohms or in

500 ohms. The value chosen was preferred by a majority of a large number of people who were consulted and in addition was found to be the only value to which all could agree. Some of the reasons for choosing 1 milliwatt (10^{-3} watt) were: (1) It is a simple round number, easy to remember; (2) 10^{-3} is a preferred number; [6] (3) 1 milliwatt is a much used value for testing power for transmission measurements, especially in the telephone plant, so that choice of this value therefore permits the volume indicators to be used directly for transmission measurements.

The choice of the standard impedance of 600 ohms was influenced by the fact that, considering all of the plants involved, there is more equipment designed to this impedance than to 500 ohms.

The question may very well be raised why the reference volume has been related to a calibrating *power* rather than to a calibrating *voltage*, inasmuch as a volume indicator is generally a high impedance, voltage responsive device. A reference level could conceivably be established based on voltage and the unit of measurement might be termed "volume-volts." However, volume measurements are a part of the general field of transmission measurements, and the same reasons apply here for basing them on power considerations as in the case of ordinary transmission measurements using sine-waves. If the fundamental concept were voltage, apparent gains or losses would appear wherever impedance transforming devices, such as transformers, occur in a circuit. This difficulty is avoided by adopting the power concept, making suitable corrections in the readings when the impedance is other than 600 ohms.

Having chosen the zero point to which the new volume readings would be referred, the next question to be decided was the terminology to be employed in describing the measurements. As has been pointed out, the past custom of describing the volume measurements as so many decibels above or below reference level has been ambiguous because of differences in instruments and standards of calibration. It was thought, therefore, that there would be less confusion in adopting the new standards if a new name were coined for expressing the measurements. The term selected is "vu," the number of vu being numerically the same as the number of db above or below the new reference volume level. It is hoped that in the future *this new term will be restricted to its intended use so that, whenever a volume level reading is encountered expressed as so many vu, it will be understood that the reading was made with an instrument having the characteristics of the new volume indicator and is expressed with respect to the new reference level.*

[6] A. Van Dyck, "Preferred Numbers," *Proc. I. R. E.*, Vol. 24, pp. 159–179 (1936).

The procedure for reading the new volume indicator is essentially the same as that which has always been employed, with the exception that, since the instrument is very nearly critically damped, there need be tolerated fewer overswings above the prescribed deflection. One who is familiar with the use of volume indicators will instinctively read the new instrument correctly. The procedure may be described by stating that the adjustable attenuator, which is a part of the volume indicator, should be so adjusted that the extreme deflections of the instrument needle will just reach a scale reading of zero on the vu scale or 100 on the per cent voltage scale. The volume level is then given by the designations numbered on the attenuator. If, for any reason, the deflections cannot be brought exactly to the 0 vu mark or 100 per cent mark, the reading obtained from the setting of the attenuator may, if desired, be corrected by adding the departure from 0 shown on the vu scale of the instrument.

Since program material is of a very rapidly varying nature, a reading cannot be obtained instantaneously but the volume indicator must be observed for an appreciable period. It is suggested that a period of one minute be assumed for this purpose for program material, and 5 to 10 seconds for message telephone speech, so that the volume level at any particular time is determined by the maximum swings of the pointer within that period.

Summary of Characteristics

In the preceding sections of the paper the considerations which led to the selection of the more important characteristics of the new volume indicator have been discussed in some detail. In this section a summary will be made, first of the fundamental requirements which must be conformed to by any instrument if it is to be a standard volume indicator according to the new standards, and secondly, of other requirements which have been specified for the new volume indicators which are perhaps matters more of engineering than of a fundamental nature. These requirements are a condensation of the more important features of the specifications for the new insturment. The Weston Electrical Instrument Corporation generously cooperated in the development, but it is emphasized that the specifications are based on fundamental requirements and are not written on the product of a particular manufacturer. The complete requirements are available to any interested party, and, as a matter of fact, at least one other manufacturer has produced an instrument which meets the requirements.

(A) *Fundamental Requirements*

1. Type of Rectifier

The volume indicator must employ a full-wave rectifier.

2. Scales

The face of the instrument shall have one of the two scale cards shown in Figs. 16 and 17. Both cards shall have a "vu" scale and a "percentage voltage" scale. The reference point at which it is intended normally to read the instrument is located at about 71 per cent of the full scale arc. This point is marked 0 on the vu scale and deviations from this point are marked in vu to + 3 and to − 20. The same point is marked 100 on the other scale which is graduated proportionately to voltage from 0 to 100.

3. Dynamic Characteristics

If a 1000-cycle voltage of such amplitude as to give a steady reading of 100 on the voltage scale is suddenly applied, the pointer should reach 99 in 0.3 second and should then overswing the 100 point by at least 1.0 and not more than 1.5 per cent.

4. Response vs. Frequency

The sensitivity of the volume-indicator instrument shall not depart from that at 1000 cycles by more than 0.2 decibel between 35 and 10,000 cycles per second nor more than 0.5 decibel between 25 and 16,000 cycles per second.

5. Calibration

The reading of the volume indicator (complete assembly as shown schematically in Fig. 18) shall be 0 vu when it is connected to a 600-ohm resistance in which is flowing one milliwatt of sine-wave power at 1000 cycles per second, or n vu when the calibrating power is n decibels above one milliwatt.

(B) *Specific Requirements*

1. General Type

The volume indicator employs a d.-c. instrument with a non-corrosive full-wave copper-oxide rectifier mounted within its case.

2. Impedance

The impedance of the volume indicator arranged for bridging across a line is about 7500 ohms when measured with a sinusoidal voltage sufficient to deflect the pointer to the 0 vu or 100 mark on the scale. Of this impedance 3900 ohms is in the meter and about 3600 ohms must be supplied externally to the meter.

3. Sensitivity

The application of a 1000-cycle potential of 1.228 volts r-m-s (4 decibels above 1 milliwatt in 600 ohms) to the instrument in series with the proper external resistance causes a deflection to the 0 vu or 100 mark. The instrument therefore has sufficient sensitivity to be read at its normal point (0 vu or 100) on a volume level of + 4 vu.[7]

4. Harmonic Distortion

The harmonic distortion introduced in a 600-ohm circuit by bridging the volume indicator across it is less than that equivalent to 0.2 per cent (r-m-s).

5. Overload

The instrument is capable of withstanding, without injury or effect on calibration, peaks of 10 times the voltage equivalent to a deflection to the 0 vu or 100 mark for 0.5 second and a continuous overload of 5 times the same voltage.

6. Color of Scale

The color of the scale card, expressed according to the Munsell system of color identification, is $2.93Y\ \frac{9.18}{4.61}$.[8]

7. Presence of Magnetic Material

The presence of magnetic material near the movements of the instruments as now made will affect their calibrations and dynamic characteristics. This is because it has been necessary to employ more powerful magnets than usually required for such instruments to obtain the desired sensitivity and dynamic characteristics, and any diversion of flux to nearby magnetic objects effectively weakens the useful magnetic field beyond the point where these characteristics can be met. The instruments should not, therefore, be mounted on steel panels. (The effect is only slight if they are mounted on 1/16 inch panels with the mounting hole cut away as far as possible without extending beyond the cover of the meter.)

[7] There should be no confusion because the instrument deflects to a scale marking of 0 vu when a level of + 4 vu is applied to it. As in previous volume indicators, the 0 vu point on the vu scale is merely an arbitrary point at which it is intended nominally to read the instrument, and the rest of the vu scale represents deviations from the 0 vu point. The volume level is read, not from the scale, but from the indications on the associated sensitivity control when the latter is so set as to give a scale deflection to the 0 vu mark. If a deflection other than 0 vu is obtained, the volume level may be corrected by the deviation from 0 vu shown on the instrument scale. In the present art, it is difficult to make an instrument of the desired characteristics having a sensitivity greater than that indicated.

[8] Munsell Book of Color, Munsell Color Company, Baltimore, Maryland, 1929.

8. Temperature Effects

In the instruments now available, the deviation of the sensitivity with temperature is less than 0.1 decibel for temperatures between 50° F. and 120° F., and is less than 0.5 decibel for temperatures as low as 32° F.

Description of Circuits

The new instrument by itself does not constitute a complete volume indicator but must have certain simple circuits associated with it. Two forms which these circuits may take are illustrated in Fig. 18.

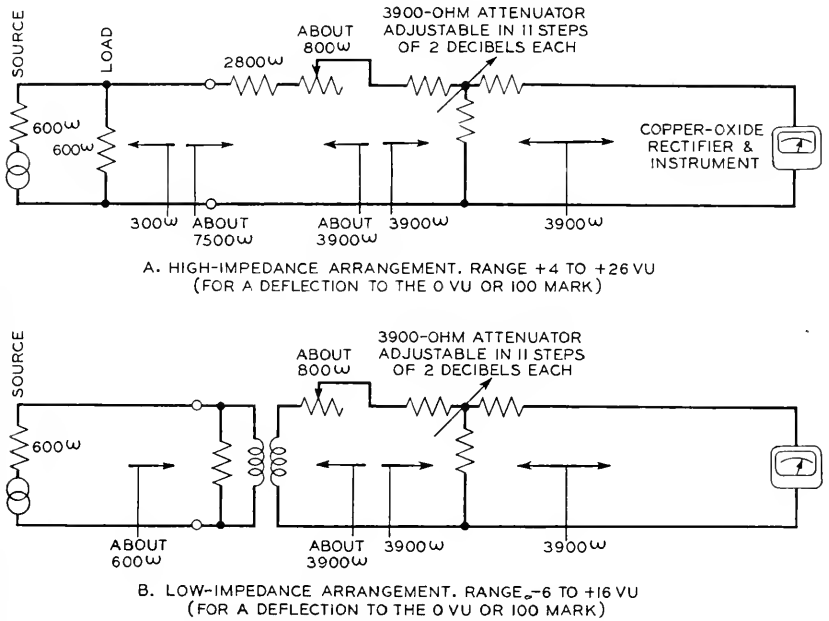


Fig. 18—Circuits for new volume indicator.

One volume indicator may, of course, have both circuits with arrangements to select either by means of a key or switch.

Diagram 18A shows a high impedance arrangement intended for bridging across lines. As noted above, about 3600 ohms of series resistance has been removed from the instrument and must be supplied externally in order to obtain the required ballistic characteristics. This was done in order to provide a point where the impedance is the same in both directions, for the insertion of an adjustable attenuator. A portion of the series resistance is made adjustable as shown by the slide wire in the diagram. This is for the purpose of facilitating ac-

curate adjustment of the sensitivity to compensate for small differences between instruments and any slight changes which may occur with time. The particular arrangement shown in the diagram has an input impedance of about 7500 ohms and a range of + 4 to + 26 vu for readings at the 0 vu or 100 mark on the instrument scale.

Diagram 18B shows a low impedance arrangement in which by adding a transformer the sensitivity has been increased by 10 vu at the expense of decreasing the input impedance to 600 ohms. The circuit is so designed that the impedance facing the instrument is the same as in diagram A, so that the proper dynamic characteristics are obtained. This arrangement, being low impedance, cannot be bridged across a through line, but must be used where it can terminate a circuit. It is

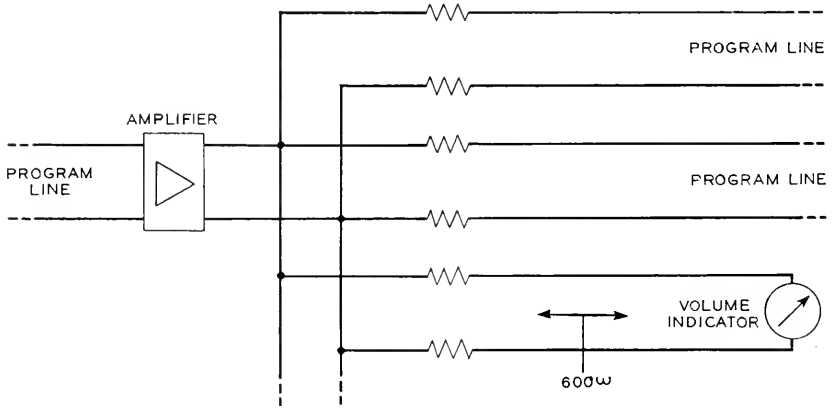


Fig. 19—Program bridge for feeding several lines from one line.

useful for measuring the transmission loss or gain of a circuit on sine-wave measuring currents, and also for measurements of volume level where it is connected to a spare outlet of a program bridge circuit, as shown in Fig. 19. Program bridge circuits, one form of which is illustrated in the figure, are commonly employed in the Bell System when it is desired to feed a program from one line simultaneously into a number of other lines. The bridge circuit which is illustrated consists of a network of resistances so designed that the volume level into each of the outgoing lines is the same, that the impedance presented to each is the correct value of 600 ohms, and that the attenuation through the network between any two of the outlets is great.

A picture of a volume indicator which is provided with both of the circuits shown in Fig. 18 is illustrated in Fig. 20.

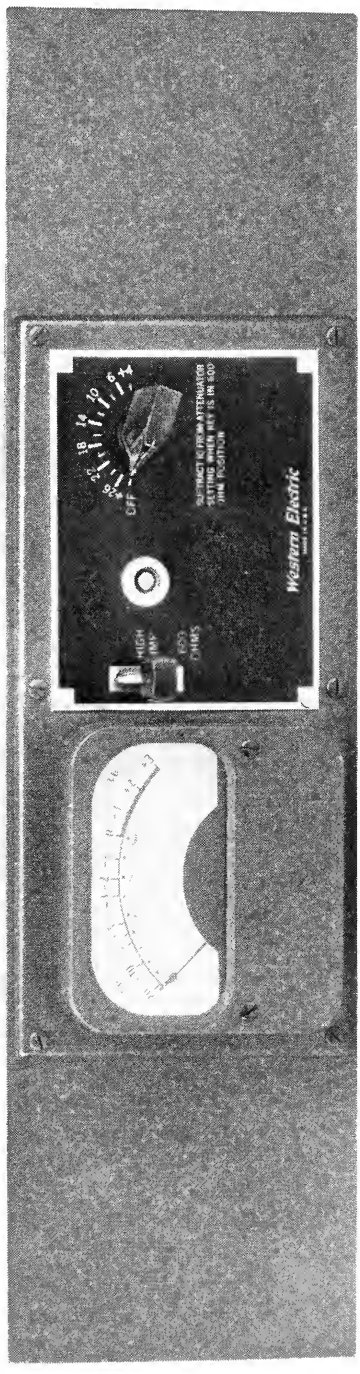


Fig. 20—754B volume indicator equipped with new standard instrument having "A" (Bell System) scale.

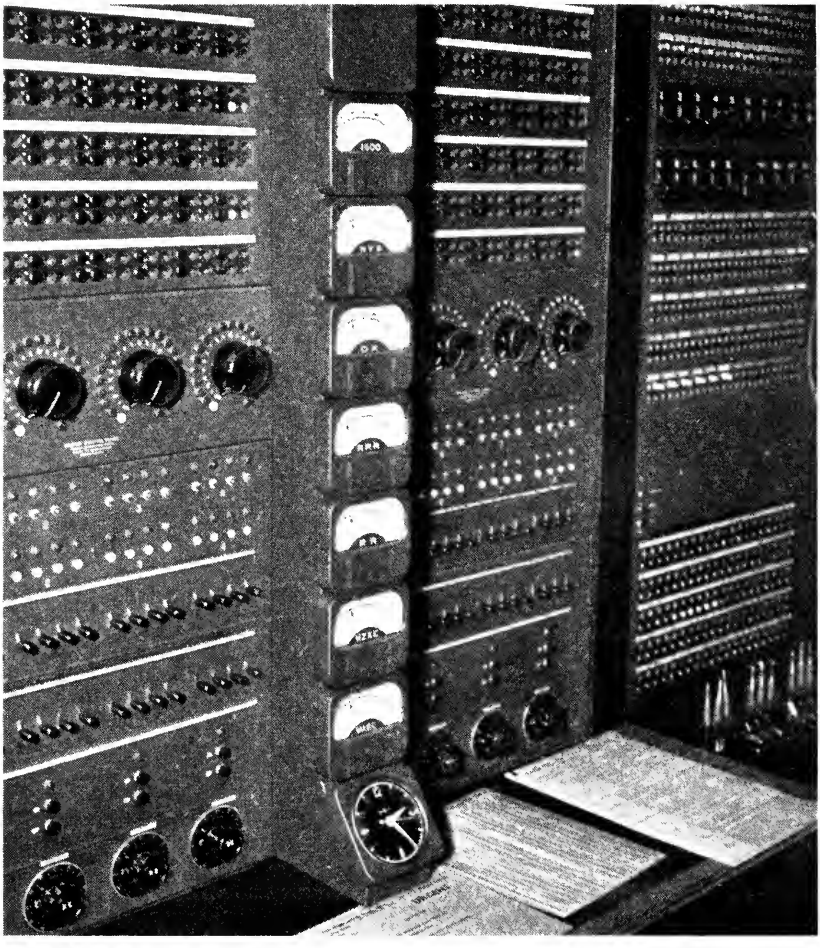


Fig. 21—New standard volume indicators installed at a network key station.

Fig. 21 shows a group of new standard volume indicators installed in a network key station.

Conclusion

This paper has described a new volume indicator which is inexpensive and whose characteristics are thought to represent a good practical compromise for a general purpose instrument of this kind. It has been commented upon favorably by all who have had any experience with it. It has been adopted as standard by the two largest broadcasting companies and the Bell System, and it is hoped that other users of volume

indicators will be sufficiently impressed by the merits of the new instrument and by the desirability of standardization in this field, to join in its adoption. The new standards are being submitted to the standards committees of the various national organizations for adoption.

Many people contributed to the development which has been described. In particular the authors wish to express their appreciation to Messrs. Robert A. Bradley of the Columbia Broadcasting System, George M. Nixon of the National Broadcasting Company, and S. Brand and Iden Kerney of the Bell Telephone Laboratories, for their important share in the work, and to the Weston Electrical Instrument Corporation for its valued cooperation.

Metallic Materials in the Telephone System*

By EARLE E. SCHUMACHER and W. C. ELLIS

IN the development of electrical communication, metals and alloys have played a noteworthy part. To emphasize specifically the utilization of metallic materials the telephone handset serves as an admirable example. The assembly of intricate parts in this small piece of apparatus, shown sectionalized in Fig. 1, contains seventeen metallic elements, either alone or in combination as alloys.

The Bell System has therefore conducted extensive metallurgical researches, and the discoveries and developments have been numerous. Space permits a discussion of only a few of the developments relating to the more extensively used materials. These comprise the alloys of lead, copper, zinc and aluminum, and the precious metals, and magnetic materials.

Lead and Alloys of Lead

Lead alloys are used principally as sheathing for cable, and as solders for joining cable sheath and making electrical connections in apparatus.

Cables represent one of the largest single items of investment; approximately ninety-five per cent of the Bell System's total wire mileage is contained in lead or lead alloy sheath and this sheath requires an enormous amount of lead annually in its production. The largest size cable made by the System contains 4242 copper wires. The same number of open wires on telephone poles would take 70 rows of poles each carrying 60 wires. Under one street today in New York City there are 282 cables containing about 560,000 wires.

Since the wires in the cable are insulated from one another only by the paper or textile wrappings or sheaths and by the dry air contained in the cable, the presence of even a slight amount of moisture will interfere with transmission by drastically reducing the insulation resistance. A positive pressure of dry nitrogen is maintained in some cables as additional protection against moisture entrance and to disclose sheath breaks. Continued efforts are made, therefore, to improve cable sheath so as to keep sheath failures to a minimum.

* Based upon a paper published in *Metal Progress*, November 1939.

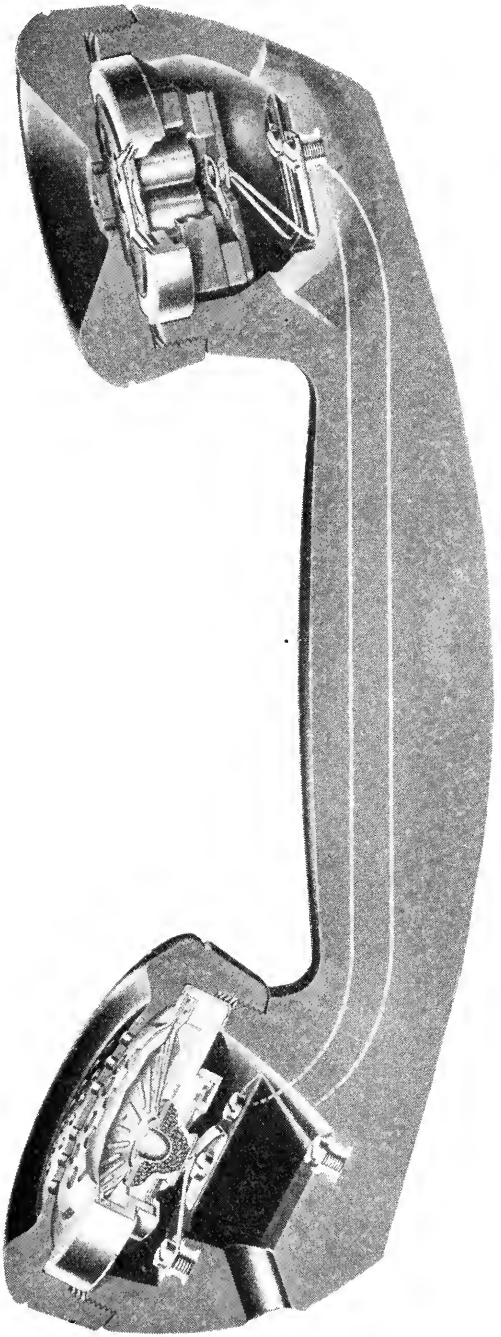


Fig. 1—Schematic cross-section of handset showing utilization of metallic materials.

The history of cable sheath development illustrates the value of metallurgical research to the telephone system. Unalloyed lead was first used because it was pliable, resistant to corrosion and could easily be manufactured into pipe. Nevertheless, it has serious shortcomings. Brittleness would not be expected in a material so soft and ductile, yet repeated stresses caused by wind sway, mechanical vibrations, and

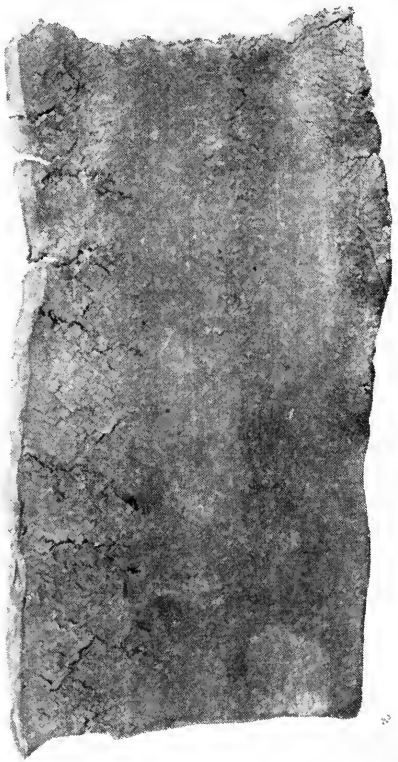


Fig. 2—View of piece of old cable sheath made of commercially pure lead, which failed in service from intercrystalline fracture.

movements due to temperature changes produce fine cracks in the cable sheath through which moisture may enter the cable. An advanced stage of such cracking is shown in Fig. 2. In fact this effect is so serious that, unless precautions are taken to minimize vibration, cables sheathed with unalloyed lead cannot be shipped for long distances by rail or boat without serious damage.

It was early found that the addition of three per cent of tin to lead greatly decreased the susceptibility to this type of failure. This alloy

was also stronger than lead and more resistant to abrasion and the cutting action of the galvanized steel rings which usually fasten aerial cable to its supporting strand. As the quantity of alloy required for cable sheathing increased, however, it became evident that a large portion of the world's supply of tin would be needed, and this would cause a prohibitive rise in its price. A search was made, therefore, for an alloy of at least equal quality which would be less expensive.

As a result of investigation of the properties of twenty or more different alloys, an alloy of lead containing one per cent antimony was selected. After extensive manufacturing and field trials this alloy was adopted in 1912 as the standard for Bell System use. Had the lead-tin alloy been continued as a sheathing material to the present time the cost would have been twenty-five million dollars greater (figured on the amount of cable sheath used during the intervening years and on the price of tin which actually prevailed during this time).

Standardization of an alloy of lead with one per cent antimony for cable sheath was not accomplished without the appearance and solution of many technical problems. For example the extrusion of sheath around the cable core has been an intermittent process, since the cylinder of the extrusion press is not large enough to contain sufficient lead to cover a full length of cable. It was necessary, therefore, to stop extrusion to recharge the cylinder with the molten lead alloy which must weld to the previous charge, a slug of solid metal. If a layer of dross was present on the surface of this material remaining in the cylinder, a faulty weld was formed which would be subsequently extruded into the sheath. Also, during the recharging interval, the lead alloy remaining in the extrusion die receives a different thermal treatment from that of the previously extruded sheath. Since the properties of the lead-one per cent antimony alloy are markedly affected by thermal treatment, there were frequently abrupt differences in stiffness of the sheath extruded just before and just after the charging interval. When this change in stiffness was sufficiently great, serious buckles occurred during reeling and installation of the cable.

Through a knowledge of the constitution and characteristics of the alloy, and by continual improvement in the extrusion process, it has been possible to overcome obstacles such as these and to manufacture cable sheath of improved quality from the one per cent antimony alloy.

The telephone metallurgist is also concerned with the life of the alloy in service. Many samples from sheath which has failed are examined annually and compared with samples from sheath which is giving satisfactory service. Microscopic examination in some in-

stances reveals a clue to the causes producing early failure and thus suggests methods by which the failures may be eliminated.

In developing new alloys such as have been described and in studying the causes of failure of these alloys in service, extensive laboratory facilities are required. For example, the Bell Telephone Laboratories possess an extrusion press, shown in Fig. 3, for experimental studies

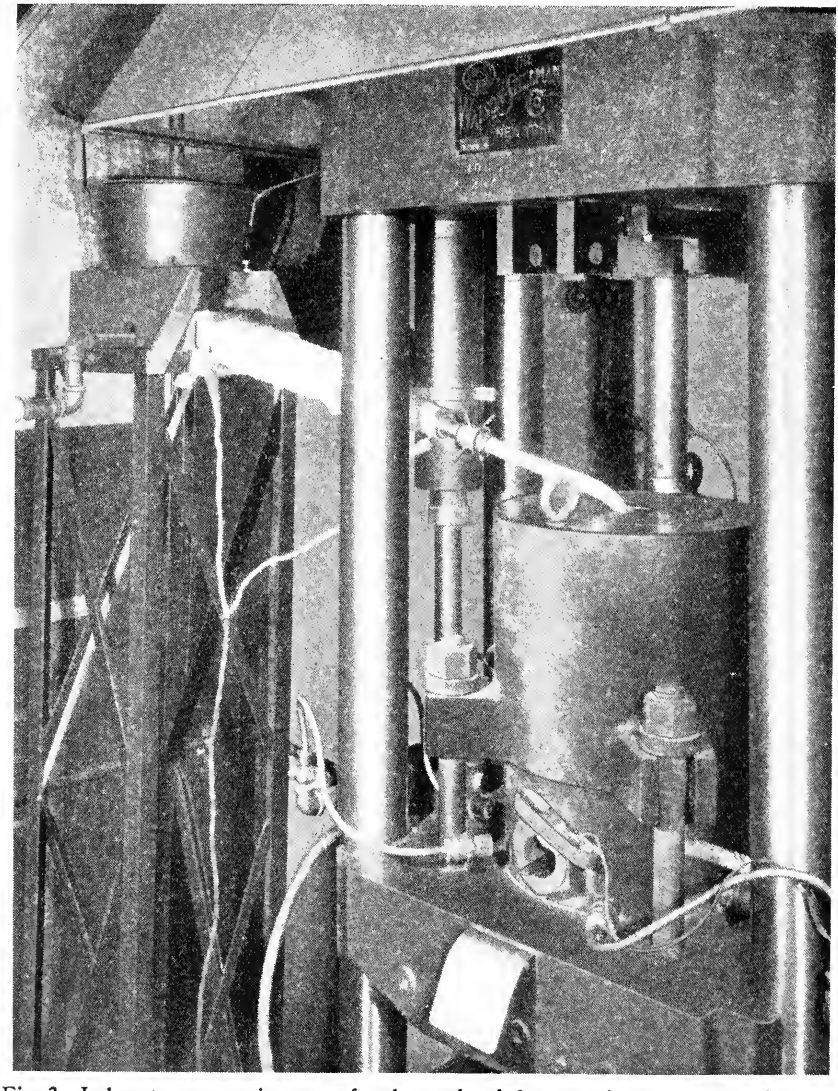


Fig. 3—Laboratory extrusion press for the study of the extrusion process and for the production of experimental cable sheathing alloys.

and for the preparation of new cable sheath alloys. With this equipment commercial extrusion conditions can be investigated or, when desired, extrusion conditions can be varied to determine the effect on the properties of the alloy.

The general layout of the metallurgical microscopic laboratory is shown in Fig. 4. In the foreground is a metallurgical microscope and camera equipped with facilities for examination with polarized light and dark field illumination. The preparation of specimens and photographic processing are done in conveniently arranged adjoining rooms. The microscopic equipment is complemented with X-ray diffraction apparatus shown in Fig. 5. This equipment consists of a demountable X-ray tube so arranged that targets can be readily interchanged. Cameras are provided for structure identification, precision determination of lattice constants, and texture and orientation studies.

Microscopic and X-ray diffraction equipment are both extremely valuable in a great diversity of metal problems. Some examples are given here of the utilization of microscopic equipment in cable sheath development studies. The possibilities of prolonging the life of cable sheath which has developed a weakened structure in service have been established through microscopic examination after a heat treatment consistent with the alloy structure. Again, the results of thermal treatment incident to the soldering and repair operations on cable in the field can be observed and used as a guide to the value of certain procedures. An interesting example is concerned with the opening of splices in installed cable sheathed with lead-antimony alloy, a procedure frequently necessary. During aging in service the antimony-rich particles coalesce into relatively large lumps. When material in this condition is heated by pouring hot solder over the joint, pools of liquid are formed around each lump of antimony, and if an attempt is made to pry the sleeve of the splice open at once, the sleeve crumbles. If heating is prolonged a few minutes, however, the tiny antimony-rich liquid pools diffuse into the surrounding solid material; at this time the sleeve can be opened without injury.

A few years ago, a new lead alloy containing from three to four hundredths per cent of calcium was produced and is being extensively studied now for cable sheathing and other applications. Laboratory tests indicate that under some conditions this material excels lead-antimony in resistance to fatigue failure. To illustrate the careful consideration given materials before making changes which might vitally affect telephone service, about one hundred miles of cables sheathed with a lead-calcium alloy have been installed for a commercial field test. In addition, thirty-six thousand feet of experimental

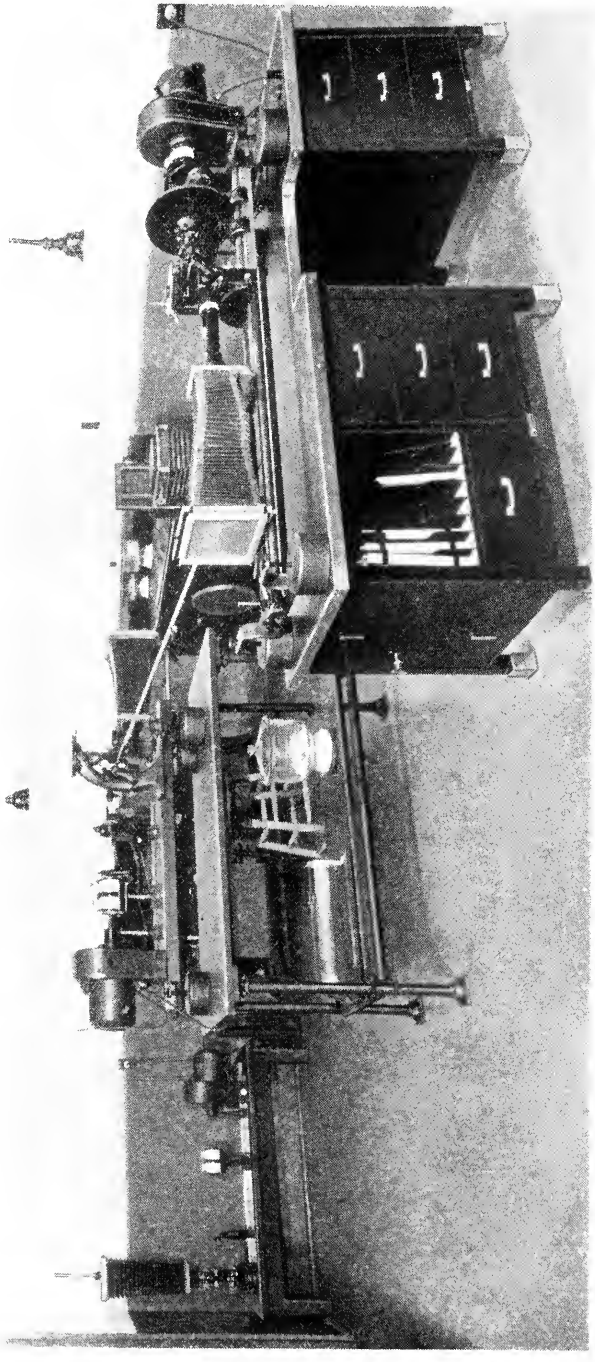


Fig. 4—Metallographic equipment used in the study of metal problems.

lead-calcium sheathed cable were installed on poles alongside of similar lengths of cable with standard lead-antimony sheath. Various sheath thicknesses ranging from .075 to the standard .125 inch were installed for comparison and to expedite early failure. In addition to the comparison between alloys this test will also give information regarding the minimum thickness of sheath which may be employed with both the standard and the experimental alloys.

Fig. 5—X-ray diffraction apparatus showing cameras mounted for identification of structure and precision measurement of lattice constant.

Besides their application as cable sheathing materials, lead alloys are also extensively used by the Bell System as solders, storage battery plates, fuses and as corrosion protection coatings.

Copper and Copper Alloys

Unalloyed copper finds application as wire in the lead-sheathed cables already discussed, in open wire circuits and in central office equipment. In the telephone plant there are eighty million miles of it—enough to span the distance from the earth to the moon three hundred thirty-five times. To obtain the lowest transmission losses, cable conductors consist of high conductivity annealed copper wire.

For line wire in open wire circuits, hard drawn copper wire is used in order to take advantage of the conductivity of copper and the inherently greater strength resulting from strain-hardening. Line wire is subject to ice and wind loads, vibration fatigue, and in some localities, severe corrosion. Where loading conditions are severe the copper-cadmium and other high-conductivity, high-strength materials have attractive possibilities but require further evaluation before their introduction for general use.

For drop wire—the conductor running from the telephone poles to subscribers' buildings—a material with somewhat different properties is required. Here lower conductivities can be tolerated but higher strengths are necessary since the wire is smaller in size and long spans are sometimes necessary. Several materials have been utilized. The alloy most generally in service in the Bell System is composed of 98.25 per cent copper and 1.75 per cent tin. This is being replaced now as a result of research development with a higher strength copper alloy containing 3 per cent tin. This substitution makes possible a reduction in gauge size of conductor from 17 to 18 without sacrifice in the strength characteristics of the conductor.

For most purposes ordinary electrolytic copper containing a fraction of a per cent of oxygen is satisfactory. There are some limited applications, however, where the copper is subjected to high temperatures in the presence of reducing atmospheres at some stage in the manufacturing process. Under these conditions, the presence of oxygen in the ordinary copper produces a well-known embrittling effect. For these applications a copper free from oxygen is used.

A small but important application of copper in telephone circuits is in the production of copper-oxide rectifiers. For this purpose a copper imported from Chile is ordinarily used; for some obscure reason domestic brands of copper have not generally proved so satisfactory.

Copper in the alloyed form also is used extensively in the telephone plant. One application, that for drop wire, has already been mentioned. Other extensive applications are for springs and contacting members in electrical circuits and for structural parts where corrosion resistance or other desired physical properties justify their use. Nickel silver and to a lesser extent phosphor bronze find application for springs. Brass is used primarily for wiper contacts since it lacks the desirable spring properties of nickel silver and bronze. Included in satisfactory spring requirements is long service life which depends upon good fatigue characteristics and freedom, in many instances, from the tendency to season crack.

Die Casting Alloys

The demand in the Bell Telephone System for the economical production of large quantities of small complex parts has led to an extensive and growing use of die castings. If the past is a guide to the future, further expansion can be expected. Although the zinc base alloys represent the major proportion of all alloys consumed, other materials find application where specific properties are desired. High dimensional accuracy is obtained with tin base alloys; light weight is a notable property of aluminum base alloys. Lead base die castings are used principally in coin collectors where their sound and mechanical damping characteristics are important. To produce the desired properties consistently the metallurgical characteristics of these materials must be known and specific procedures followed.

Electrical Contact Alloys

Requirements of a suitable contact are many, and vary with the use to which the contact is subjected. Two requirements that are universal and paramount are that the contact material must provide an electrical path of a low resistance and must not wear away too rapidly. (Some contacts are expected to give satisfactory performance for more than 150 million operations.) In the communication systems both precious and base metal contacts are extensively used. Of the former class, platinum, palladium, silver, platinum-gold-silver, gold-silver, palladium-copper, or platinum-iridium, have given good service performances. Wiping contacts are widely employed in dial central offices. These consist generally of brass and bronze although silver is being used to an increasing extent.

Some idea of the extent to which our modern communication systems are dependent upon electrical contacts is illustrated by the number of pairs of precious metal contacts that must operate reliably to complete an ordinary dial system call between subscribers in a large city. Such a call brings into operation about three hundred relays involving over one thousand pairs of contacts. In a long distance call between New York and San Francisco about 1500 additional pairs of precious metal contacts must perform dependably for satisfactory transmission. In some years our communication systems have required more than 100 million pairs of contacts furnished on different kinds of telephone apparatus.

It may be readily appreciated, therefore, that knowledge of the factors governing contact performance is of vital importance.

Magnetic Materials

Telephone apparatus presents a great diversity of applications for magnetic materials. Both soft * and permanent magnet materials are extensively used. The soft magnetic materials are employed both as sheet and rod and in a finely divided form for compressing into cores for inductance coils. Previous to 1920 the primary soft magnetic material was iron; small quantities of silicon steel also were used. Since that date a large number of new soft magnetic materials have been developed with superior properties for particular applications. The discovery of permanent magnet characteristics in dispersion-hardening iron alloys containing no intentional carbon has resulted in a number of new permanent magnet materials of superior properties.

At this time, in the field of soft magnetic materials, iron and silicon steel find by far the most extensive application. The iron is a high grade commercial iron. The silicon steel used is the grade normally containing about 4 per cent silicon. For applications requiring higher permeabilities and lower losses, alloys of iron and nickel, known as the *permalloys*, are used. There are two principal *permalloys*, one containing about 80 per cent nickel and another 45 per cent. The higher nickel composition is also modified by molybdenum or chromium additions to increase electrical resistivity and improve magnetic properties. Sheet and rod stock are used in relays, transformers, miscellaneous coils, and ringers.

In investigating magnetic materials in the laboratory it is desirable frequently to fabricate the alloy into extremely thin sheet. The twenty roll cold-reduction mill shown in Fig. 6 is of value for this purpose. It is equipped with small diameter working rolls, each backed by a cluster of nine supporting rolls. With this arrangement high unit pressures are obtained and sheet a fraction of a mil thick can be produced readily.

In the form of 120-mesh powder and even in finer sizes certain of the *permalloys* find application in loading coils, filter coils and associated equipment. To secure low losses the powder particles are each insulated with a high resistivity, heat resistant material prior to pressing into cores. Manufacture of this fine alloy powder is a unique metallurgical process taking advantage of the effects of small amounts of added elements to achieve a desired result. The presence of a few thousandths per cent of sulphur in the iron-nickel alloys in the range of 80 per cent nickel results in a structure which can be rolled to small

* The term *soft* is used to designate materials of relatively high permeability and low magnetic loss. Likewise, permanent magnet materials are frequently referred to as "hard."

section when hot, but when cold it is exceedingly brittle and can be pulverized to fine powder. The manganese content of the alloy must also be controlled since it has an effect opposite to that of sulphur.

The iron-cobalt system yields a useful magnetic material, the one

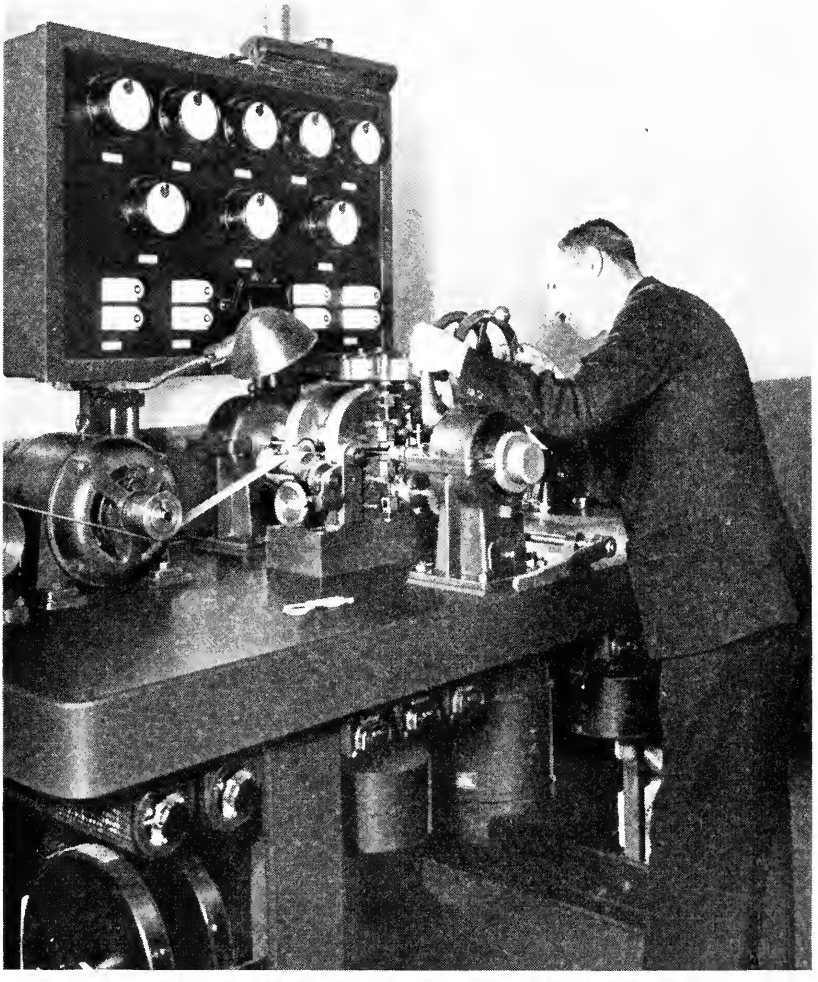


Fig. 6—Twenty roll cold-reduction mill for producing thin sheet materials for experimental studies.

containing approximately equal percentages of iron and cobalt. This alloy, called *permendur*, is characterized by high permeability at high flux densities and by a high reversible permeability when subjected to superposed direct current magnetizing forces. The binary alloy can-

not be fabricated cold and this appeared at first to limit seriously the applications for the otherwise promising material. Brittleness in cold-rolling was overcome through the addition of approximately 2.5 per cent of vanadium, whereupon the alloy can be cold rolled after a quench from a high temperature. Fortunately the vanadium does not materially impair the useful magnetic characteristics. The alloy finds its chief application in the form of .010 inch sheet in the telephone receiver diaphragm.

Substantial tonnages of permanent magnet materials are also used in telephone apparatus per year. Of this most is 3.5 per cent chromium and other permanent magnet steels of low cost and low maximum energy product ($B \times H$ maximum for the demagnetization curve). Much of the remainder used is a material with high maximum energy product for receivers and other applications where space and weight limitations prevail. For this purpose 36 per cent cobalt steel has been used but it is now replaced in new apparatus by an iron-cobalt-molybdenum alloy, *remalloy*, which has superior magnetic properties and is of lower cost.

This iron-cobalt-molybdenum alloy, which contains approximately 12 per cent cobalt and 17 per cent molybdenum, has no intentional carbon addition and is of a dispersion hardening type. The hardening heat-treatment consists of quenching from 1180°–1300° C. in oil (after which the material is mechanically and magnetically soft) followed by aging at 670°–700° C. for one hour (which induces mechanical and magnetic hardness). The material can be hot-worked and machined except in the hardened condition, and welds readily, but is somewhat brittle.

Magnets of the iron-nickel-aluminum type are increasingly used in telephone apparatus. These alloys may be ternary compositions or may be modified by a number of additional elements; cobalt and copper additions have been found advantageous. The high coercive force, high maximum energy product, and light weight make them attractive. Disadvantages are non-workability and lack of machinability.

In addition to magnetic purposes, ferrous alloys are used extensively in other applications. Considerable quantities of carbon and alloy steels are used for structural purposes, and high alloy steels for installation and maintenance tools.

Prospective Developments

In concluding a discussion of metallic materials in telephone equipment interest naturally is directed toward the future developments.

The trends in the use of new metallic materials in the telephone service are difficult to predict. A large class of applications includes the incorporation of improved materials in existing apparatus with some modification in design resulting in a cost saving or in improved service. Such materials originate from developments by the metallurgical industry and from investigations by the System's engineers. Examples of this type have already been mentioned; for example, improved cable sheathing materials, electrical conductors, and magnetic alloys. This evolution in application of materials will undoubtedly continue and constitute a large part of the telephone metallurgists' activities.

There is another field of application for metallic materials, applications in newly designed apparatus or systems of communication. Here the properties of existing materials are frequently inadequate to perform the required duties and new materials must be developed with the necessary properties. One example already cited is the preparation of magnetic powder for inductance coil cores. A new system of transmission, a million-cycle system, requires newly developed materials in the coaxial cable and the associated equipment. Special properties are usually involved which are of interest only in connection with communications, and hence the development of such materials is dependent almost wholly on the activities of the System's research groups.

An Interesting Application of Electron Diffraction *

By L. H. GERMER and K. H. STORKS

SILICOSIS develops rather quickly in rabbits exposed to air containing moderate concentrations of quartz particles finer than about 5×10^{-4} cm, but is completely prevented if aluminum powder is also present in the air to the extent of about one per cent by weight of the quartz powder. This protective action of aluminum powder was discovered at the McIntyre-Porcupine Mines, and has been studied experimentally by Denny, Robson and Irwin.[1]

It has been established that aluminum forms, in the lungs, a protective film upon the surface of silica particles which prevents them from dissolving, and thus prevents toxic effects. From the relative amounts of aluminum and silica, and diameters of silica particles, one can deduce that this protective film need never be so thick, on the average, as 2×10^{-6} cm, and is, in general, many times thinner than this.

The action of the aluminum is sufficiently striking and important to justify a fuller understanding of the nature of the film which it forms upon quartz particles and Dr. Frary, Director of the Aluminum Research Laboratories, suggested to us that the answer might be forthcoming through a study of electron diffraction patterns.

In our experiments, electron diffraction patterns were obtained from thin films of silica, about 2×10^{-6} cm thick, which had been previously treated with water containing metallic aluminum powder. A beam of high speed electrons was sent through such a treated film and the resulting diffraction pattern recorded upon a photographic plate. From studies of such patterns, and comparisons with X-ray and electron patterns of known substances, materials composing layers upon silica surfaces were identified.

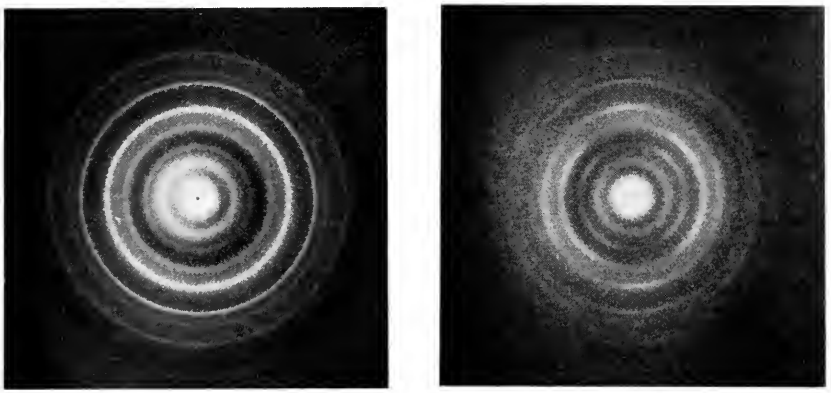
Silica films for these studies were prepared in the following manner. A glass microscope slide was first covered by gold vaporized in high vacuum from a V-shaped tungsten ribbon; then immediately in the same apparatus silica was vaporized upon the gold from a second tung-

* Digest of a paper entitled "Identification of Aluminum Hydrate Films of Importance in Silicosis Prevention," published in *Industrial and Engineering Chemistry*, Anal. Edition, 11, 583 (1939).

[1] J. J. Denny, W. D. Robson and D. A. Irwin, *Canadian Medical Association Journal, 37*, 1–11 (1937); *40*, 213–228 (1939).

sten ribbon, the distances and the quantities of gold and silica having been adjusted so that the resulting composite film consisted of a layer of silica of thickness 2×10^{-6} cm lying upon a layer of gold of thickness 30×10^{-6} cm. This composite film was large enough to supply a great many samples of silica which could be used in a large number of experiments. Each sample was prepared, as and when required, by stripping from the glass slide a small piece of the composite film, dissolving the gold in a nitric-hydrochloric acid mixture, and then washing the remaining tiny silica film in several changes of distilled water.

Films prepared in this manner were floated upon distilled water containing aluminum powder, for various lengths of time and at two differ-



A *B*

Fig. 1—Electron diffraction patterns from a relatively thick layer of oriented aluminum alpha-monohydrate crystals formed upon a silica film as a result of exposure of the film to metallic aluminum and water at 38° C. *A*—Electron beam normal to film surface. *B*—Beam inclined 45° to film surface.

ent temperatures. In some experiments the pH of the water was adjusted by the addition of HCl or various salts.

Films treated at 38° C. (approximately body temperature), and at medium and high pH values [2] (6 to 9), gave sharp electron diffraction patterns which were identified with oriented crystals of that hydrated oxide of alumina known as aluminum alpha-monohydrate (Boehmite). Typical patterns are reproduced in Fig. 1. At a low pH value (pH 4) monohydrate crystals were not discovered even after long reaction times. Although the crystal structure of aluminum alpha-monohydrate is not known it was possible to make the identification by

[2] The term pH is defined as the logarithm of the reciprocal of hydrogen ion concentration, hydrogen ion concentration being expressed for purposes of this definition in terms of grams of hydrogen ions in a liter (or more strictly 1000 grams) of solution. In a neutral solution pH = 7; in acid pH < 7 and in alkali pH > 7.

comparison of the electron patterns with X-ray and electron patterns obtained from the bulk material (Fig. 2).

Electron diffraction patterns from alpha-monohydrate formed on silica surfaces were found to vary markedly with pH of the aluminum-water solution and with the reaction time. From these patterns the following conclusions were drawn. Monohydrate crystals formed after short reaction times (4 hours to 20 hours) were sharply oriented with a particular crystal plane parallel to the silica surface; the individual crystals were on the average fairly large (from 5 to 10×10^{-7} cm) in directions parallel to the surface, and thin (2×10^{-7} cm or less) normal to the surface. As the reaction time was increased, the crystals became, on the average, thicker normal to the surface (but seldom

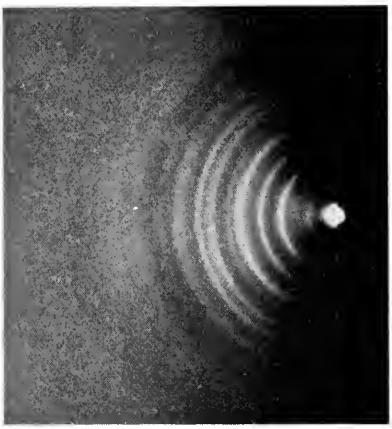


Fig. 2—Electron diffraction pattern obtained by the reflection method from finely pulverized aluminum alpha-monohydrate ($Al_2O_3 \cdot H_2O$).

as thick as 5×10^{-7} cm), and at the same time other crystals of monohydrate were formed which were less nearly perfectly oriented although still showing the same strong preference. For long reaction times layers of completely unoriented alpha-monohydrate crystals were sometimes produced.

In the presence of traces of organic acids oriented soap crystals were formed as a result of the reaction of aluminum and water. These crystals were produced at all pH values. They appeared as scum upon the water surface, and were not readily adsorbed upon silica. This fact proves that the action of aluminum in preventing development of silicosis cannot be attributed to an aluminum soap. Figure 3 exhibits a typical diffraction pattern from oriented crystals of an aluminum soap.

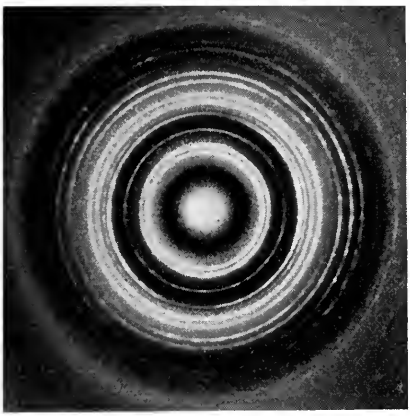


Fig. 3—Electron diffraction pattern produced by a layer of oriented crystals of an aluminum soap, which had been formed as scum upon a water surface as a result of the reaction of powdered aluminum, water and traces of organic acid present as an impurity.

Our experiments prove that aluminum hydrate is precipitated fairly rapidly upon silica at pH values lying within a range in which lie also the pH values of body fluids of men and of animals. Since in these experiments aluminum hydrate is not formed upon silica at pH 4, it seems highly probable that aluminum would not afford protection from silicosis to a hypothetical animal with body fluids of pH 4.

Abstracts of Technical Articles by Bell System Authors

Remaking Speech.[1] HOMER DUDLEY. Speech has been remade automatically from a buzzer-like tone and a hiss-like noise corresponding to the cord-tone and the breath-tone of normal speech. Control of pitch and spectrum obtained from a talker's speech are applied to make the synthetic speech copy the original speech sufficiently for good intelligibility although the currents used in such controls contain only low syllabic frequencies of the order of 10 cycles per second as contrasted with frequencies of 100 to 3000 cycles in the remade speech. The isolation of these speech-defining signals of pitch and spectrum makes it possible to reconstruct the speech to a wide variety of specifications. Striking demonstrations upon altering the pitch of the remade speech stress the contribution of the pitch to the emotional content of speech. Similarly the spectrum is shown to contribute most of the intelligibility to the speech.

Deviations of Short Radio Waves from the London–New York Great-Circle Path.[2] C. B. FELDMAN. During the past year experiments have been made to determine the frequency of occurrence and extent of deviations of short radio waves from the North Atlantic great-circle path. For this purpose the multiple-unit steerable antenna (Musa), described to the Institute at its 1937 convention, has been used to steer a receiving lobe horizontally. This is accomplished by arraying the unit antennas broadside to the general direction from which the waves are expected to arrive. The Musa combining equipment then provides a reception lobe in the horizontal plane, steerable over a limited range of azimuth. Two such Musas have been used, one of which possesses a wide steering range but is blunt, while the other is sharp but is restricted in range. Transmissions from England have been studied with this equipment at the Holmdel, N. J., radio laboratory of the Bell Telephone Laboratories. Comparisons of results obtained on transmission from antennas directed toward New York with those from antennas otherwise directed have, to a limited degree, given results representative of the effects of horizontally steerable transmitting directivity. Observations made on these British transmissions during the past eight months have disclosed the following characteristics:

[1] *Jour. Acous. Soc. Amer.*, October 1939.
[2] *Proc. I. R. E.*, October 1939.

1. During "all-daylight" path conditions, the usual multiplicity of waves distributed in or near the great-circle plane, which constitutes normal propagation, has been predominant. Usually neither ionosphere storms nor the catastrophic disturbances associated with short-period fade-outs seem to affect the mode of propagation.

2. In contrast to 1, during periods of dark or partially illuminated path conditions, the great-circle plane no longer provides the sole transmission path. The extent to which other paths are involved varies greatly. Propagation during inosphere storms of moderate intensity usually involves paths deviated to the south of the great circle, during afternoon and evening hours, New York time.

An Experimental Investigation of the Characteristics of Certain Types of Noise.[3] KARL G. JANSKY. The results of an investigation of the effect of the band width on the effective, average, and peak voltages of several different types of noise are given for band widths up to 122 kilocycles. For atmospheric noise and that due to the thermal agitation of electric charge in conductors, both of which consist of a large number of overlapping pulses, the peak, average, and effective voltages were all proportional to the square root of the band width. For very sharp, widely separated, clean, noise pulses, the average voltage was independent of the band width and the peak voltage was directly proportional to the band width. For noise of a type falling between these two the effect of the band width depended upon the extent of the overlapping.

The ratio of the peak to effective voltage of the noise due to the thermal agitation of electric charge in conductors was measured and found to be 4. The ratio of the average to effective voltage of this type of noise was found to be 0.85.

The experiments showed that when a linear rectifier, calibrated by a continuous-wave signal having a known effective voltage, is used to measure the effective voltage of this type of noise the measurements should be increased by ½ decibel to obtain the correct result.

Insulation of Telephone Wire with Paper Pulp.[4] J. S. LITTLE. The paper presented here covers the history and development of wood pulp insulation for telephone circuits. The development involved the study of wood pulps and their preparation, the methods of applying such pulp to wire, and the development of the necessary properties within the insulation to make it suitable for telephone use. The use

[3] *Proc. I. R. E.*, December 1939.
[4] *Wire and Wire Products*, October 1939.

of this insulation has made it possible to increase greatly the number of telephone circuits in a given cable by using finer wires and thinner insulations.

A General Radiation Formula.[5] S. A. SCHELKUNOFF. In this paper a general formula is derived for the power radiated in non-dissipative media by a given distribution of electric and magnetic currents. Magnetic currents are included not only for the sake of greater generality but also because in problems involving diffraction through apertures and radiation from electric horns, the radiation intensity can be made to depend upon fictitious electric- and magnetic-current sheets covering the apertures or horn openings.

Part I consists of an introductory discussion, summary of the formulas, and examples illustrating the convenience of the general formulas. Part II contains a mathematical derivation of the radiation formulas.

A Transmission System of Narrow Band-Width for Animated Line Images.[6] A. M. SKELLETT. A new method of transmission and reproduction of line images, e.g., drawings, is described which utilizes a cathode-ray tube for reproduction, the spot of which is made to trace out the lines of the image 20 or more times a second. The steps of the complete process are: first, the transcription of the line image into two tracks similar to sound-tracks on moving picture film; second, the production from these tracks of two varying potentials by means of photoelectric pick-up devices; third, the transmission of these potentials; and fourth, their application to the cathode-ray deflector plates to effect reproduction. Satisfactory transmission of fairly complex images, e.g., animated cartoons, could be effected within a total band-width of 10,000 cycles.

[5] *Proc. I. R. E.*, October 1939.
[6] *Jour. S. M. P. E.*, December 1939.

Contributors to this Issue

R. M. Bozorth, A.B., Reed College, 1917; U. S. Army, 1917–19; Ph.D. in Physical Chemistry, California Institute of Technology, 1922; Research Fellow in the Institute, 1922–23. Bell Telephone Laboratories, 1923–. As Research Physicist, Dr. Bozorth is engaged in research work in magnetics.

W. P. Mason, B.S. in Electrical Engineering, University of Kansas, 1921; M.A., Columbia University, 1924; Ph.D., 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged in investigations on carrier systems and in work on wave transmission networks both electrical and mechanical. He is now head of the department investigating piezoelectric crystals.

H. Nyquist, B.S. in Electrical Engineering, North Dakota, 1914; M.S., North Dakota, 1915; Ph.D., Yale, 1917. Engineering Department, American Telephone and Telegraph Company, 1917–19; Department of Development and Research, 1919–34; Bell Telephone Laboratories, 1934–. Dr. Nyquist has been engaged in transmission work, particularly telegraph transmission. He is at present Engineer of Transmission Theory.

K. W. Pfleger, A.B., Cornell University, 1921; E.E., 1923. American Telephone and Telegraph Company, Department of Development and Research, 1923–34; Bell Telephone Laboratories, 1934–. Mr. Pfleger has been engaged in transmission development work, chiefly on problems pertaining to delay equalization, delay measuring, temperature effects in loaded-cable circuits, and telegraph theory.

D. K. Gannett, B.S. in Engineering, University of Minnesota, 1916; E.E., University of Minnesota, 1917. American Telephone and Telegraph Company, Engineering Department, 1917–19; Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. As Toll Transmission Engineer, Mr. Gannett is concerned principally with the transmission features of toll systems, particularly program systems, toll signaling systems, and vacuum tube applications in these and other systems.

Earle E. Schumacher, B.S., University of Michigan; Research Assistant in Chemistry, 1916–18. Engineering Department, Western

Electric Company, 1918–25; Bell Telephone Laboratories, 1925–. As Associate Research Metallurgist, Mr. Schumacher is in charge of a group whose work relates largely to research studies on metals and alloys.

W. C. Ellis, Ch.E., Rensselaer Polytechnic Institute, 1924; Ph.D., 1927. Bell Telephone Laboratories, 1927–. Dr. Ellis has been engaged in metallurgical studies on magnetic materials and copper alloys.

A. M. Curtis, 1907–13: Radio Operator; Supervisor of Radio System of Brazilian Lloyd; Exploration Work for Brazilian Government. Western Electric Company, 1913–17. Captain, Signal Corps, A.E.F., 1917–19. Western Electric Company and Bell Telephone Laboratories, 1919–. Mr. Curtis took part in the pioneer transatlantic radio telephone tests of 1915 and was associated with the development of permalloy loaded submarine cables and terminal apparatus for their operation. As Circuit Research Engineer he is now in charge of researches dealing with radio telephone terminals and similar "voice-operated" apparatus and part of the research work on contact operation.

L. H. Germer, A.B., Cornell, 1917; M.A., Columbia, 1922; Ph.D., Columbia, 1927. Engineering Department, Western Electric Company, 1917–25; United States Army, 1917–19; Bell Telephone Laboratories, 1925–. Dr. Germer has been engaged upon work in thermionics and electron scattering, and in more recent years upon applications of electron diffraction to investigation of surface films and surface chemistry.

K. H. Storks, B.S. in Chemistry, Coe College, 1930. Bell Telephone Laboratories, 1930–. Mr. Storks has been engaged in studies of applications of electron diffraction to chemical problems.

VOLUME XIX APRIL, 1940 NUMBER 2

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XIX *April, 1940* *No. 2*

Advances in Carrier Telegraph Transmission

By A. L. MATTE

Introduction

IN the comparatively short period which has elapsed since its commercial introduction in a practical form, the voice-frequency carrier method of operating telegraph has risen to a position of preeminence and is becoming the outstanding means for providing telegraph facilities over main toll routes.

Since the original installation, improvements have been made in the carrier supply, level-compensating devices, maintenance facilities, and in numerous other specific physical parts of the system. Operating speeds have also gone up, and the number of telegraph channels per telephone circuit has been increased. Furthermore, this system, originally designed for cable circuits operating at voice frequencies, has been applied to open-wire lines and adapted by remodulation to other frequency ranges, in particular to those occupied by existing carrier-telephone systems. Some of the chief advances, however, have been of a more intangible nature, not the least of these being the clearer insight which experience and extended tests have given into the possibilities and limitations of carrier-telegraph systems with respect to interference and other causes of signal distortion.

As a result of the success attained by the carrier-telegraph system for open-wire lines,[1, 2] which had been in commercial service in the Bell System since 1918, this company's engineers turned their attention to the adaptation to cable circuits of the carrier method of transmission for telegraph purposes. Following this work and extensive field trials, the voice-frequency carrier-telegraph system went into commercial use in the Bell System in 1923.[3] The initial installation consisted of ten two-way or duplex channels between New York and Pittsburgh, giving an aggregate channel mileage of 3800 miles (6120 km.) including both directions of transmission. Since then, the application of this

mode of transmission has spread rapidly so that in spite of the intervening period of retarded business activity it now provides about $1\frac{1}{2}$ million miles (2.4×10^6 km.) of high-grade circuits throughout the Bell System. Its use with variations has extended to other countries so that it bids fair to become an outstanding means for providing overland telegraph facilities, particularly for long distances, where the service is exacting. As indicating the general trend, it may be stated that in England alone about 1700 voice-frequency telegraph channels were reported as available for operation at the end of 1938.[4]

It is interesting to note the role played by carrier telegraph in the evolution of the art of telegraphy. The three major telegraph systems up to about 1890 were those of Hughes on the Continent, Wheatstone in England, and the manual Morse system in the United States. As electrical communication reached out to greater and greater distances, the desire to utilize costly lines more effectively led inventors to concentrate their efforts in two different directions; namely, the development of high-speed systems and of multiplex systems. In high-speed systems, the object, as the name implies, is to secure increased line output by speeding transmission well beyond the ability of a single operator. These devices are characterized by automatic transmitters which can be supplied with perforated tape prepared in advance by a number of individuals. Typical high-speed systems are the Wheatstone automatic, the Murray automatic, the Siemens and Halske high-speed, and the Creed high-speed.[5]

The first efforts at multiplexing circuits were based upon the suggestions of Gintl and Highton, who proposed to take advantage of directional and magnitude effects respectively, and whose ideas were brought together by Edison in his invention of the quadruplex. The multiplex system as we know it, however, was the invention of Baudot, who, putting into practical form a suggestion made by Moses G. Farmer as far back as 1853, produced a system whereby the line was assigned successively to a number of operators. This process had the great advantage that while maintaining the line speed which economy made imperative, it permitted a number of messages to be transmitted simultaneously without delay and with each operator working at his normal pace. The chief examples of the multiplex are the Baudot, Murray, and American.[5, 6]

Owing to the advantages of these higher output systems, the older methods of operation were gradually supplanted for the longer commercial message circuits. This was particularly true in Europe, since certain conditions operating in America tended to favor the survival of the simple Morse arrangement; the chief of these being the avail-

ability of large numbers of composited and simplexed circuits, most of which were used in private line service,[7] and the low traffic density on many long multi-section circuits, making it desirable to provide intermediate operating points. By about 1920, the weight of evidence definitely favored the multiplex method of exploitation over the use of the high-speed printer.[6] It was at about this point in telegraph history that the carrier telegraph method of subdividing the line capacity made its appearance and, through its superior flexibility and lesser intricacy of operation, began gradually to supersede the distributor methods of multiplexing circuits for many types of services.

While voice-frequency telegraphy was foreshadowed by Elisha Gray's harmonic telegraph,[8] which was exhibited at the Third French International Exposition in 1878 and at the Electrical Exhibition in Paris in 1881,[9] its practical embodiment had to await the invention of the electrical filter by Campbell, that of the audion by DeForest, and the production of effective means for generating alternating currents of acoustic frequencies.

The success of this system rests mainly upon its adaptability to economical operation over telephone circuits by making effective use of the whole frequency band usually allocated to the voice, and in requiring similar transmission characteristics. Henceforth, every advance in telephony directed to an improvement of the transmitting medium contributes as well towards the improvement of telegraphy; the economies of wide-band carrier telephony, the improved equalization and regulation of circuits, the reduction of interference and the elimination of crosstalk, all tend to make the telegraph a more dependable and efficient tool for modern industry and modern living. Thus telegraphy, one of the oldest of the electrical arts, having fathered the telephone, now finds, within the great technical structure which the latter has created, a fertile medium for the development of its usefulness, not as a competitive but as a complementary service. Thanks to voice-frequency telegraphy, wherever the telephone reaches, a high-speed, reliable, record-form of telegraph may follow. This has brought about a great simplification of the problem of interconnection in such large communication networks as the international postal area in Europe and the Bell System in our own country.

Furthermore, carrier telegraphy has doubtless been a means of advancing the fortunes of the start-stop teletypewriter,[10] by subdividing the frequency band to such an extent that one channel may be economically assigned to a single operator working at normal speed. It has also been a factor in simplifying the switching problems presented by the extensive introduction of teletypewriter exchange (TWX) service.[11]

The purpose of this paper is to describe the principal transmission developments which have taken place in the voice-frequency system since the first commercial installation, to present some of its operating characteristics, and to outline advances in maintenance methods which have developed during this period.

There has been marked and steady improvement in the quality of the service rendered by the voice-frequency telegraph circuits during the last decade within the Bell organization, and while it would be unfair to overlook the part which imaginative management, employee cooperation, and similar factors have played in securing this desirable result, it appears certain that a good deal of it is to be credited to those physical improvements and advances in testing and operating procedures which we are about to recite.

Extensions in Utilized Frequency Range

While the greater part of past experience has been had with the application of voice-frequency systems to extra-light-loaded four-wire cable circuits,* a considerable mileage now utilizes other types of telephone facilities, particularly high-frequency carrier open-wire lines [13] through the less densely populated regions. These latter applications are interesting to the transmission engineer because the association of telephone and telegraph in the same repeaters brings into view new problems which will doubtless grow in importance as the use of broadband carrier systems becomes more extensive. A typical voice-frequency carrier-telegraph circuit is shown in Fig. 1. It consists of a section of four-wire cable connected in tandem with a three-channel type "C" carrier-telephone circuit,[14, 15] without mechanical repetition at the junction point. The telegraph power is appropriately modified to suit the requirements of the two media by means of pads and amplifiers *P* at the point where they join. In addition to this, the remaining telephone circuits operating through the same carrier repeaters are equipped with volume limiters [16] to prevent voice-energy peaks, which contribute little if any to telephone quality, from overloading the amplifiers, thereby causing excessive distortion to the telegraph.

Shortly after the initial voice-frequency telegraph installation in cables the number of channels was extended from 10 to 12 by the addition of two channels at the upper end of the frequency range, corresponding to carrier frequencies of 2125 and 2295 cycles.

* These circuits are designated as H44. They consist of 19 AWG conductors, loaded with 44-millihenry coils 6000 ft. (1830 m.) apart, with four-wire repeaters spaced at approximately 50-mile (80.5 km.) intervals.[12]

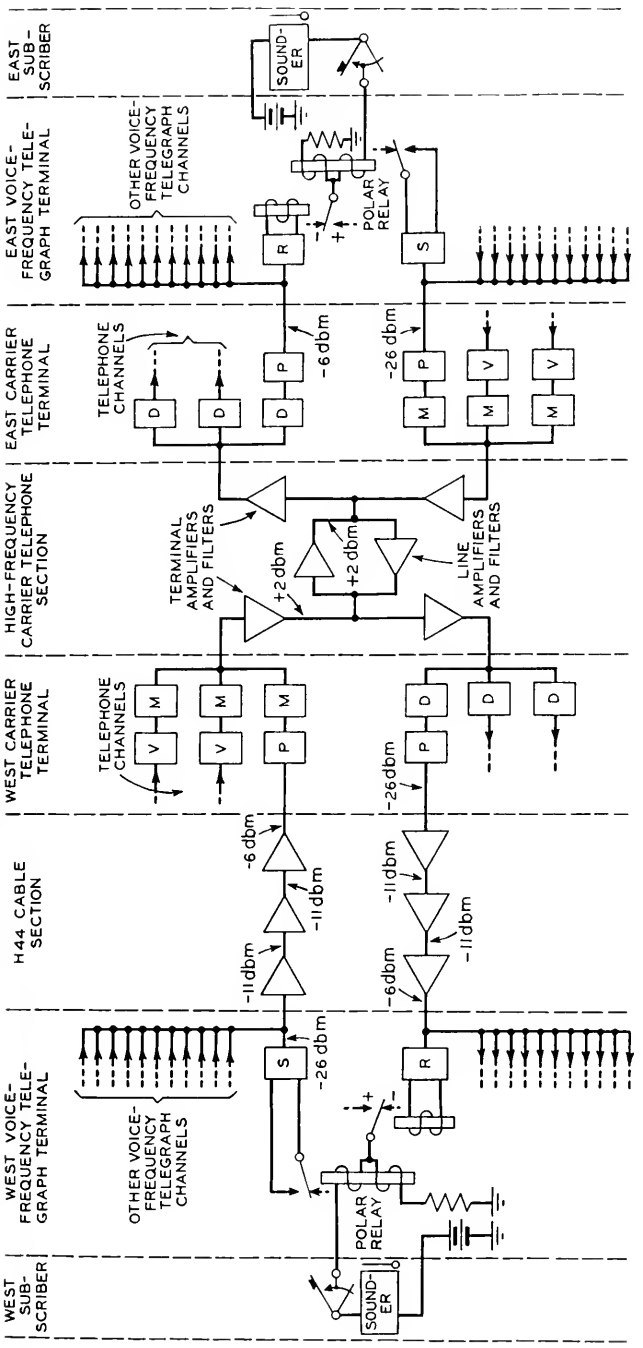


Fig. 1—Typical carrier-telegraph circuit.

At a later date the extensive introduction of H44 circuits, with their relatively high cut-off, was thought to make it desirable to consider a further extension of the frequency band utilized by the telegraph. However, in view of the fact that the then existing state of the art of filter design made it impractical to produce economical filters of the required narrow band-width but having the desired high mid-band frequency, it was decided to develop a carrier-telegraph system suitable for use between dense traffic centers by superposing two standard 12-channel systems over the same cable pair. This was accomplished by causing the various signaling frequencies of one system to modulate a single secondary carrier, thereby transposing all the frequencies of this voice-frequency system to a frequency range above that of a normal voice-frequency system operating over the same cable pair.

The line circuit used with this double system was required to transmit a range of frequencies from about 350 cycles to 4400 cycles, and the stability within this range had to be such as not to cause excessive bias in any channel with the regulating methods available at that time. In order to secure this result, it was necessary to change the transmission characteristic of all repeaters from that used for ordinary four-wire telephone or voice-frequency telegraph transmission and, furthermore, to modify somewhat the regulating repeaters in order to maintain the desired transmission characteristics with changes in temperature.

No changes of any kind were required in the voice-frequency telegraph terminals. The channel frequencies on the line were arranged to extend uninterruptedly at 170-cycle intervals from 425 cycles to 4335 cycles.

This arrangement was called the "double-modulation" system, because operation of two voice-frequency carrier telegraph systems over the same circuit was realized by causing all the channel frequencies of one of these two systems to pass through a common modulator, where a second modulation took place; the individual frequencies of each channel being already considered as having been modulated by the sending relays. A single secondary carrier-frequency was used which was common to all the channels. The allocation of frequencies, which was identical for both directions of transmission, will be readily understood by referring to Fig. 2*B*. The general principle of operation is illustrated in Fig. 2*A*, which shows transmission from west to east, it being understood that the arrangement from east to west is identical. The voice-frequency system denoted as No. 2 will be seen not to differ in any way from the earlier arrangement except that signals from all twelve channels traverse low-pass grouping filters at the sending and receiving terminals. In the case of voice-frequency system No. 1,

however, all the frequencies are passed through a modulator, where they are transposed as a group to a position above those of system No. 2. The lower sideband of the secondary carrier is used, so that the order of the channels is reversed on the line. After modulation, this

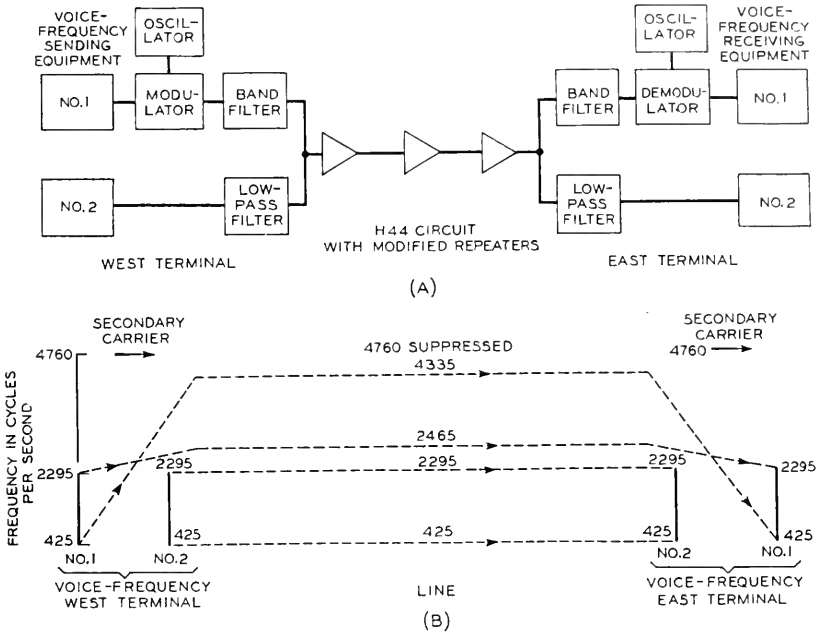


Fig. 2—Double-modulation telegraph system. *A*. Block diagram for one direction of transmission. *B*. Frequency relations at terminals and on the line.

group of frequencies passes through a sending band filter, which elminates all the unwanted frequencies, thereby preventing useless overloading of the repeaters and the creation of undesired modulation products therein. The two groups of frequencies pass through common repeaters over the modified H44 circuit and are then separated at the receiving terminal by a combination of filters similar to the one at the transmitting end of the circuit. The signals pertaining to voice-frequency system No. 1 are next demodulated by a secondary carrier having the same frequency as that used at the sending end and thereby reduced to a frequency range adaptable to the standard terminal equipment. The modulators and demodulators were provided with separate oscillators at both ends of the circuit.

Both modulators and demodulators were of the push-pull type and were arranged as grid-current modulators [17] instead of as plate-current

modulators such as were then current in telephone practice. In the case of grid-current modulators, the necessary non-linear characteristic is obtained by so constructing the input circuit that the voltage between the grid and filament of the modulating tubes does not vary directly with the voltage impressed upon the modulator input, while plate modulation is suppressed; in the case of plate-current modulators, on the other hand, there is a linear relation between the grid-to-filament voltage and the voltage impressed upon the modulator input, but advantage is taken of the fact that the plate current does not vary directly with the grid voltage to secure the desired modulation effect. The reason for using grid-current modulators of this type was that the increased power output secured thereby made it possible to produce the required output levels without auxiliary amplifiers.

The trial double-modulation system performed satisfactorily under commercial conditions although the upper group or remodulated system was somewhat less satisfactory than the standard. This wide range system has not been used, however, partly because of reduced demand due to economic conditions and partly because advances in the art of filter design now make possible a considerable extension on a single modulation basis; furthermore, the increasingly wide use of carrier telephone circuits makes it desirable to restrict the band width used by a voice-frequency telegraph system to the frequency range normally assigned to a telephone channel.

Signal Distortion

Improvements in transmission amount essentially to reductions in signal distortion. The principal sources of such distortion [18, 19] are:

Type of Distortion	Source
Characteristic	Filter characteristics.
	Wave shaping characteristic of detectors.
Bias	Variations in circuit net-loss.
	Battery variations at terminals.
	Variations in carrier-current generator voltage.
	Gradual frequency changes.
	High-resistance sending-relay contacts.
	Asymmetrical relay adjustments.
Fortuitous	Noise.
	Lightning.
	Functional switching operations.
	Change in repeater gain with load.
	Intermodulation of several channels in repeaters.
	Infiltration from adjacent telegraph channels.
	Relay contact troubles.
	Irregularities in relay operation.
	Rapid variations in carrier frequencies.

Characteristic distortion attributable to the telegraph channel filters is not an important limitation at the speeds now generally used and need not detain us as it has been discussed at length elsewhere.[20, 21, 22] It might be stated, however, that while the frequency band used at the present time, which provides an effective width of about 110 cycles, allows some margin of transmission with present speeds, the proposition of reducing the spacing of carriers is not very attractive for a number of reasons, among which may be mentioned the reduction in cost per cycle of band width due to the development of carrier-telephone systems, the possibility that higher speed requirements may ultimately make the present spacing desirable, and the greater degree of maintenance demanded by a system designed with less liberal operating margins where a number of sections are operated in tandem. With respect to the speed factor, it may be observed that service is already being rendered commercially in several cases at 75 words per minute, and still higher speeds have been used.

Bias

Variations in circuit loss consequent upon changes in temperature, battery voltages, etc., are a major factor in determining signal distortion, as will be seen by reference to Fig. 3 which shows graphically the

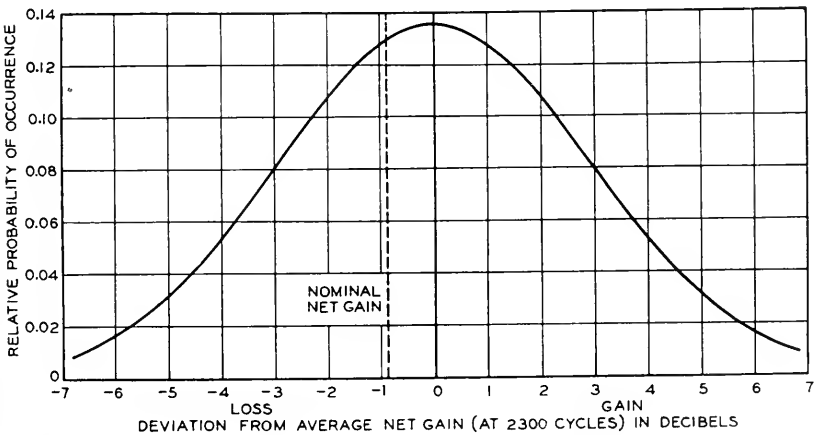


Fig. 3—Probable distribution of net-gain variations at 2300 cycles. H44 circuits about 1000 miles (1600 km.) long. Includes both differences between circuits and variations with time.

approximate manner in which a particular group of 19 gauge H44 circuits about 1000 miles (1600 km.) in length varied through a one-year cycle at the frequency of channel 12 (2295 cycles). The effective

range of variations with respect to the optimum input level to the detector depends on the method by which the sensitivity of the latter is adjusted. Two methods have particular advantages: In one of these the detectors are adjusted for a nominal received level which is made the same for all the channels. This is the method usually employed for cable circuits. It permits adjusting the detectors at any time without reference to the particular line with which they are to be used and without the assistance of an attendant at the distant station. It will be evident that the departures from optimum line gain must then be reckoned from this nominal net circuit-gain, which in the illustration is shown as being .9 db below the actual mean value. This discrepancy is principally due to imperfect equalization of the line. In general it is least in the neighborhood of 1000 cycles and increases progressively as one goes away from this frequency. Furthermore, the standard deviation of the distribution curve increases generally in the same manner, so that for a 12-channel system the condition illustrated is perhaps the most unfavorable one.

A second method of lining up is to adjust the detector sensitivity so as to give unbiased operation with signals transmitted from the distant station and with the line loss whatever it happens to be at the moment. On the average, and in the long run, the effect of this procedure is to restore the symmetry of the variations, but the standard deviation is multiplied by a factor equal to the square root of 2. This is because the occurrence of a given departure from the optimum level is then further conditioned by the particular net gain which happens to exist at the time the detector is adjusted, and the chance of a gain of this particular value is given, of course, by the same distribution which has just been discussed. Most of this increased latitude in variations can be eliminated, however, by seasonal adjustments; a procedure which is evidently of no help when the first method is followed. This second method has been found useful on open-wire circuits because the average net loss of telephone channels over these facilities depends somewhat on their frequency allocation and varies more widely than is the case with cable circuits.

If no provision were made to compensate for these line-variation effects the result would be a rapid change in bias as the level at the input of the detector departs from its optimum value. This is shown for a typical telegraph channel by the dotted line in Fig. 4. By the use of a *level compensator* associated with each individual detector a great improvement may be obtained, however, the bias variations being reduced to those illustrated typically by the full line in the same drawing. The resulting changes in teletypewriter orientation

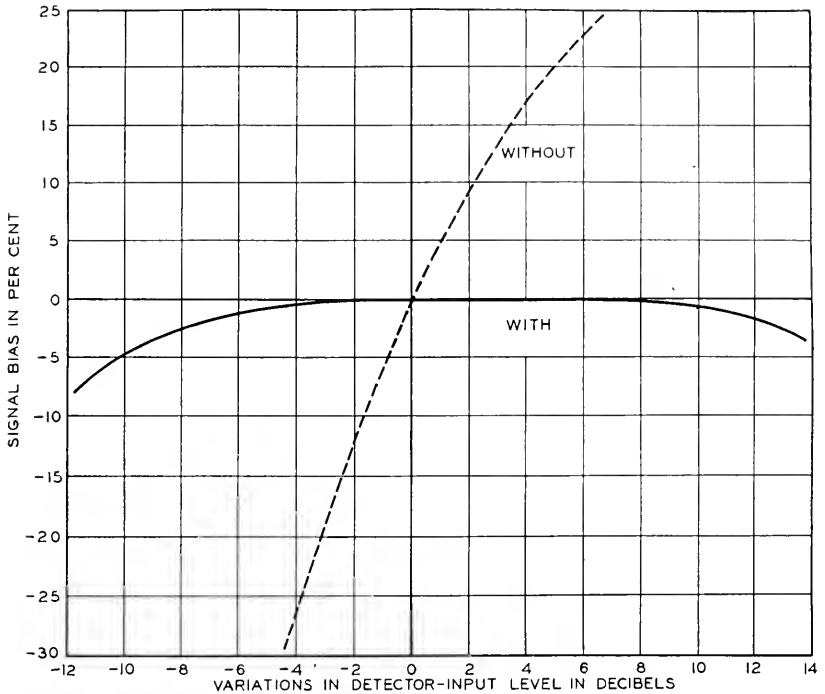


Fig. 4—Effect of level compensator. Signal bias vs. variations in detector input level. Signaling speed 23 dots per second.

range when the level compensator is used are of the order shown in Fig. 5, indicating satisfactory operation over extensive changes in circuit loss. It will be seen that in the absence of a level compensator a single-section telegraph circuit operating at a speed of 23 d.p.s. (46

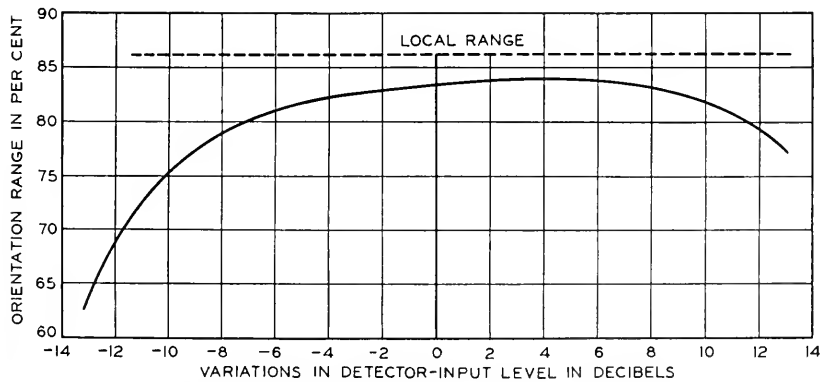


Fig. 5—Teletypewriter orientation-range vs. variations in detector input-level.

bauds) may exhibit a change in signal bias of 4 per cent or more for each db change in input level. With a level compensator of the type described herein, the bias may be kept within a range of ± 2 per cent for a variation in input level of ± 8 db.

The elimination of any considerable bias variations in individual telegraph sections due to level changes which usually occur in practice is particularly important in the case of multi-section * circuits. As a result of this improvement it has been found feasible under test conditions to operate satisfactorily as many as 10 telegraph sections in tandem at 60 words per minute for long periods without objectionable bias variations due to level changes and without the use of regenerative telegraph repeaters. In practice, however, other considerations usually make it desirable to use a regenerative repeater when the number of sections in tandem exceeds 4.

The greater part of the effective changes in received level in a given circuit is due to temperature changes which are imperfectly compensated for by the regulators, aggravated by the fact that the conditions prevailing when the circuits are adjusted within the limits specified by the maintenance routines may depart considerably from the average. In addition to this, there are variations of considerable magnitude between individual circuits due to structural differences. In view of the fact that the variations over the whole frequency range are not the same, there is a material advantage in a compensator which adjusts the gain of each detector independently, a feature which could not be secured with a pilot-channel regulator.

Level Compensator

The level compensator,[23] shown diagrammatically in heavy lines in Fig. 6, may be considered as functionally divided into two parts, one of which is in series with the grid of the detector tube and the other of which is connected to the armature of the receiving relay. The first of these will be referred to as the *grid-bias circuit* and the second as the *compensator-relay circuit.*

The grid-bias circuit consists essentially of a condenser C shunted by a high resistance R_c, in series with a biasing battery E_0 of fixed voltage, the secondary of the interstage transformer T, and the grid-filament terminals of the detector tube V. This arrangement functions to keep the effective grid-filament voltage due to the signals nearly constant, irrespective of their magnitude, by setting up a voltage on the condenser which adds algebraically to the grid-bias battery and whose magnitude

* By a multi-section telegraph circuit is meant a connection made up of 2 or more telegraph lines in tandem with mechanical repetition between them.

is automatically adjusted to be proportional to that of the incoming signals.

At any instant the actual voltage between filament and grid is therefore equal to the algebraic sum of (1) the fixed bias voltage E_0;

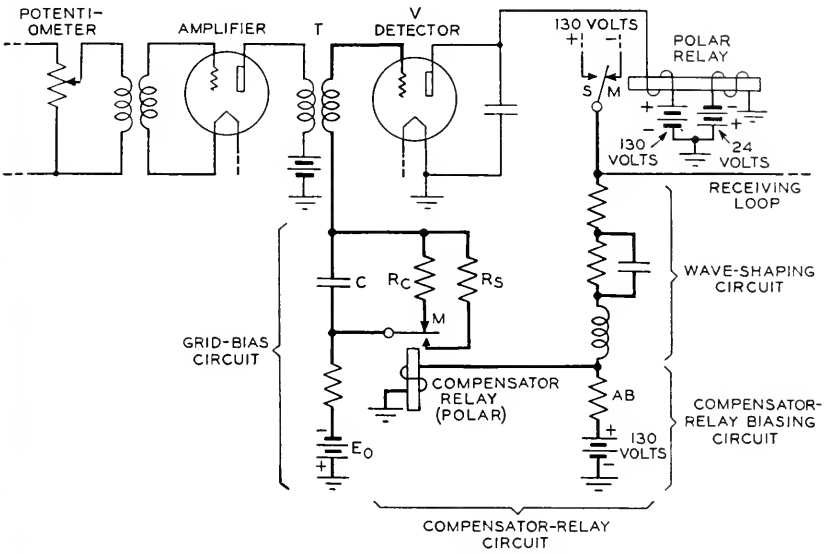


Fig. 6—Schematic diagram of level compensator.

(2) the voltage due to the charge on the condenser C; (3) the signal voltage across the secondary of the interstage transformer T; and (4) the drop in voltage across the transformer. By making the detector sensitivity sufficiently great, the signal voltage is caused to overcome the opposing negative bias during a portion of each positive half of the carrier cycles composing a marking pulse, so that a net positive voltage is periodically impressed on the grid causing a flow of current between it and the filament and consequently through the resistance R_e and the condenser C in parallel.

The resulting voltage across condenser C is in the same direction as that of the fixed grid battery E_0 and adds thereto. The condenser voltage is determined by the amplitude of the received carrier current, increasing with increased input level and decreasing with decreased input level. By a proper selection of the constants of the circuit, the desired compensation action may be obtained. This action will be such that the change in voltage across the condenser will always, within the effective range of compensation, produce the proper negative grid voltage for unbiased reception of telegraph signals by the

receiving relay. A quantitative discussion of the operation of the compensator is given in the Appendix.

In order that the voltage across the condenser may not decrease during spacing signals, an auxiliary polar relay, called the compensator relay, is provided which derives its operating current from the armature of the receiving relay and serves to disconnect the resistance R_c during spacing signals. As discussed in the appendix, the unbiased operation of the compensator relay would cause a noticeable decrease in the condenser voltage during the rapid transmission of signals, because the wave shape of the signals impressed on the grid circuit of the detector tube is not square but considerably rounded. In other words, there is a portion of a marking signal during which the receiving relay is operated to marking but the grid is non-conducting. Hence more charge would leak from the condenser during the time the resistance is connected across the condenser than would be replaced by rectification. To prevent this, the compensator relay bridges the discharge resistance R_c around condenser C for a period of time which is shorter than that during which the receiving relay is on its marking contact. The amount by which the compensator condenser must be biased depends on the signaling speed and other factors. It is determined by observing the "drift" in bias suffered by reversals when these are suddenly switched on after a long marking interval.

The operation of the level compensator will be more readily understood by referring to Fig. 7, which shows diagrammatically the manner in which received impulses of different magnitudes are made to operate the receiving relay for equal time intervals. In this diagram, the positive halves of the envelopes of three received marking impulses * of different amplitudes are shown in relation to the grid-voltage plate-current characteristic of the detector tube. For normal input level the sensitivity of the detector is made sufficiently great so that the amplitude of the impulse which is impressed on the interstage transformer is of the magnitude shown at N. The envelope of the received carrier current is symmetrical about the line E_N, which is located at the net value of grid biasing voltage due to the battery voltage E_0 and the grid condenser voltage e_c. The latter voltage, as previously noted, is produced in the grid circuit by rectification of that part of the received carrier current which lies on the positive side of the zero grid-voltage axis OO. By properly adjusting the bias of the receiving relay the latter may be made to operate at a value AA which is one-half the crest value of the envelope. Signals having amplitude N will thus be repeated unbiased by the receiving relay, since the ascending and de-

* In practice such pulses would usually reach the steady state.

scending curves bounding the envelope represent identical time sequences of events.

If the level increases so that the amplitude of the received signals has a value indicated by curve *H*, a much larger part of the signal current momentarily flows through the grid-filament space and is rectified, thereby increasing the charge on the condenser in such a

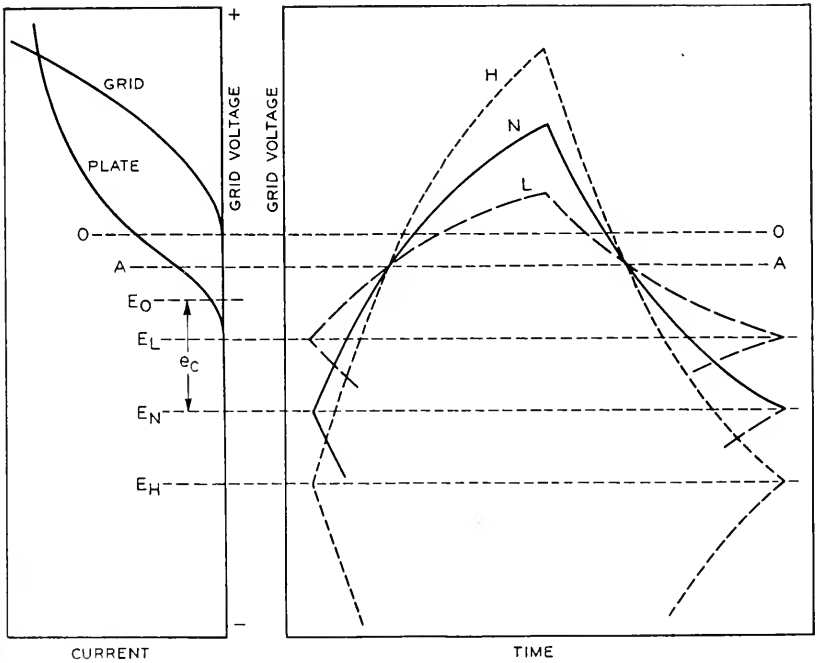


Fig. 7—Principle of level-compensator action.

way as to increase the negative bias of the grid and shift the signal bodily towards the negative side of *OO*. This shift continues until the total effective grid voltage has reached a point at which the rectified current is just sufficient to compensate for the increased discharge through resistance R_c due to the higher voltage across the condenser. The system is then in equilibrium at the new input level. If the constants of the circuit have been properly chosen, the new grid voltage E_H will be such that the middle of the positive envelope of the received signal again passes through the line *AA*, thereby giving unbiased signal reception.

The corresponding condition for an input level below normal is shown by curve *L*.

The condenser is the essential element of the compensator, and it is, of course, able to accumulate a charge in the absence of the resistance R_c; the only function of the latter being to dissipate this charge quickly when the input level drops. If there were only a resistance and no condenser there would be a simple rounding off of that part of the positive crests of the carrier waves which cause grid current to flow but no transfer of the operating axes as a whole to new bias values, such as E_L, E_N and E_H in Fig. 7.

Before leaving Fig. 7, it may be interesting to note parenthetically that a signal with large bias, such as H is much more immune to distortion due to changes in its own magnitude, or variations in relay bias, than a signal such as L. This can readily be seen if we imagine the line AA, which corresponds to the operating point of the receiving relay, to be moved up or down and note the relative lateral displacement of the intersections on these two envelopes. Another advantage of the stronger signal is that the rate of change of energy at the moment when the relay operates is much greater, insuring a more positive operation of its armature. The concave character of the detector characteristic is also favorable to securing a desirably shaped pulse for relay operation. Advantage was taken of these wave shaping possibilities in the design of detectors antedating the use of the level compensator to minimize the effect of circuit and battery variations.

In the absence of resistance R_s, Fig. 6, there would be a tendency during a long spacing interval for leakage in the wiring connected to the grid, to reduce the grid bias to ground potential. Before such discharge had gone very far, however, the receiving relay would close and recharge the condenser. This would give rise to periodic operation or *pulsing* of the relays. The purpose of R_s is to prevent this undesirable effect by making the negative bias voltage approximately equal to E_0 during long spaces. Since this resistance is large compared to R_c, it has a negligible effect during the reception of signals.

Where a circuit is exposed to transient additions of energy from external sources such as lightning, the operation of the level compensator may be stabilized by bridging a large capacitance in series with a resistance around condenser C.

Sending Circuit

In the voice-frequency telegraph system the spacing signals are produced at the transmitting end by short circuiting that portion of the circuit which supplies the sending filter with power, and the marking signals by allowing the current from the generator to flow through freely. This operation is performed by the sending relay.

Experience has shown that owing to the small a-c. voltages involved, there is a tendency for the contacts of the sending relay to increase in resistance, sometimes reaching a value as high as 1,000 ohms or more. In view of the low impedance of the circuit originally used, this caused considerable residual current to flow during spacing intervals. In order to remedy this condition, the sending relay circuit was modified to the form shown in Fig. 8, in which an auto-transformer is so con-

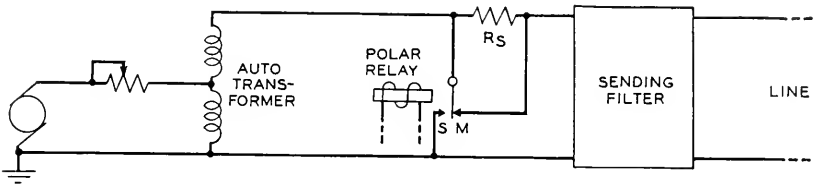


Fig. 8—Relay sending-circuit.

nected as to give a high impedance looking towards the generator, while R_s, which is of the order of 50,000 ohms, provides a correspondingly high resistance towards the output. The sending filter input is suitably padded to insure a satisfactory termination. It will be evident that with this arrangement the contact resistances in both the spacing and marking positions may vary considerably without seriously affecting the transmitting efficiency. Another advantage of R_s is that it eliminates the bias due to the transit time of the sending-relay armature, which may therefore be increased, and need not be kept within such precise limits: a matter of considerable convenience where demountable relays are used.

A trial has also been made of various schemes using varistors (copper-oxide rectifier-elements) to control the flow of carrier current by means of the changes in voltage in the loop circuit, thus dispensing with sending relays of the electromagnetic type. Figure 9*A* shows an arrangement which has been in actual operation for a number of years at several central offices and has given satisfaction. The loop circuit is provided with two equal apex resistances RR; hence when the key is closed the point x is positive relative to y regardless of the position of the receiving relay. This follows from the fact that the current through the loop is twice that through the loop-balancing resistance. On the other hand, if the receiving relay is on its marking contact, opening the loop key reverses the relative voltage between points x and y so that x becomes negative relative to y. In other words, polar signals are impressed between points x and y as a result of the transmission of signals in the loop. The a-c. part of the circuit contains a

bridge-like arrangement between the carrier source and the sending filter, which is balanced at all times with respect to the d-c. pulses so that these do not tend to be propagated into the line or towards the generator. When x is positive relative to y, rectifier elements a_1 and a_2

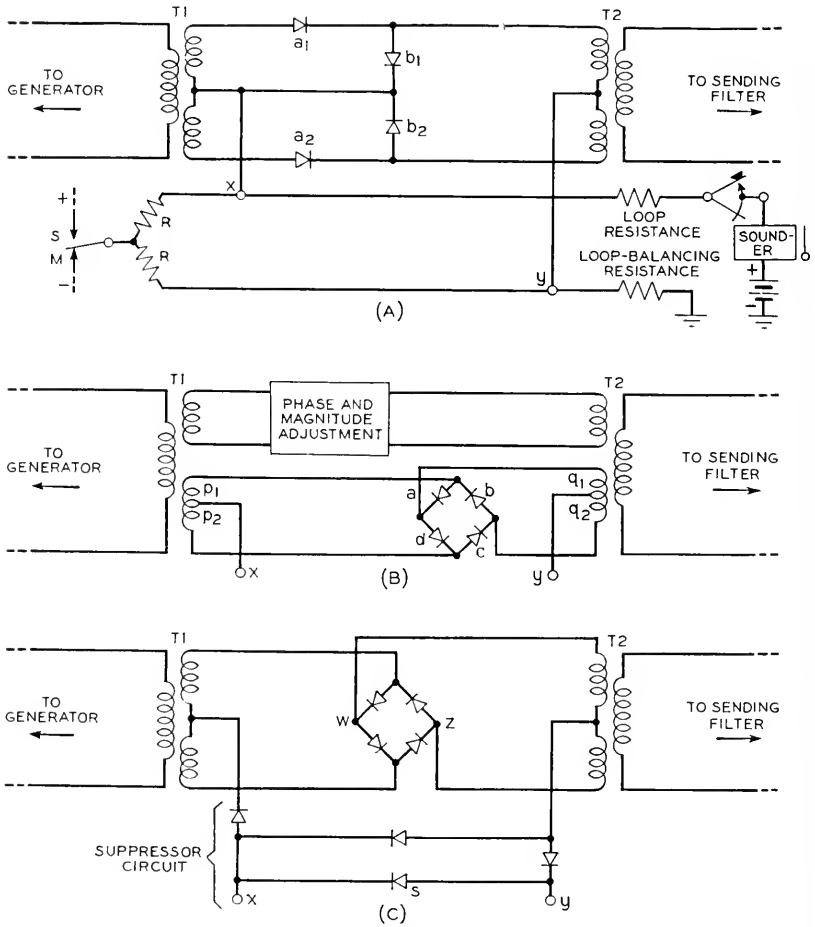


Fig. 9—Varistor sending-circuits. *A*. Series-parallel arrangement. *B*. Phase inverter. *C*. Non-polar arrangement.

are conducting while b_1 and b_2 are non-conducting. This allows a free path for the carrier between transformers T_1 and T_2 and thus from the generator to the sending filter. If, however, point x is negative relative to y, a_1 and a_2 acquire a high resistance, thereby greatly impeding the passage of current between T_1 and T_2, while b_1 and b_2 become conducting and effectively shunt the primary of T_2.

Alternative arrangements have also been tried. Two of these, which were used in actual installations, are shown in Figs. 9*B* and 9*C*. In both cases, the loop arrangement is the same as in Fig. 9*A*. The circuit of Fig. 9*B* consists of two parallel paths between generator and sending filter, one of which impresses a steady a-c. voltage on transformer T_2 while the second path serves to impress a second a-c. voltage of the same magnitude at the same point, but this latter voltage may be either in phase aiding or in phase opposing to the first, depending on the polarity of the d-c. voltage impressed through the varistor bridge. Thus it will readily be seen that if x is positive relative to y, elements a and c are conducting, while b and d are not. Transmission of the carrier then takes place around the path p_1, q_1, q_2, p_2, and the voltages from the two parallel circuits are additive in T_2. If, however, x is negative relative to y, the conducting condition of the varistor elements is reversed so that the carrier path becomes p_1, q_2, q_1, p_2, and the net voltage impressed on the primary of T_2 is zero.

The direct-transmission branch contains a phase and magnitude adjustment network to permit exact neutralization of the carrier voltage for the spacing condition.

Figure 9*C* is much like Fig. 9*B* except that the direct-transmission branch is omitted and a "suppressor circuit" is added, which may be thought of as changing the signals impressed on the varistor bridge from polar to neutral. This is done by inserting element s, which equalizes the voltage between points x and y whenever y is positive with respect to x. This effect is further enhanced by adding other series and shunt elements as shown. The bridge conditions for marking are the same as in Fig. 9*B*, while for spacing, all the elements are normal and alike so that the bridge is balanced for a-c. as well as for d-c.; thus no voltage appears between points w and z, and the carrier is suppressed.

While all these schemes involve balance between groups of varistors, recent advances in design have made it possible to fulfill this requirement to the desired extent and to maintain it over long periods of time.

The limited use to which varistor sending circuits have been put in the telegraph plant of the Bell System is not due to unsatisfactory operation in their present applications, but rather because they have imposed certain operating limitations on recently developed arrangements for interconnecting telegraph circuits.

Grid Bias

The fixed grid bias required by the level compensator exceeds the filament battery voltage, hence the latter cannot be used as a bias source and recourse is had to the negative 130-volt telegraph battery.

It follows that any variations in the latter will cause signal bias. To minimize this effect, several schemes have been used to stabilize the voltage applied to the level compensator. One of these is shown in Fig. 10*A*. In this arrangement the 130-volt battery discharges through

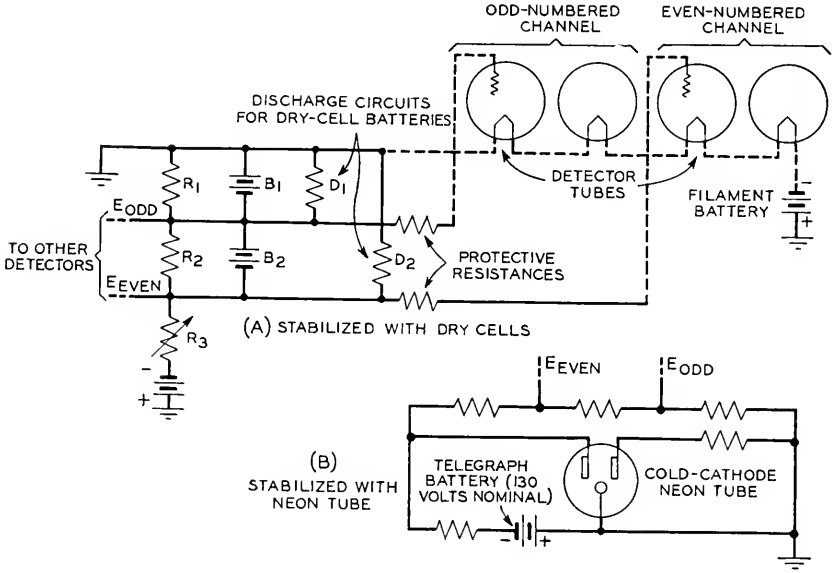


Fig. 10—Grid-bias supply circuits.

a series of resistances R_1, R_2 and R_3 so proportioned as to provide suitable taps giving the required bias when this battery is at its average voltage. Inasmuch as the filament circuits of two detectors are in series, two different voltages to ground are required. Dry cell batteries B_1 and B_2 are bridged between the taps which provide the desired biases and ground, of such values that they would give the proper voltages in the absence of the telegraph battery. These dry cells insure constant bias voltages; they supply no current when the telegraph battery is at its average value; discharge when it is low, and charge when it is high. Resistance R_3 is sufficiently large so that these charging and discharging currents are kept down to very small values and the life of the cells is consequently long. Since the bias batteries are part of a rectifying circuit there is a tendency for the signal current passing from grid to filament to charge the dry cells. To compensate for this, adjustable discharge circuits D_1 and D_2 are bridged respectively across the two grid-bias taps and ground in the manner shown, and their resistances are varied according to the number of detectors deriving their bias from this source.

A second method for securing a stable grid-bias voltage is shown in Fig. 10*B*. Here, advantage is taken of the fact that the voltage required to maintain discharge in a cold-cathode neon-tube is constant, by bridging such a tube across the negative 130-volt telegraph battery in series with a resistance. The desired voltages for the even and odd numbered detectors are then derived by tapping off at suitable points on a second resistance which is connected across the neon tube.

Interference

Interference in a particular channel may manifest itself either by the presence of current when none is intended or by a diminution of the signal current during a marking condition. The former is called *spacing interference*, and tends to change spacing units to marking units; the latter is termed *marking interference*, since it is observed during marking units, tending to change them to spaces.

The principal sources of spacing interference are:

1. Unsuppressed carrier.
2. Noise, lightning, etc.
3. Infiltration from adjacent channels.
4. Modulation products.

For marking interference, these are:

1. Crowding (Saturation effects).
2. Out-of-phase parasitic currents.

Parasitic currents (noise, modulation, etc.) are usually not an important source of marking interference, as their phase relative to that of the carrier forming the signal must fall within a rather narrow range to be effective.

If there exists some unsuppressed carrier during spacing intervals which is due to the design of the sending circuit, it will be fixed in value and may therefore be taken care of in the initial adjustments of the receiving circuit. All the other effects are of a chance character, being for the most part dependent on the transmission circumstances on associated channels or circuits. These effects, therefore, lead to fortuitous distortion.

The effectiveness of all forms of interference is dependent upon the ratio of their magnitudes to that of the signals with which they interfere. On the other hand, the absolute magnitude of the greater part of this interference depends upon the signal level. To establish a balance between these two tendencies, the telegraph power per channel which is transmitted over the circuit is selected so as to minimize the

effects of interference on telegraph signals and at the same time cause as little disturbance as possible to associated telegraph or telephone circuits. In the Bell System, the four-wire cable circuits used for telegraph purposes are for the most part devoted to this use exclusively and are organized with terminal repeater gains adapted for this special service. In the case of open-wire carrier-telephone circuits, on the other hand, the overall gain from modulator input to demodulator output is fixed by the telephone requirements, and the telegraph must be adapted thereto. The power per telegraph channel in dbm.* now used on cable and open-wire circuits is shown in Fig. 1 at various points.

The effect of changing the power on the line is illustrated qualitatively in Fig. 11, in which the variations of the received current with increasing transmitted current are sketched diagrammatically for various operating circumstances. By nominal power (*a*), is meant the power which would be received if transmission took place over a linear network having a fixed gain equal to the nominal gain of the circuit. Owing principally to the reduction of repeater gain which takes place with increasing load, the current actually received with all channels marking, is less than this (*c*). If only one channel is marking, some intermediate values (*b*) will be obtained of course, while if, as in the case of regular operation, some of the channels are spacing and others marking, still other values (*c'*) will result. This saturation effect is sometimes called "crowding."

One of the contributions to spacing interference consists of ambient noise due to the combined crosstalk from all the other circuits in the cable and to external induction. This current is represented by curve *e*, which is shown as independent of the power transmitted; this corresponds to the situation existing where telegraph is a small part of the total traffic in the cable under consideration, for evidently if the power were increased on all the circuits the noise power would increase almost proportionally.

A more serious source of spacing interference consists of parasitic currents due to third-order modulation products arising directly or indirectly from the interaction of the several channels of the same system when passing through the non-linear elements of the circuit. Second-order modulation products are taken care of quite effectively by the receiving filters, due to the fact that the carriers are odd harmonics of 85 cycles, while these products being even harmonics thereof fall midway between channel frequencies. Since the attenuation of the receiving filters in the frequency range occupied by neighboring

* The symbol dbm. as used in this paper may be read "db referred to 1 milliwatt." It is intended to denote the ratio expressed in db of the power under consideration to 1 milliwatt; e.g., − 6 dbm. = .2512 milliwatt.

channels is finite, some infiltration of undesired frequencies takes place. Thus if a given channel is spacing and the adjacent ones marking, some small fraction of the flanking carriers will find their way into the idle

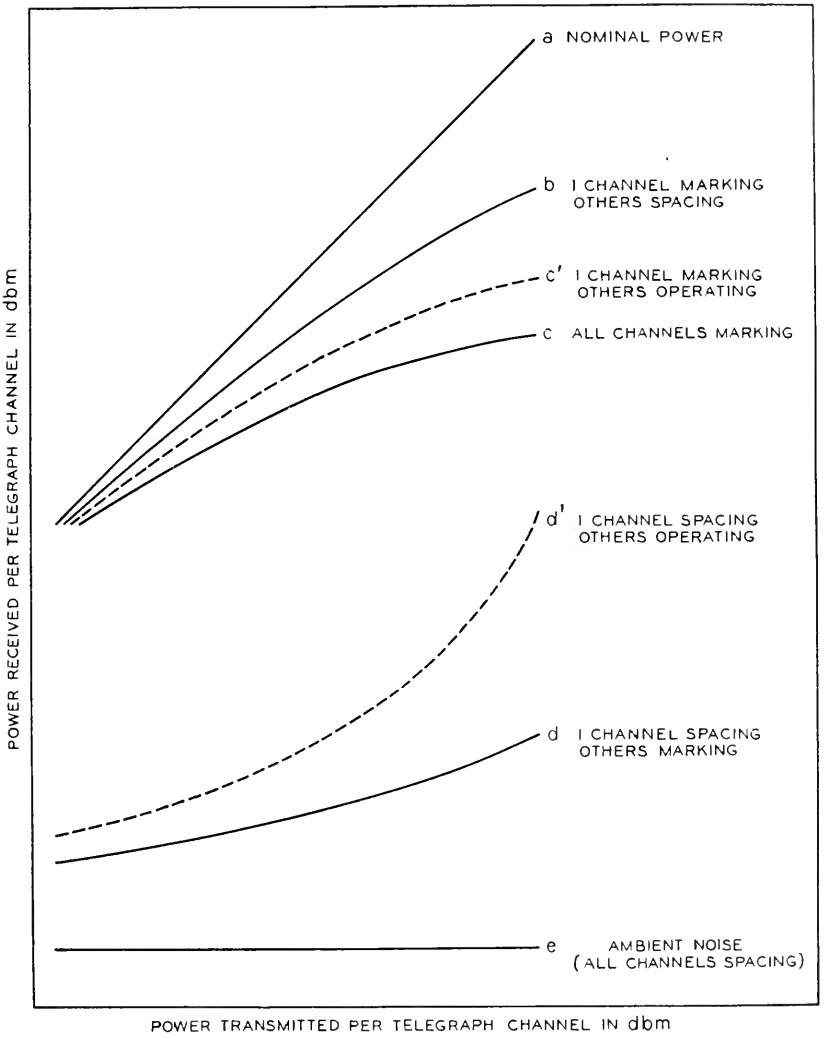


Fig. 11—Interference in cable circuits.

channel. Furthermore, if a channel is in operation, the sending filter does not completely suppress the sideband components lying outside the band assigned to it, and these pass freely through the receiving filter of nearby channels. Unlike modulation, the effectiveness of

filter action does not depend, to any large extent, upon the amount of power per channel, since the ratio of the disturbing currents to the signal remains approximately constant. Finally, such residual carrier as exists during spacing intervals may be variable in amount due to changes in the resistance of the sending relay contacts. If the power levels are increased sufficiently the total spacing interference will usually become due preponderantly to the modulation effects and hence a function of total power on the line, as indicated in curves d and d'.

In order to estimate the quality of transmission to be expected from circuits in view of these various interfering factors, it is desirable to establish the following two definitions:

The signal-to-interference ratio * of a circuit is the ratio, expressed in db, of the normal marking current plus the interference, to the interference alone.

The marking interference is the ratio, expressed in db, of the normal marking current alone, to the marking current plus the interference.

A variety of results may be obtained under these definitions depending upon the methods of observation: it is customary, therefore, to adopt the following practical specifications:

Signal-to-interference Ratio: The change in sensitivity, expressed in db, required in the receiving circuit of a given channel, all other channels being in a marking condition, to just cause the armature of the receiving relay to go to its marking contact; first, with steady marking current transmitted over the channel under test, and second, with the channel under test opened at the sending end. It is understood that the currents are turned on and off by operating the sending relays and moreover that the receiving relay operates on one half the steady marking current. (E.g., no interference $= \infty$ db; complete failure = 6 db, approximately.)

Marking Interference: The change in sensitivity, expressed in db, required in the receiving circuit of a given channel to just cause the armature of the receiving relay to go to its spacing contact; first, with steady marking current transmitted over the channel under test only, and second, with the interference added thereto. (E.g., no interference = 0 db; complete failure = 6 db, approximately.)

While the above method for measuring spacing interference is the one used in practice owing to the ease with which it can be applied,

* More precisely the signal to spacing-interference ratio.

a more significant characteristic is the signal-to-interference ratio obtained by observing the aforesaid change in receiving-circuit sensitivity required to operate the receiving relay in a given channel, as we go from a marking to a spacing condition in that channel with all other channels transmitting uncoordinated signals. This is not only a more practical consideration but a more severe condition, as indicated by curve d' in Fig. 11.

Carrier telegraphy, as here considered, is a marginal system of operation in which the current received for a marking condition corresponds on the average to that shown by curve c', while that for a spacing condition is the one represented by d'. The difference between these two characteristics is not all available for operation, however, since the receiving circuit must be made sufficiently sensitive to operate when the marking current has risen to half its final value, thus bringing the threshold of operation 6 db closer to the spacing interference. From these considerations it follows that the actual *operating margin* is an essentially variable quantity whose approximate value is less by about 6 db than the signal-to-interference ratio measured as specified above, since the indicated procedure takes account of marking as well as of spacing interference effects.

The following definitions are also useful in reporting and interpreting test results:

The *interference margins* of a circuit are the ratios, expressed in db, of the actual interference and the amount of this interference which will cause failure. More specifically:

(*a*) The spacing-interference margin is the increased sensitivity, expressed in db, required in the receiving circuit of a given channel adjusted to receive unbiased signals in the absence of interference, to just cause the receiving relay to close when interference alone is present (e.g., no interference = ∞ db; complete failure = 0 db).

(*b*) The marking-interference margin is the decrease in sensitivity, expressed in db, required in the receiving circuit of a given channel adjusted to receive unbiased signals in the absence of interference, to just cause the receiving relay to open when interference is added (e.g., no interference = 6 db, approximately; complete failure = 0 db).

Clearly the various effects which have been defined above, being of a variable and indeterminate character, contribute to the fortuitous distortion of signals. This is illustrated in Fig. 12, which shows what

may happen to a single dot. The extreme conditions of operation on the particular channel under consideration are those where all of the channels are marking or spacing. If they are all spacing, there is only noise and unsuppressed carrier, and the change in received current

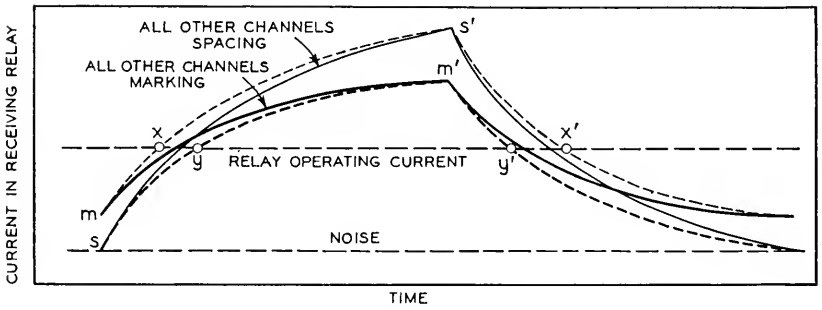


Fig. 12—Effect of interference on signal distortion.

when the one operating channel changes from space to mark is the vertical distance between points s and s', while when all the other channels are marking the change is from m to m'. The ratio of the currents corresponding to either of these pairs of points, when expressed in db, may be conveniently called the *marking-to-spacing ratio.* Practically, of course, the power over the line will be constantly and fortuitously varying, so that the actual arrival curves will lie somewhere between certain limiting values indicated by the dotted lines ms' and sm', giving rise to a range of fortuitous distortions which, if the receiving equipment was adjusted with all the other channels marking, will depend on the length of the extreme intervals xx' and yy'.

The results obtained in the course of experimental observations of some of the above quantities are given below. They were secured on a 700-mile H44 cable circuit of the type described at the beginning of this paper. In order to obtain uniform results, the output of each of the 17 repeaters in tandem in this circuit was adjusted to the same level, so that each output tube contributed about equally to the total modulation effects and a similar uniformity existed relative to the most heavily energized loading coils. In practice, the saturation effects would not be likely to exceed this, and generally would be somewhat less.

Figure 13*A* shows typically the effect on the received current of increasing the total power transmitted over the line when all channels are marking simultaneously. This phenomenon (which may be called crowding) is a measure of the intermodulation which is caused by the

circuit during normal operation, since the power transmitted over the circuit then varies fortuitously over a range of 10.8 db for a 12-channel circuit and 13.8 db for a 24-channel system, depending on the number of channels which happen to be marking simultaneously.

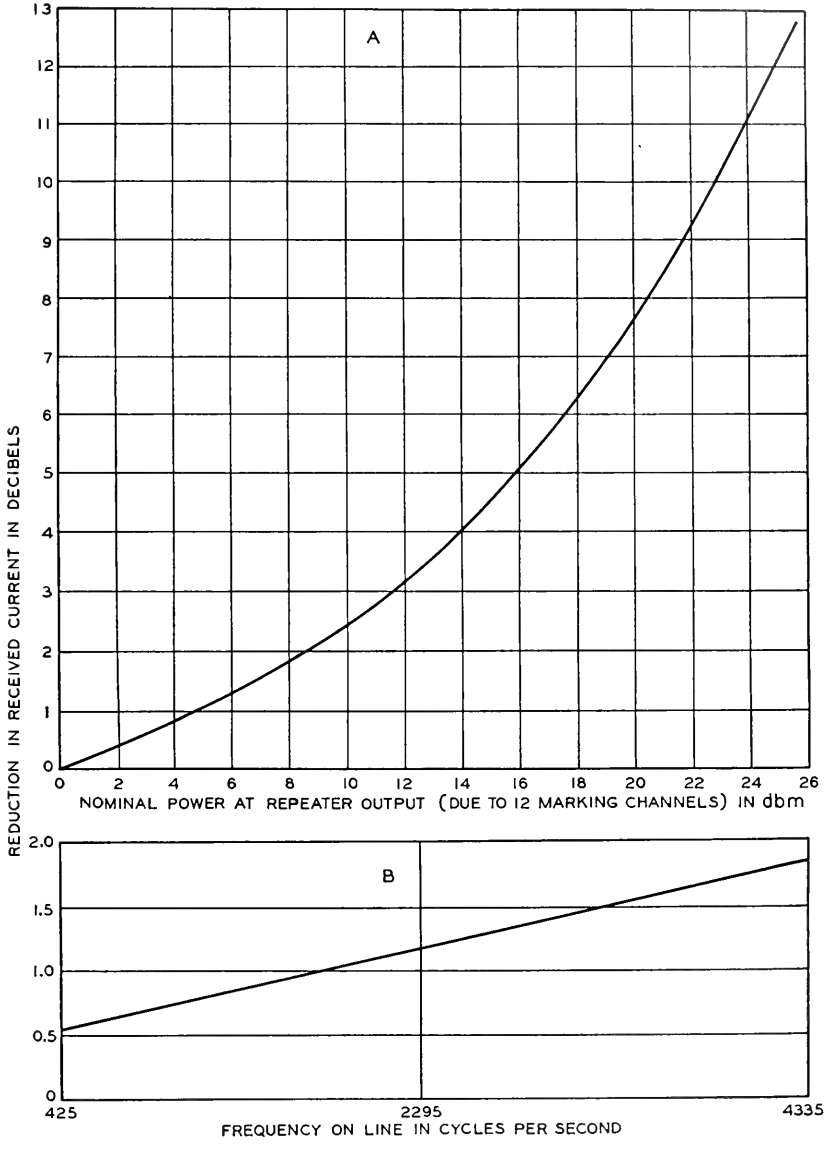


Fig. 13—Crowding. H44 circuits. *A*. Relative crowding vs. power. 12 channel system. Channel No. 5 (1105 cycles per second). *B*. Crowding vs. frequency. 24 channel system. Total power at repeater output increased from 2.2 to 9.2 dbm.

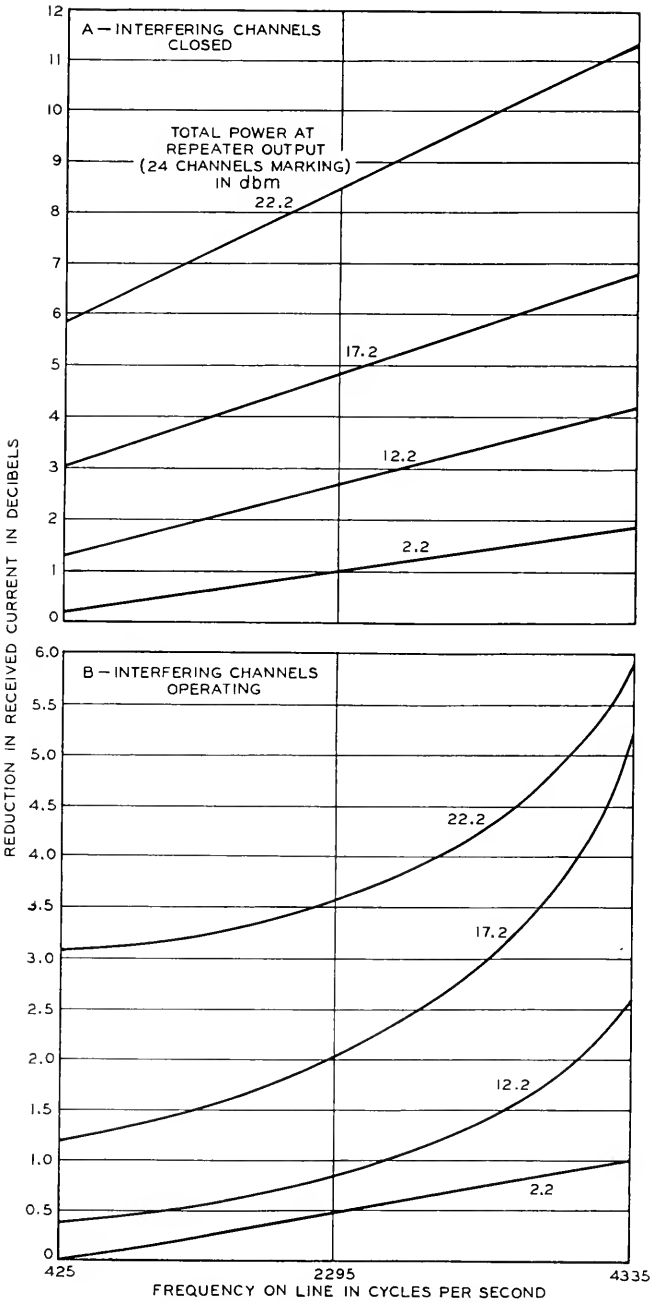


Fig. 14—Marking interference vs. frequency. H44 circuits. 24-channel system.

As will be seen by reference to Fig. 13*B* there is a marked frequency effect; the variations becoming greater for the higher frequency channels. This is particularly noticeable where there is a large number of channels operating simultaneously, as otherwise individual variations between channels tend to obscure the gradual trend.

Figure 14*A* shows the marking interference, expressed in db, due to changing from one channel marking to all channels marking for various changes in repeater output ranging from 2.2 to 22.2 dbm. and a frequency range extending from 425 to 4335 cycles.

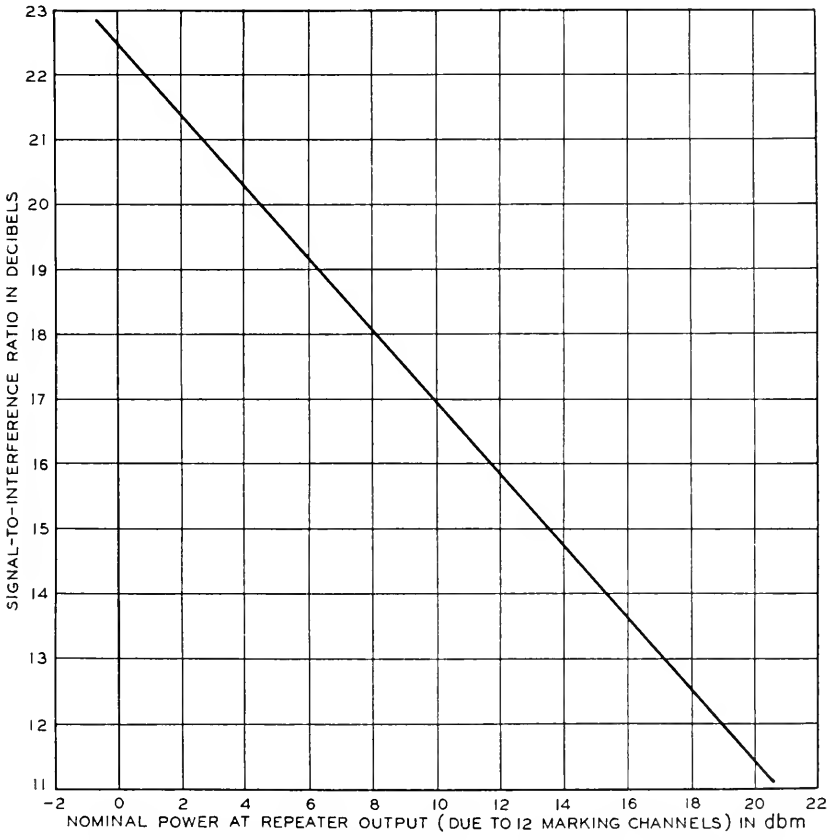


Fig. 15—Spacing interference. H44 circuits. Interfering channels operating. 12 channel system. Channel No. 5 (1105 cycles per second).

Figure 14*B* shows the reduction in received current which results on any given channel over the same range of conditions as Fig. 14*A* except that in this case the associated channels are transmitting uncoordinated reversals.

These interfering effects are subject to considerable variations from channel to channel in an irregular manner depending on repeater spacing, phase relations between the carrier sources, etc., so that the characteristics given in Figs. 13 and 14 must be interpreted as indicating average values and trends rather than specific amounts.

The erratic character of the results due to such fortuitous circumstances is particularly noticeable in spacing interference measurements made as a function of channel frequency, or in those relating to parasitic currents produced by steady currents in the remaining channels of the system. The general effect of repeater load on spacing interference is illustrated for a particular channel in Fig. 15. This refers to the case where the interference is caused by non-synchronized signals on the remaining channels.

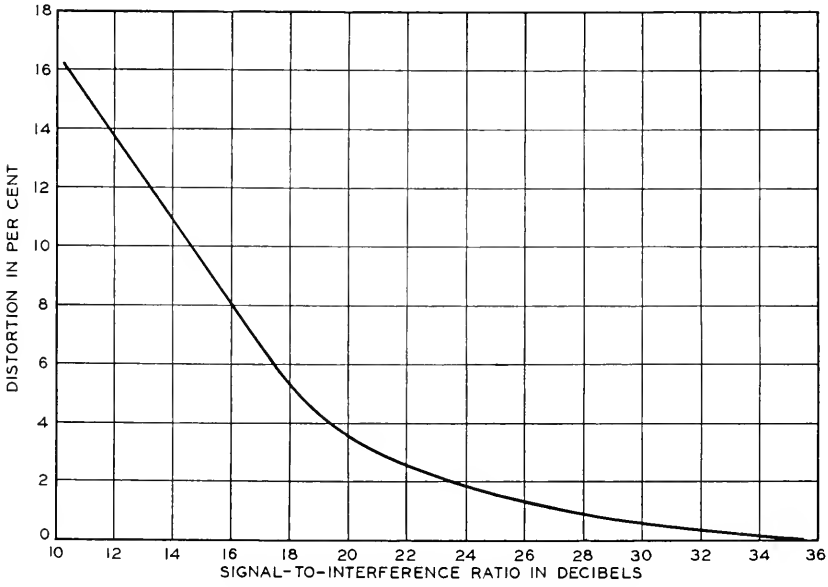


Fig. 16—Increase in signal distortion caused by spacing interference.

As far as the toll line itself is concerned, noise from other circuits is an almost negligible factor in the distortion of signals. It is of a highly fortuitous character and shows no definite trend with frequency except perhaps as it may be somewhat greater in the effective voice range, as modified, however, by variations in cross-talk efficiency with frequency.

The effect of spacing interference on telegraph distortion is given in Fig. 16, which shows that a residual current as small as 30 db below signaling begins to degrade transmission.

Operation Over Carrier Telephone Circuits

Where v-f. telegraph operates through the same repeaters as one or more telephone circuits, as in the case of carrier-telephone systems, a new form of variable interference arises due to the changing load conditions introduced by variations in voice volume. Where there are comparatively few telephone channels involved, as in the type C carrier-telephone system for instance, there is very little averaging and the voice peaks determine the repeater load which is effective in

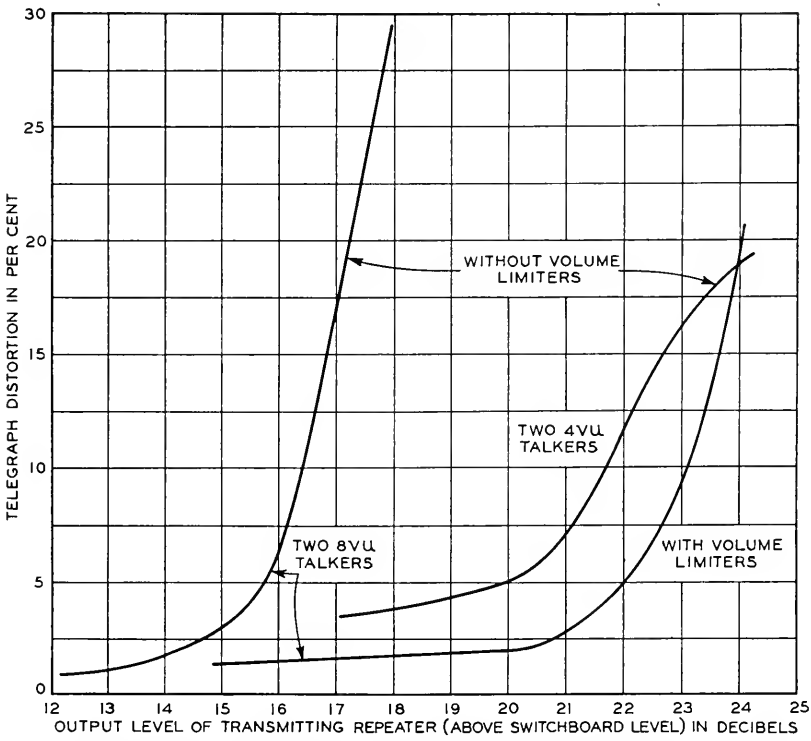


Fig. 17—Voice-frequency telegraph operated over an open-wire carrier-telephone circuit. Effect of associated telephone channels on telegraph distortion.

causing interference to telegraph. This interference is due in part to changes in net-loss of the circuit resulting from the non-linearity of amplification and in part to intermodulation between the various currents passing through the repeaters simultaneously. Figure 17 gives an example of the distortion produced in a telegraph circuit operating over one channel of a 3-channel telephone system when the other two are occupied by two equal-volume talkers. The marked

reduction in permissible repeater amplification as we change from two 4-vu to two 8-vu talkers is evident. This figure also shows that by the use of the volume limiters mentioned at the beginning of the paper, the repeater gains with the higher volume talkers may be made at least as great as for the lower volume. This makes it possible to use the present telephone circuits for carrier telegraph purposes without change.

Drainage

The occurrence of atmospheric disturbances, and particularly lightning, constitutes a potential hazard to the operation on open-wire circuits of a service as exacting as carrier telegraph. Where such circuits have been transposed for the operation of carrier telephone, the transverse or metallic-circuit effects due to lightning discharges in the neighborhood of the line are in general not serious, but the voltages generated to ground are very often of sufficient magnitude to cause a breakdown of the protectors. Since it has been found impracticable to devise protectors with such precise limits that they will operate at the same voltage and possess the same discharge characteristics, a transient transverse current is set up in such cases which may cause telegraph errors.

The remedy adopted consists in bridging drainage coils [24] across the line at all points where protectors are required, and in so connecting them that they will either prevent a breakdown of the protectors or assure simultaneous operation with equal discharge currents from either wire to ground.

In Fig. 18*A* the drainage coil is shown bridged directly across each end of the open-wire line between two sections of entrance cable. These coils consist of two carefully balanced windings with the mid-point grounded. They present a high impedance to voice or carrier currents transversely, but offer only a small resistance to ground for longitudinal currents compared with that across the adjacent protector blocks. The chief disadvantage of this method, which is called "direct drainage," is that it prevents the use of grounded telegraph and interferes with the testing of the line by means of direct current. To obviate this, the scheme shown in Fig. 18-*B* has been devised, which is termed "protector drainage." In this case, the drainage coil is connected to the line wires through protectors having a low breakdown-voltage. This combination is backed by high-voltage protectors to insure unimpeded discharge in case of large disturbances. With this arrangement the drainage coil does not come into operation unless there is a severe disturbance, and furthermore owing to the mutual inductance between the two halves of its windings it tends to cause

such discharge to occur at the same moment and be of equal magnitude from both wires of the pair to ground whenever it does operate.

Extensive tests and practical experience have shown that the above arrangements are quite effective, affording a reduction of about 95

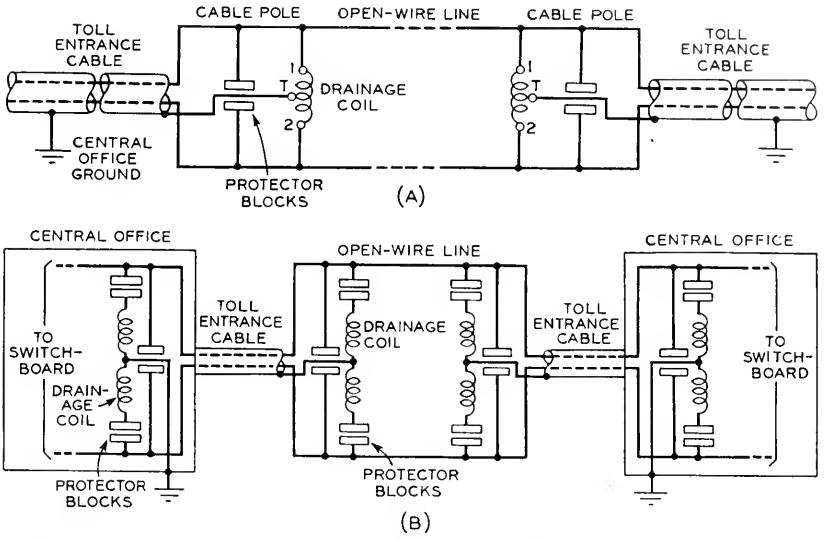


Fig. 18—Drainage arrangements. *A*. Direct drainage. *B*. Protector drainage.

per cent in the number of disturbances occurring on the line as well as in the number of errors in transmission resulting therefrom.

Carrier Supply

The inductor-alternator source for carrier frequencies used in the original installation has been retained except for the addition of two channel frequencies, and the substitution of improved means of speed (frequency) control. The first mechanical governor was soon superseded by a center-contact device which is much less erratic in operation.

The carrier frequency can be shifted over quite a few cycles from its correct value without affecting signal distortion appreciably but rapid speed variations have a more serious effect. Two methods have been used to overcome this difficulty. The first consists effectively in making the generator part of a system having great overall inertia, the second in subjecting it to very rigid control.

To secure the first end the VF generator is driven by a synchronous motor operated from the lighting circuit. This arrangement can, of course, be used only where the frequency of the commercial power is regulated within narrow limits. The required generator speed being

1700 r.p.m. a belt drive is used to reduce the motor speed of 1800 r.p.m. to the proper value; small adjustments in speed being provided by means of a V pulley having an adjustable diameter.

In order to insure continuity of operation in case of failure of the power supply the generator is also arranged to be driven by a d-c. motor equipped with a mechanical governor. This motor is supplied from the central office battery and is automatically switched thereto in case the voltage of the commercial power supply falls below a predetermined value.

In the second method the VF carrier-supply unit is governed by means of a vacuum-tube circuit whose output is controlled by frequency variations in the highest frequency channel, and the resulting d-c. current is applied to the motor field. The operation of this device, which has been described elsewhere,[25] may be briefly explained as

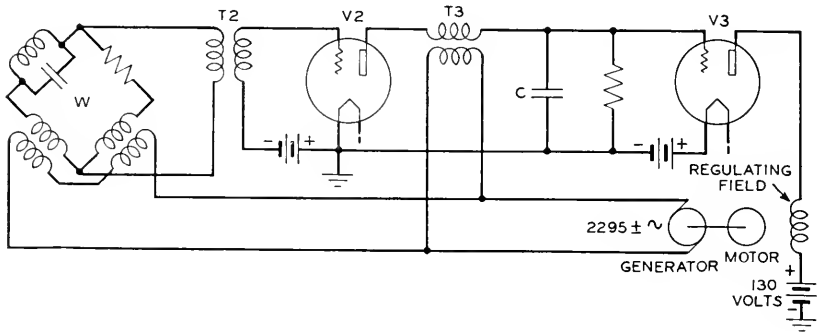


Fig. 19—Principle of tuned-circuit speed-regulator.

follows: It depends for accomplishing its purpose on producing variations in the strength of the current through an auxiliary regulating field associated with the motor which drives the multi-frequency generator. The essential features are shown in Fig. 19. The action is as follows: The voltage of channel 12 (2295 cycles) of the generator whose speed is to be controlled is applied simultaneously through a divided circuit to the input and output of tube V_2 which may be called the phase-detector tube. The plate voltage is applied directly through transformer T_3, there being no B battery in the ordinary sense. The grid voltage on the other hand is applied through the bridge W, one of the arms of which is an anti-resonant circuit tuned to 2295 cycles. The result is that the magnitude of the grid-filament voltage and its phase relative to the plate voltage are dependent on frequency. When the latter has its correct value the anti-resonant arm exactly balances the bridge, and the a-c. voltage across T_2 is nil. At higher frequencies

this circuit acts like a capacitance, while at lower frequencies it acts like an inductance. There is thus a rapid change in both the voltage and in the phase thereof across T_2 whenever there is a variation in frequency on either side of the specified value. The output current from V_2, combining with the current impressed directly through T_3, produces corresponding abrupt changes in the d-c. component of the resultant which in turn varies the bias of tube V_3 and hence the current through the regulating field of the motor.

In order to assure a rapid change in impedance with frequency, the anti-resonant circuit referred to above comprises a carefully shielded air-core coil having a very small resistance relative to its inductance.

The frequency indicator shown in Fig. 20*A* provides means for easily

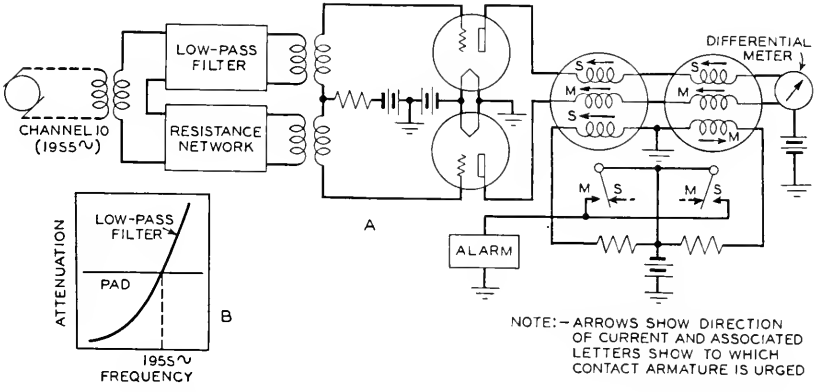


Fig. 20—Frequency indicator. *A*. Schematic diagram. *B*. Attenuation of two input-paths.

observing any departures of the carrier frequency from its nominal value, as well as an automatic maximum-minimum alarm to warn the attendant if these variations become excessive. Since all the carrier currents are derived from generator elements which are mounted on a common shaft, it is sufficient to observe the frequency of a single channel. Current from channel 10 (1955 cycles) is impressed simultaneously on two vacuum tube circuits which are identical except for the fact that there is a low-pass filter in the input of one while there is a simple pad in the input of the other. As indicated in Fig. 20-*B*, the loss through the pad is the same at all frequencies and equal to that of the filter when the generator speed is correct. If the frequency increases, the attenuation of the filter branch goes up; while if it decreases, its attenuation goes down; but in any case the loss through the pad remains constant, of course. The net resulting ampere-turns

tend to move one or the other relay armature to the opposite contact according as the frequency is high or low, and thereby ring the alarm. This resulting current also indicates the amount by which the frequency departs from its nominal value.

The frequency indicator may be used with either the mechanical governor or with the synchronous drive, but it is of little use with the vacuum tube tuned circuit control as the latter is too precise to register any indications.

The frequency indicator does not permit a very close adjustment of the mechanical governor nor does it provide a satisfactory check for the correctness of the frequency of the commercial power when a synchronous motor is used, so a stroboscopic method has been adopted as an ultimate standard of comparison. This stroboscope consists of a cylindrical target made up of three distinct peripheral rows of alternate black and white segments mounted on the end of the generator shaft. These segments may be viewed by means of a tuning fork fitted with overlapping metal plates attached to the ends of the tines. Slits cut in these plates lie opposite each other when the fork is at rest. When it is set vibrating, vision through the slits can therefore be established momentarily twice during each complete oscillation. By illuminating the target with a steady source of light the apparent direction of motion of the dots can thus be observed by looking through the slits. The middle row of segments on the target is so proportioned as to appear stationary if the speed is practically correct while the outer rows appear respectively stationary if the speed is approximately 1 per cent above or below the nominal value.

For offices where the frequency of the commercial supply is sufficiently stable, an additional and somewhat more convenient method for checking the speed has been made available. It consists of a special target mounted on the generator shaft, which is illuminated by a neon lamp associated with a wave-shaping circuit which makes the flashing time a very brief portion of each pulse of the 60-cycle current. In other words, the attendant in this case looks at the target constantly under intermittent illumination, while in the former case he views it intermittently under constant illumination.

Testing Facilities

While developments in carrier telegraph equipment have resulted in considerable economies, the ever increasing demand for service which is freer from errors and interruptions and is adaptable to circuits of greater length and complexity has tended to render the maintenance problem increasingly difficult and time consuming. Voice-frequency

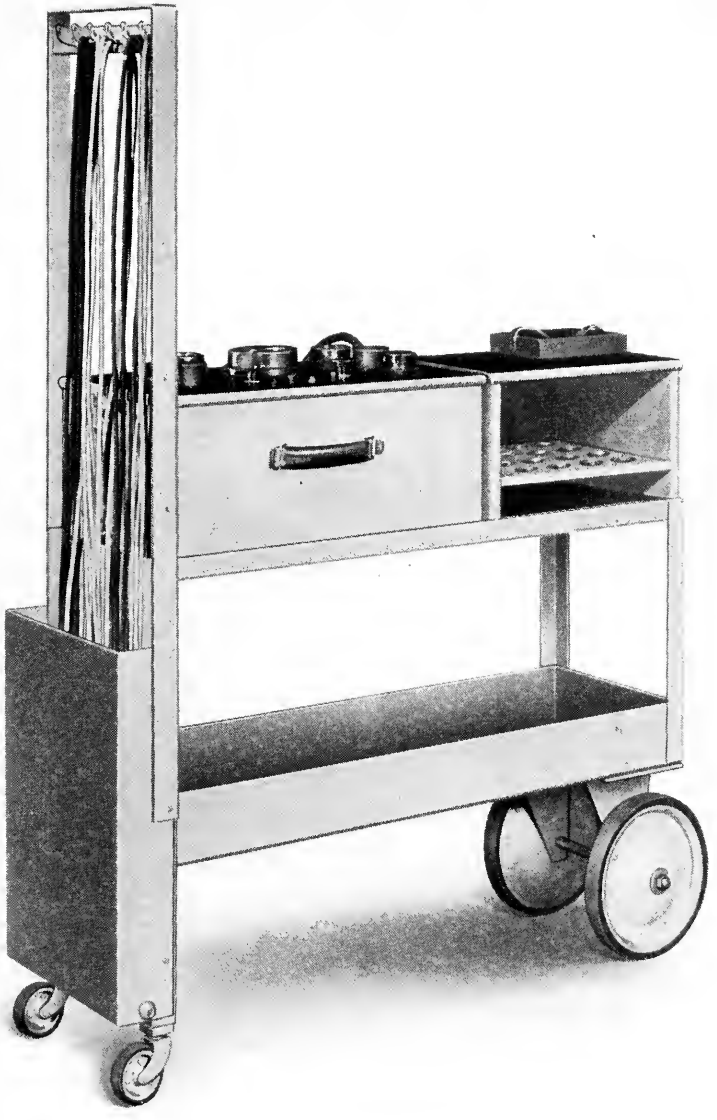


Fig. 21—Carrier-telegraph test set. General view.

telegraph circuits equipped with level compensators must be subjected to a series of specialized tests and adjustments whenever placed in service and at periodic intervals thereafter. To provide for this, a special testing set comprising in readily available form all the test equipment required for this purpose, has recently been introduced. It includes the following features:

1. Bias measuring circuit.
2. Filament-activity test-circuit and filament-current measuring circuit.
3. Drift measuring circuit.
4. Test amplifier.
5. Adjustable attenuator.

With this set all terminal tests may be made from one end of the circuit without the use of a line or external line-simulating repeater as

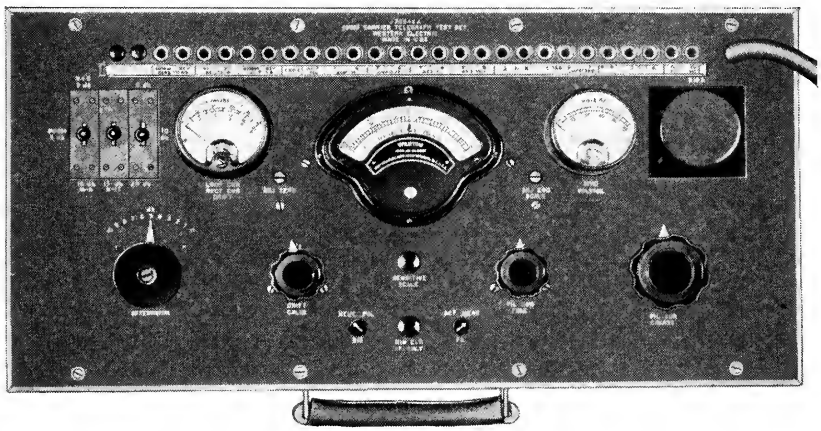


Fig. 22—Carrier-telegraph test set. Instrument panel.

was done in the past. It is mounted on a small wagon which may be wheeled into position adjacent to the terminals to be tested, then connected by means of a long cord and multiple contact plug to the necessary battery supplies, grounds, etc. The general appearance of the set is shown in Fig. 21, while a more detailed view of the face equipment may be obtained from Fig. 22.

The bias measuring circuit is shown in simplified form in Fig. 23. It uses a 215 type polar relay,[26] or its equivalent, with a meter connected in the armature circuit. This meter, which permits measuring to an accuracy better than 1 per cent, is specially designed to have the proper ballistic and damping characteristics to permit measuring dot

signals having a rate of 11 d.p.s. The reason for providing 11-cycle reversals is that experience has shown that these more nearly simulate miscellaneous teletypewriter signals, both with respect to bias and drift, than the higher speed reversals used in the past. An end scale

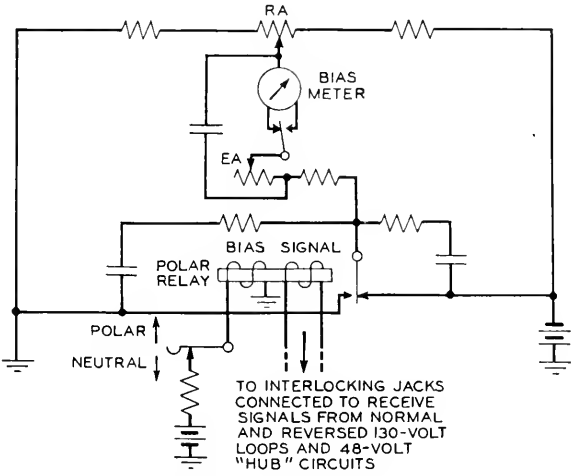


Fig. 23—Carrier-telegraph test set. Bias measuring circuit.

adjustment *EA* permits correction for battery variations, while a second adjustment *RA* allows for correcting any residual bias which may be present in the 215 type relay or in the dot signals used for the tests.

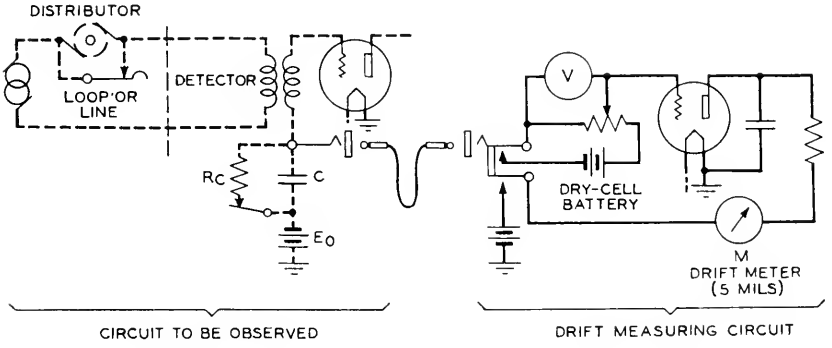


Fig. 24—Carrier-telegraph test set. Drift measuring circuit.

In order to observe and correct for the drift effect discussed previously, a special circuit is provided which is illustrated in Fig. 24. The skeletonized diagram of the carrier-terminal circuit to which it is applied, shown to the left in dotted lines, will help to understand the

function of this measuring device. The object is to observe the change in voltage experienced by the upper plate of the level compensator condenser when incoming signals are changed from steady marking to dots. The circuit provided for this purpose is essentially a vacuum-tube electrostatic voltmeter of the conventional type.

Acknowledgment

The advances in the art which have been described in the foregoing pages have taken place over a number of years and cover a considerable variety of subjects. Much of the material is to be found only in test reports and unpublished memoranda. Since such work is, of necessity, the product of the cooperation of many minds, it is impracticable in most cases to apportion credit equitably and the author must therefore confine himself to a general expression of indebtedness to his associates.

APPENDIX

LEVEL COMPENSATOR THEORY

The theory of the level compensator may best be considered in two stages, first taking the steady marking condition, and second the condition where signals are being received.

Steady Marking Condition

Figure 25*A* shows the essential elements involved when the steady current I_0 flows through the receiving relay. Figure 25*B* is the

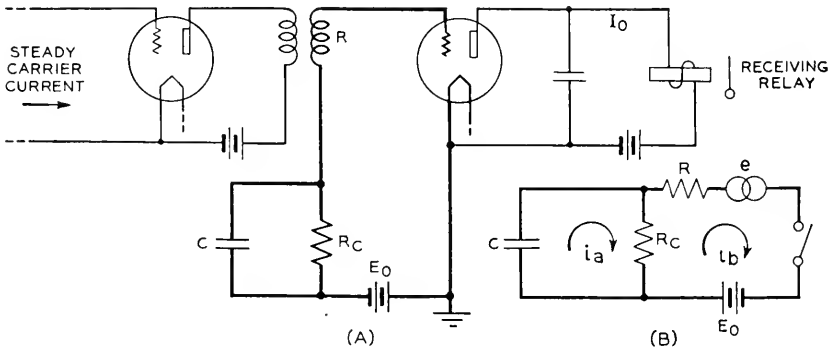


Fig. 25—Theory of level compensator. *A*. Simplified circuit. *B*. Equivalent circuit.

equivalent of the grid circuit shown in heavy lines on Fig. 25*A*, on the assumption that the grid-filament space is effectively a switch which turns the grid current on or off when the grid goes positive or negative,

respectively, and that the interstage transformer is perfect. The tube resistance which is effective during the conducting period may be included in R and considered constant, as it is relatively small compared to the other resistances involved.

In describing the principle of action of the level compensator, it has been pointed out that for proper compensation the point of relay operation AA (Fig. 7) should bisect the crest value of the signal pulse in each case. In this figure, the three envelopes shown correspond to very short signal pulses; if the signals were sufficiently long so that a steady marking condition were reached, as is usually the case, the crest of the signal would coincide with the peaks of the steady carrier wave, and AA would bisect its positive loops.

For long signals the condition for proper compensation is therefore that the relay will just operate at one-half the steady-state carrier voltage. This is true for all carrier voltages E which equal or exceed the value required to make the grid positive. In the particular case where E just fails to cause the condenser to charge, $E = -E_0$ and the relay operates when a value $-E_0/2$ is reached. For any other greater value of E it is necessary in addition to overcome the voltage due to the charge on the condenser before the relay will operate. The criterion for perfect compensation where signals are of sufficient duration so that the steady state is reached is therefore:

$$\frac{E}{2} + \frac{E_0}{2} - e_c = 0, \tag{1}$$

where E is the maximum value of the instantaneous steady state carrier voltage, and E, E_0 are arbitrarily taken with such polarities as to urge the mesh current i_a in the direction indicated in Fig. 25B, while e_c is negative because it opposes the current i_a which gives rise to it.

Substituting $e_c = Q/C$ in (1) we have

$$\frac{E}{2} = -\left(\frac{E_0}{2} - \frac{Q}{C}\right), \tag{1a}$$

where Q is the average charge existing on the condenser during a long mark for any particular value of E. The problem of compensation then resolves itself in adjusting the constants C, R, R_c, and E_0 so that relation (1a) will hold; in other words an expression for Q must be found in terms of these parameters. The problem is not susceptible of explicit solution, but expressions will be derived which are believed to clarify the operation of the compensator and permit computation.

Referring to Fig. 26 let A represent the positive halves of the open circuit voltage across the secondary of the interstage transformer due to the steadily impressed carrier; and B represent the instantaneous

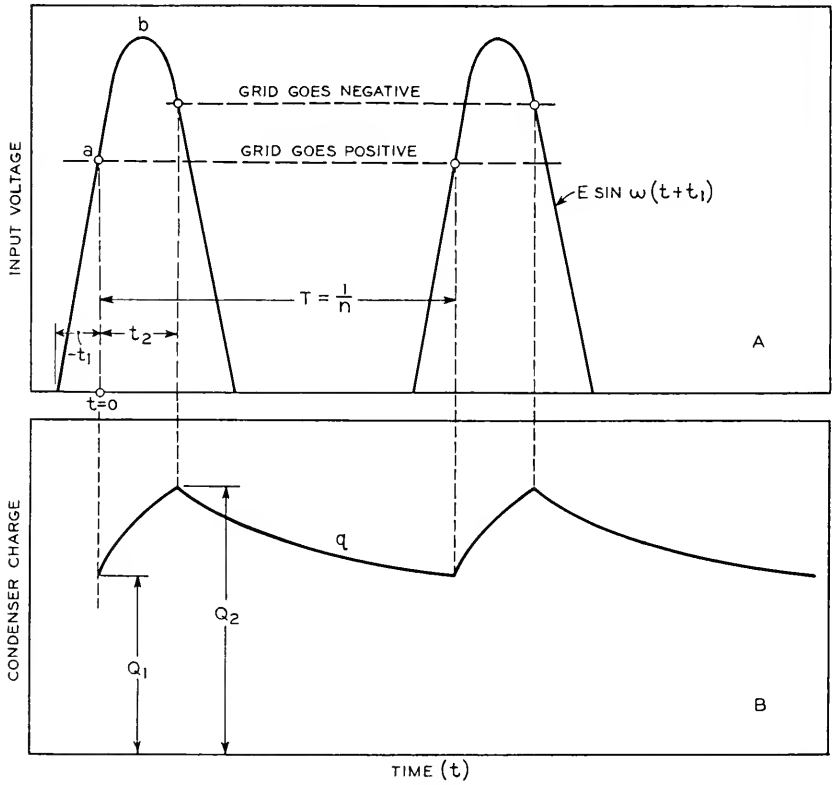


Fig. 26—Theory of level compensator. A. Carrier voltage. B. Condenser charge.

charge on the condenser. Reckoning time from the instant a when the grid goes positive, we have:

$$e = E \sin \omega(t + t_1), \tag{2}$$

where $\omega/2\pi$ is the carrier frequency.

Consider the instant at which $t = -t_1$: e is increasing through its zero value, the switch representing the grid-filament space (Fig. 25B) is open and the condenser is discharging through R_c. This state of affairs continues until e reaches a value which nullifies the condenser voltage plus the steady grid bias so that the voltage across the switch

is zero. At $t = 0$, the switch closes and current i_b starts to flow. If Q_1 is the charge on the condenser at $t = 0$,

$$\frac{Q_1}{C} = E \sin \omega t_1 + E_0. \tag{3}$$

The switch remains closed while e increases to its crest value at b and then decreases until, at time $t = t_2$, the grid goes negative. If Q_2 is the charge on the condenser at $t = t_2$

$$\frac{Q_2}{C} = E \sin \omega(t_1 + t_2) + E_0. \tag{4}$$

The switch now opens, and the condenser discharges through R_c until the switch closes once more at $t = T$ to begin another cycle, T being the carrier period $2\pi/\omega$. During this interval the charge on the condenser is given by

$$q = Q_2 \epsilon^{-(t-t_2)/CR_c}. \tag{5}$$

In order that steady-state conditions may prevail, that is, in order that every cycle be the same as its predecessor, q must again equal Q_1 when $t = T$. Hence

$$Q_1 = Q_2 \epsilon^{-(T-t_2)/CR_c}$$

or, setting

$$D = \epsilon^{-(T-t_2)/CR_c}$$

$$Q_1 = Q_2 D. \tag{6}$$

In connection with equation (1a) it has been pointed out that what is sought is an expression for the average charge Q in terms of the parameters. Referring to Fig. 26*B*, it will be seen that this average lies somewhere between Q_1 and Q_2, but since Q_1 is very large compared with $Q_2 - Q_1$, we may take either Q_1 or Q_2 as equal to the average charge Q in (1a), when considering the effect of bias on the grid of the detector tube in producing compensator action. For the sake of definiteness, let:

$$\frac{E}{2} = -\left(\frac{E_0}{2} - \frac{Q_1}{C}\right). \tag{1b}$$

Equations 3, 4 and 6 contain the five unknowns E, Q_1, Q_2, t_1, t_2; hence one more relation must be established before we can get the desired formulation between E and Q_1. The circuit equations for Fig. 25*B* when the switch is closed (charging period) furnish the required relation. Thus:

$$\frac{q}{C} + R_c(i_a - i_b) = 0$$

$$(R + R_c)i_b - R_c i_a = E \sin \omega(t + t_1) + E_0.$$

Eliminating i_b, setting $i_a = \frac{dq}{dt}$ and $\alpha = (R + R_c)/RR_cC$ we have

$$\frac{dq}{dt} + \alpha q = \frac{E}{R}\sin \omega(t + t_1) + \frac{E_0}{R},$$

whence

$$q = EPF(t + t_1) + \frac{E_0}{\alpha R} + A\epsilon^{-\alpha t}, \tag{7}$$

where A is the constant of integration and

$$P = 1/R(\alpha^2 + \omega^2), \qquad F(x) = \alpha \sin \omega x - \omega \cos \omega x.$$

Since q obeys (7) from $t = 0$ until $t = t_2$, we have for the initial and final charges during the charging portion of the cycle:

$$Q_1 = EPF(t_1) + \frac{E_0}{\alpha R} + A$$

$$Q_2 = EPF(t_1 + t_2) + \frac{E_0}{\alpha R} + A\epsilon^{-\alpha t_2}.$$

Eliminating A

$$Q_1 - Q_2\epsilon^{\alpha t_2} = EP\left[F(t_1) - \epsilon^{\alpha t_2}F(t_1 + t_2) \right] + \frac{E_0}{\alpha R}\left[1 - \epsilon^{\alpha t_2} \right], \tag{8}$$

which is the additional equation required.

It now merely remains to eliminate some of the unknowns. To this end we first eliminate Q_1 and Q_2 from (3) and (4) by means of (6) and, solving for E, obtain

$$E = -E_0\frac{D - 1}{D \sin \omega(t_1 + t_2) - \sin \omega t_1}. \tag{9}$$

A second expression for E involving the same variables is next obtained by replacing Q_1 and Q_2 in (8) by their values as given by (3) and (4). Equating these two expressions for E leads to:

$$\tan \omega t_1 = \frac{G_2 + G_3 \sin \omega t_2 - BG_2 \cos \omega t_2}{G_1 - G_3 \cos \omega t_2 - BG_2 \sin \omega t_2}, \tag{10}$$

where

$$G_1 = (D - 1)S + II \qquad II = \left(\frac{1}{\alpha R} - C \right)\left(1 - \epsilon^{\alpha t_2} \right)$$

$$G_2 = (D - 1)P\omega \qquad S = P\alpha - C$$

$$G_3 = (D - 1)SB + IID \qquad B = \epsilon^{\alpha t_2}.$$

Referring to Fig. 26*A* it will be seen that t_2 is 0 if the input voltage is just sufficient to make the grid positive, i.e., if $E = -E_0$; while on

the other hand t_2 will not exceed π/ω as E approaches infinity. Hence, if from equation (10) we obtain corresponding pairs of t_1 and t_2 by substituting values of t_2 ranging from $\omega t_2 = 0$ to $\omega t_2 = \pi$ and solving for t_1, we may substitute the pairs of values so obtained in (9), thereby arriving at a relation between E and t_1. Corresponding pairs of these two latter quantities can finally be substituted in (3) thus obtaining a relation between E and Q_1/C.

From (1b) we may express the departure from perfect compensation as a voltage ΔV which would have to be added to the condenser voltage to bring about this ideal condition. Viz:

$$\Delta V = \frac{E}{2} + \frac{E_0}{2} - \frac{Q_1}{C}. \tag{11}$$

By inserting corresponding values of E and Q_1/C, as obtained in the preceding paragraph, the precision of compensation obtained with any given set of parameters may be calculated.

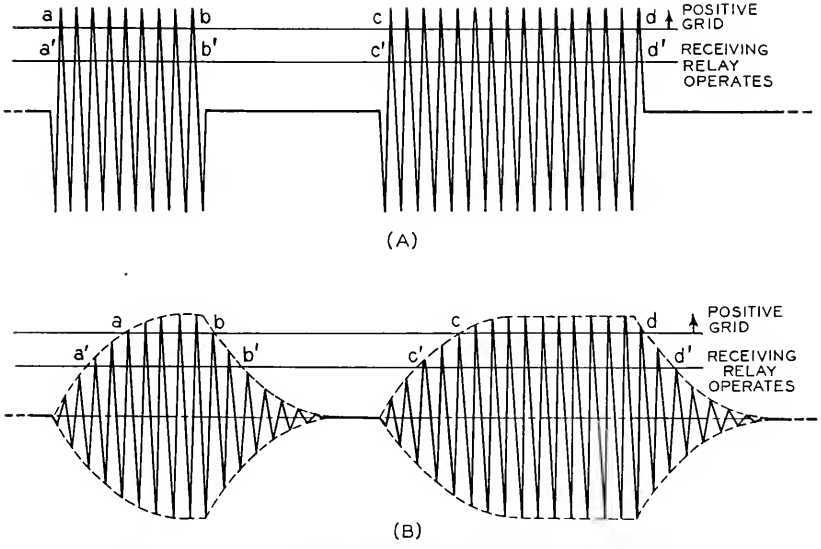


Fig. 27—Function of level-compensator relay. *A*. Square signal. *B*. Rounded signal.

Signaling Condition

Figure 27*A* shows the form which the received signals would have if the transfer admittance of the circuit were independent of frequency. If the level compensator consisted only of the simple condenser-resistance circuit shown in Fig. 25*A*, a large part of the

charge accumulated during a mark would be dissipated during the following space, and since both marks and spaces continually change in relative lengths, this would give rise to characteristic distortion. This difficulty could be taken care of by having a relay operating in unison with the receiving relay and serving to disconnect the leakage resistance R_C during spaces in the manner shown in Fig. 6. Actually, however, the presence of the channel filters causes the received signals to resemble more nearly Fig. 27*B*. This leads to a further difficulty due to the fact that the charging intervals ab, cd, etc., are shorter than the periods $a'b'$, $c'd'$, etc., during which the receiving relay is closed. To remedy this, the compensator relay is biased towards spacing by means of the resistance AB (Fig. 6) and associated battery; the operating impulses being first rounded off by the resistances, condenser, and inductance in the *wave shaping circuit* to make them susceptible of such time bias. A necessary condition to be fulfilled by this circuit is that it should supply enough energy to the compensator relay to operate it even under conditions of extreme bias.

Drift

Returning to Fig. 27*B*, we may consider the bias of the compensator relay as equal to $a'b' - ab$ for the shorter signal or $c'd' - cd$ for the longer one. These biases are, of course, equal time intervals which we will denote by δ. From this it can readily be shown that in a given period—one second for instance—the grid will be conducting during a longer time for a group of long signals than for a group of short ones. Thus, let it be assumed for example, that $c'd' = 2a'b'$ and let n be the number of marking conditions of length $a'b'$ in a given interval; we have for the cumulative charging time in the two cases:

$$T_A = n(a'b' - \delta)$$

$$T_B = \frac{n}{2}(c'd' - \delta) = n\left(a'b' - \frac{\delta}{2}\right),$$

whence

$$T_B = T_A + \frac{n\delta}{2}.$$

It follows that the charge on condenser C is greater for the longer signals. If, therefore, the receiving circuit is adjusted to give unbiased signals when dot signals at the rate of 11 d.p.s. are received, similar signals at 23 d.p.s. will be biased positively. Experience shows that an adjustment which gives zero bias with 11 d.p.s. dots will give substantially unbiased signals with the standard test-sentence; hence

circuits adjusted with 23 d.p.s. dots should be given a small initial positive bias.

It will be noted that the effective voltage available for charging the condenser is not constant throughout the interval *ab*. The maximum value which it can attain, and does attain if the signal is sufficiently long, equals the effective charging voltage during steady marking; on the other hand if the signal is very short, this value may never be reached. It follows that the charge accumulated by the condenser for any kind of intermittent signals is always less than that for steady marking; the amount of the discrepancy depending on the wave shape of the envelope.

As a result of these effects, and other similar conditions which tend to modify the average charge on the condenser depending on the character of the received signals, there is a perceptible amount of characteristic distortion manifesting itself in a fortuitous manner during the reception of ordinary text.

The change in the mean condenser charge can easily be observed if after a steady marking condition has been maintained for a few seconds, dots are suddenly impressed on the circuit and their bias observed. The latter will be found to *drift* as the charge assumes a new mean value. In practice, an adjustment is made for this by observing the change in voltage across the condenser under these two conditions and adjusting the compensator-relay bias to maintain the drift within limits which experience has shown to give minimum distortion with ordinary text.

Bibliography

1. Colpitts, E. H. and Blackwell, O. B., "Carrier Current Telephony and Telegraphy," *A. I. E. E., Trans.*, Vol. 40, pp. 205–300, 1921. (Has extensive bibliography.)
2. Bell, J. H., "Carrier-Current Telegraphy," *Bell Labs. Record*, Vol. 1, pp. 187–192, Jan., 1926.
3. Hamilton, B. P., Nyquist, H., Long, M. B., and Phelps, W. A., "Voice Frequency Carrier Telegraph System for Cables," *A. I. E. E., Trans.*, Vol. 44, pp. 327–332, 1925.
4. Nancarrow, F. E., "Line Telegraphy: More Voice Frequency Channels," *Electrician*, Vol. 122, p. 108, Jan. 27, 1939.
5. Harrison, H. H., "Printing Telegraph Systems and Mechanisms," Lond., Longmans, 1923.
6. Bell, J. H., "Printing Telegraph Systems," *A. I. E. E., Trans.*, Vol. 39, pp. 167–230, 1920.
7. Pierce, R. E., "Modern Practices in Private Wire Telegraph Service," *A. I. E. E. Trans.*, Vol. 50, pp. 426–435, 1931.
8. Gray, Elisha, "Nature's Miracles," Vol. 3, *Electricity and Magnetism*, pp. 126–128. N. Y., Fords, Howard and Hulbert. 1900.
9. "Il y a Cinquante ans: Essai du telegraphe harmonique de E. Gray," *Rev. Gen. de l'Elec.*, Vol. 32, pp. 361–362, 1932.
10. Watson, E. F., "Fundamentals of Teletypewriters Used in the Bell System," *Bell Sys. Tech. Jour.*, Vol. 17, pp. 620–639, 1938.
11. Pierce, R. E. and Bemis, E. W., "Transmission System for Teletypewriter Exchange Service," *Bell Sys. Tech. Jour.*, Vol. 15, pp. 529–548, 1936.

12. Clark, A. B., "Telephone Transmission over Long Cable Circuits," *A. I. E. E., Trans.*, Vol. 42, pp. 86–96, 1923.
13. Green, E. I., "Transmission Characteristics of Open-Wire Telephone Lines," *A. I. E. E., Trans.*, Vol. 49, pp. 1524–1535, 1930.
14. Affel, H. A., Demarest, C. S., and Green, C. W., "Carrier Systems on Long Distance Telephone Lines," *Bell Sys. Tech. Jour.*, Vol. 7, pp. 564–629, 1928; *A. I. E. E., Trans.*, Vol. 47, pp. 1360–1387, 1928.
15. O'Leary, J. T., Blessing, E. C., and Beyer, J. W., "Improved Three-Channel Carrier Telephone System," *Bell Sys. Tech. Jour.*, Vol. 18, pp. 49–75, 1939.
16. Cowley, G. W., "Volume Limiter Circuits," *Bell Labs. Record*, Vol. 15, pp. 311–315, June, 1937.
17. Peterson, Eugene and Keith, C. R., "Grid Current Modulation," *Bell Sys. Tech. Jour.*, Vol. 7, pp. 106–139, 1928.
18. Nyquist, H., Shanck, R. B., and Cory, S. I., "Measurement of Telegraph Transmission," *A. I. E. E., Trans.*, Vol. 46, pp. 367–376, 1927.
19. Shanck, R. B., Cowan, F. A., and Cory, S. I., "Recent Developments in the Measurement of Telegraph Transmission," *Bell Sys. Tech. Jour.*, Vol. 18, pp. 143–189, 1939.
20. Kupfmuller, K., "Über Einschwingvorgänge in Wellenfiltern," *E. N. T.*, Vol. 1, pp. 141–152, 1924.
21. Nyquist, H., "Certain Topics in Telegraph Transmission Theory," *A. I. E. E., Trans.*, Vol. 47, pp. 617–644, 1928.
22. Stahl, H., "Versuche über eine gunstige Verteilung der Tragerwellen in der Wechselstromtelegraphie," *T. F. T.*, Vol. 19, pp. 340–347, 1930.
23. Herman, J., "Bias Control in Telegraph Communication," U. S. Patent 1,886,808, Nov. 8, 1932; "Transmission Level Control in Telegraph Signaling Systems," U. S. Patent 1,943,478, Jan. 16, 1934.
24. Cash, C. C., "Protection against Lightning Interference," *Bell Labs. Record*, Vol. 15, pp. 125–128, Dec., 1936.
25. Trucksess, D. E., "High Precision Speed Regulator," *Bell Labs. Record*, Vol. 13, pp. 187–190, Feb., 1935.
26. Fry, J. R. and Gardner, L. A., "Polarized Telegraph Relays," *A. I. E. E., Trans.*, Vol. 44, pp. 333–338, 1925.

Electrical Drying of Telephone Cable

By L. G. WADE

DRYING equipment has recently been installed at the Kearny Works of the Western Electric Company whereby exchange area telephone cable is heated to a temperature of 270 degrees Fahrenheit by passing direct electric current through the copper wire conductors. Before discussing this installation in detail, however, it might be well first to outline the type of material handled and to review briefly the history of telephone cable drying methods and the reasons for the changes that have occurred.

Telephone cable consists of a number of individual paper ribbon or pulp insulated wires grouped together and the whole then covered with a serving of wrapping paper before being sheathed in lead. Cables may vary in length from a few feet up to several thousand feet, and in number of pairs from 6 to 2121. The size of the wire ranges from 26 American Wire Gauge to 10 American Wire Gauge, with some of the product often containing as many as two or three different sizes of wire in the same cable. One or more cable lengths are wound on a core truck which may be readily moved from one place to another by means of an electric truck.

Early cables were textile insulated and then dried by placing the cores in a heated oven, followed by boiling in a tank containing a sealing mixture or impregnant. The impregnant was used to keep the cable relatively dry in the rather imperfect lead sheath developed at that time. However, with the advent of an improved lead sheath extruded directly on the core, which would guarantee the excluding of any water, it was found desirable to go to a dry paper insulated telephone cable. This change in design reduced capacitance to about one-half of the previous values. The paper insulated cable was dried in a brick oven with gas retorts below a grille floor for maintaining the oven temperature between 215 degrees and 250 degrees Fahrenheit.

This method of drying produced satisfactory results for short-haul cables in use at that time. However, with the possibility of longer lead covered cable circuits replacing open telephone lines, it became necessary to obtain a greater degree of dryness in order to meet the new demand for transmission quality. The toll cables, which were for the longer haul, were therefore given an additional drying after the

sheathing operation. This was accomplished by passing calcium chloride dried air at a temperature of 270 degrees Fahrenheit through the heated cable for a period of twenty-four hours. About 1917 the brick ovens were replaced by steam heated vacuum tanks which reduced the drying period to about one-third that used for the brick ovens and very greatly improved the dryness of all types of telephone cable.

Calcium chloride drying of toll cable was continued until 1927. It then became necessary to provide a means of keeping the cable from regaining appreciable amounts of moisture between the vacuum drying operation and lead covering. After considerable investigation a dry core storage room was developed where the dry cable could be held at .3 per cent to .5 per cent relative humidity until ready for covering.[1] With the improvement in vacuum tank drying and with cable stored under such a dry condition until the protective sheath could be applied, transmission quality of the shorter lengths of cable approached the level of the calcium chloride drying while there was some improvement in drying the longer lengths. This change in handling cable resulted in a large reduction in drying cost.

Still further improvements in drying methods have been obtained by heating the cable electrically rather than by radiation from steam coils in the drying chamber. Considerable thought had been given to this method over a number of years and the first unit of equipment was installed experimentally at the Baltimore Works in 1931. This unit has been in successful operation since that time for drying a part of the toll cable output and has furnished most of the data used in engineering the Kearny installation. The choice of the Kearny Plant for the first large scale installation was due largely to reduced operating and maintenance costs. The following discussion covers the type of equipment used in this installation and points out in closing some of the advantages gained in reduced cost and better drying of cable.

In preliminary experimental work on electrical drying a low voltage transformer was used to supply alternating current and the cable rendered non-inductive on its core truck by short-circuiting one end and dividing the other end and attaching to the source of power. However, such a set-up places the full voltage between a considerable portion of the conductors near the clamped ends, a condition not suited for telephone cable. Since, in telephone cable, the conductors are insulated with a thin tube of pulp or ribbon paper, the insulation

[1] Drying and Air Conditioning in Cable Manufacture, J. Wells and L. G. Wade, *Chemical and Metallurgical Engineering*, March 1932.

resistance between wires is low in an undried cable, particularly when moisture is leaving the cable during the drying operation. The choice of direct current not only eliminated the danger of breakdown between conductors (all conductors are grouped in parallel) but simplified the preparation and clamping of the cable for drying. Direct current also makes it possible readily to obtain practically any starting voltage between the maximum and minimum range of the equipment, a requirement necessitated by the great variety of cable lengths.

The electric drying installation at Kearny consists of a motor generator set for supplying the direct current for heating, Fig. 1, control

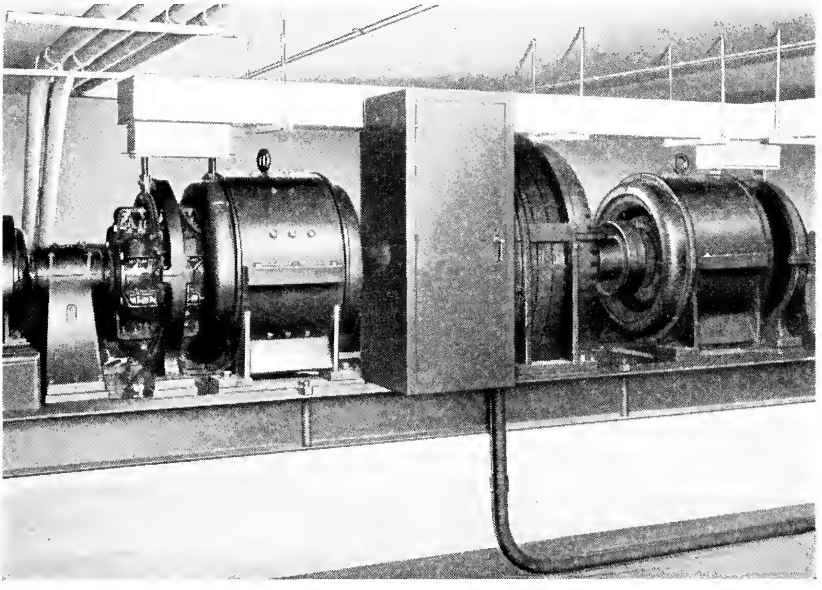


Fig. 1—Power supply set showing two generators for furnishing heating current.

equipment for starting and regulating the motor generator set, Fig. 2, and dome type dryers for holding the cable and its core truck under a vacuum of ½ inch to 1 inch of mercury back pressure during the heating and evacuation period, Fig. 3. A central control panel is located at the dryers on which is mounted apparatus for limiting the heating period to a predetermined amount, as well as visual instruments for indicating information for the operator's use in properly running the drying process.

All cables are dried at the same starting current density per unit of cross-section of the copper wire. With this as a starting point the

capacity and type of the motor generator set were determined from the sizes and lengths of cables to be dried, balancing the first cost and efficiency against loss of capacity when handling some cables requiring

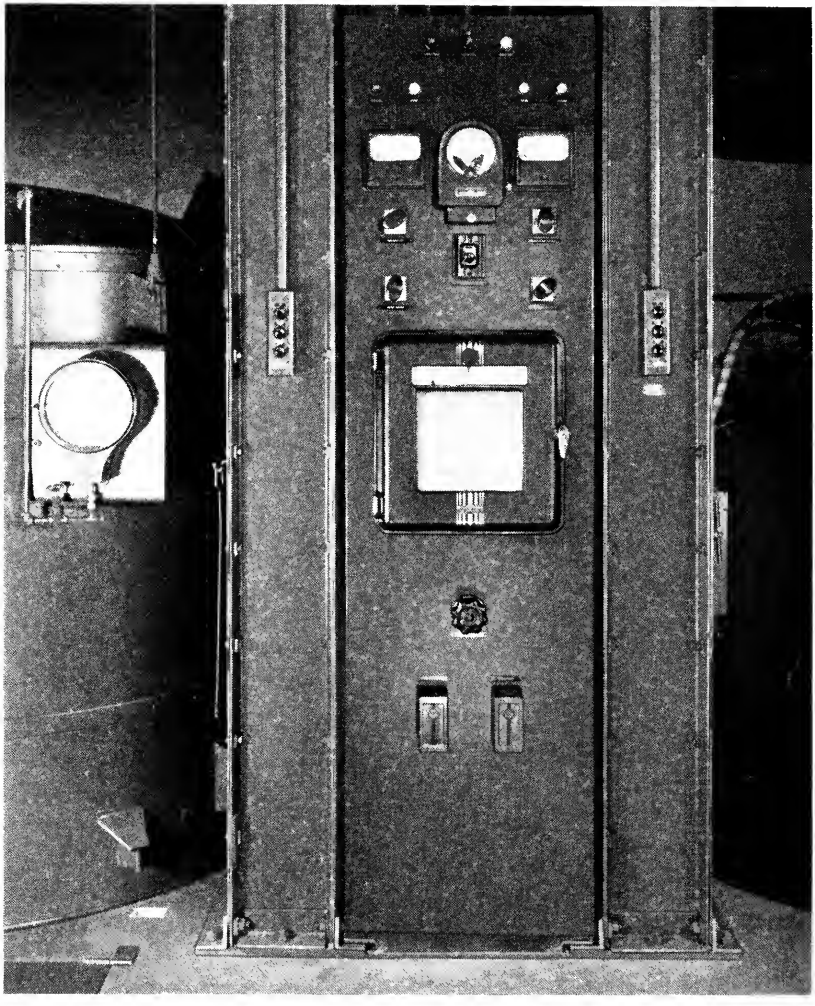


Fig. 2—Control panel for starting motor generator set and controlling the drying operation.

a longer heating period. This latter might be due to too large a cable or too long a length to be heated at the standard rate. In determining the capacity for a fixed starting current density, consideration was given to the cost of equipment, length of heating period and general

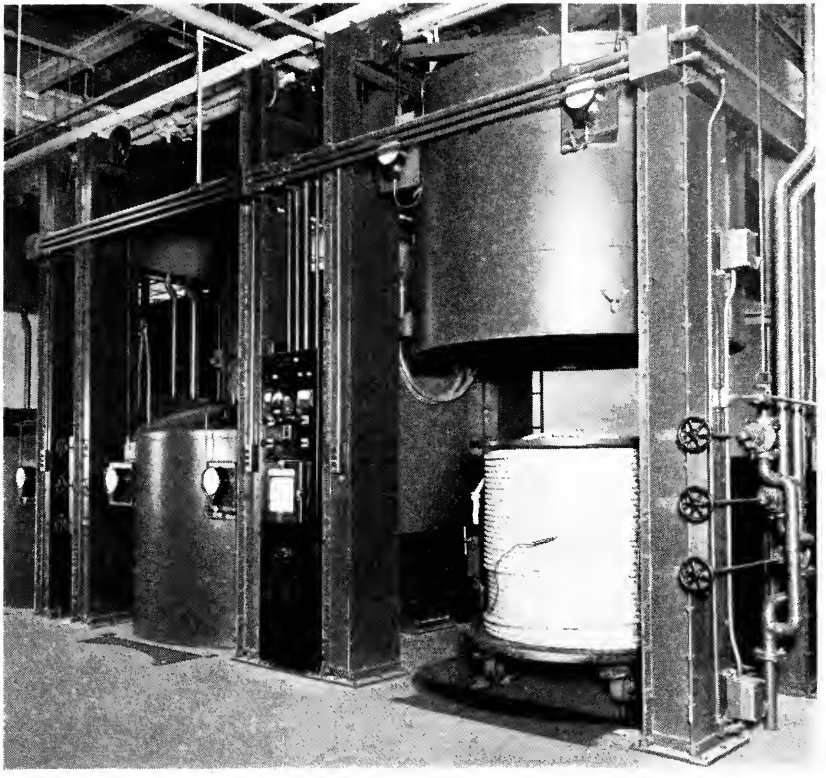


Fig. 3—Dome type dryers for holding cable and its core truck. One dome up, showing core truck in position.

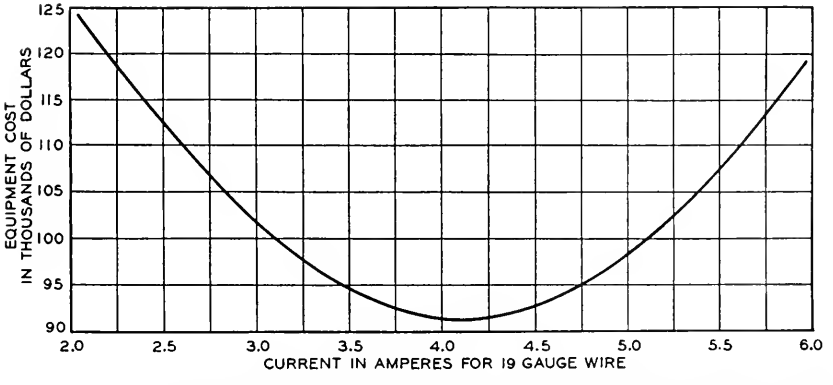


Fig. 4—Curve showing cost of equipment for drying the entire telephone cable output at Kearny Works for various starting current densities for 19-gauge wire or equivalent density for other sizes of wire.

overall efficiency of the drying operation. As shown by the curves in Fig. 4, 4 amperes per 19-gauge wire or equivalent provided about the lowest equipment cost. The heating period under this starting cur-

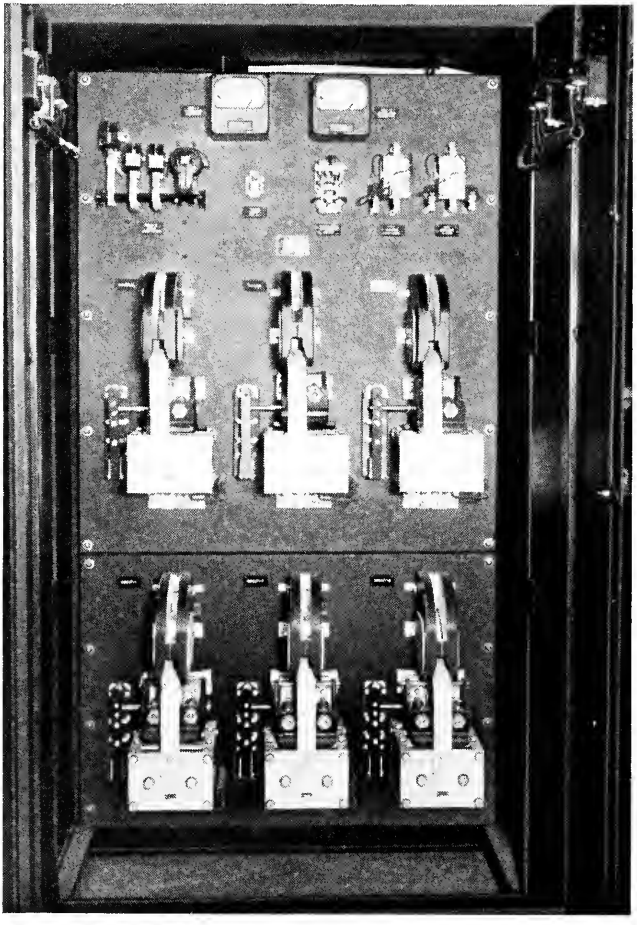


Fig. 5—D-C. power panel. The contactors in the upper row connect the generators either in parallel or in series. The contactors in the lower row connect the power supply to any of three dryers. All contactors are operated by switches on the control panel.

rent density was approximately one-half hour, which fitted very well into the plan for servicing the lead presses with dry cable cores.

In determining the type of equipment it was decided to provide two generators on the set. For the shorter lengths of large cable

where low voltage is required, the two generators are operated in parallel. For the long lengths of small cable where higher voltage is required, the generators are operated in series. Suitable switching equipment on the control board provides ready means for changing to either method of operating the generators through large contactors located on the power panel, Fig. 5. Such an arrangement permitted the maximum capacity of the unit to be reduced one-half and the average starting load to be increased to 90 per cent of full load. This not only increased operating efficiency but very greatly reduced installation costs. Three such generating units were provided for handling the entire product at Kearny, one large, one intermediate and one small.

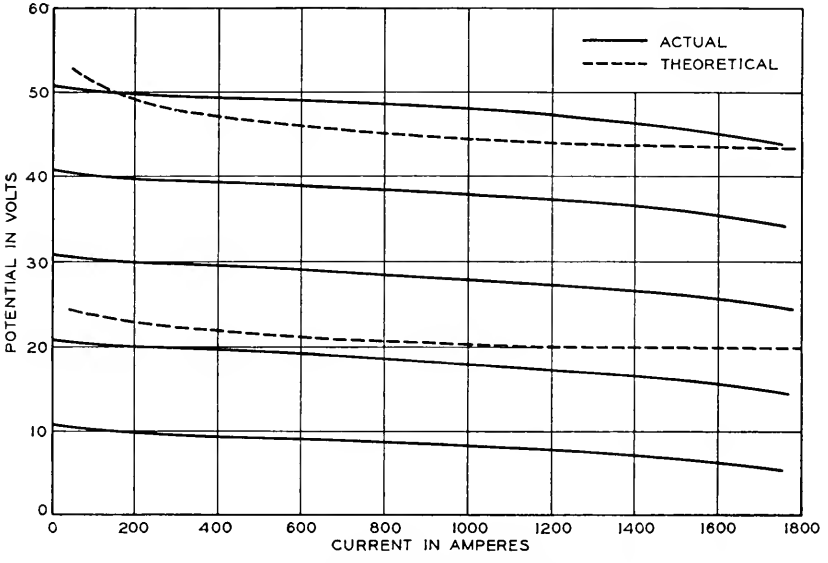


Fig. 6—Curves showing a desired theoretical voltage regulation and actual performance of the d-c. generators.

One of the chief factors contributing to simplification in operating electrical drying is to have the heating period the same for all sizes and lengths of cable. Theoretically, this requires the same voltage regulation at all conditions of the load, from the extreme of high current and low voltage to high voltage and low current. Two such theoretical curves are shown by the dotted lines in Fig. 6. The actual voltage regulation curves for one of the special generators are shown by the heavy lines on the same figure. The maximum amount of

variation in the actual voltage regulation for any two load conditions amounted to a change in heating time of approximately one minute. This small difference permits the use of one heating period for all cable as some variation in final temperature is permissible. As will be noted, the generators gave a slightly rising voltage as the current dropped, which contributed to a shortening of the heating period.

The control panel located at the dryers contains switches for starting and stopping the motor generator set, as well as indicating and control equipment for properly conducting the drying operation, Fig. 2. Included in the indicating equipment is a voltmeter and ammeter for showing the potential and current readings for the cable at any time. At the beginning of the heating cycle the operator adjusts the rheostat controlling the field amperage in the generator until the proper voltage is obtained across the cable. This reading is determined from the cable length since the voltage for the specified current density is directly proportional to the cable length. Having established the proper voltage the operator then checks the ammeter for the proper current reading which, in turn, is determined by the size and number of conductors in the cable. To simplify the work, charts are prepared which show the voltage and corresponding starting current for all lengths and sizes of cable to be dried.

As the cable increases in temperature the current falls in direct proportion to the increase in conductor resistance, which in turn is determined by the temperature coefficient of copper (voltage constant). A chart can be used, therefore, to determine the final current for the desired drying temperature. As noted above, the voltage rises slightly as the current falls and the corresponding correction is made in the chart for final current readings for the various sizes of cable.

Such a control by final current readings is all that would be needed if the set were operated manually. However, it is not only more economical to release the operator from watching the ammeter but it has proved advisable for the safety of the product to control the drying cycle automatically. This is accomplished by a time relay and a temperature controller, either of which operates at a predetermined setting to open the field circuit on the direct current generator. The opening of the field circuit operates a signal device to let the operator know that the heating cycle is over so that the next one can be started without delay. The time relay is set from actual experience for a period slightly longer than the required heating time and serves as a protection in case of failure in the heating control circuit. The temperature controller is operated from a thermocouple embedded be-

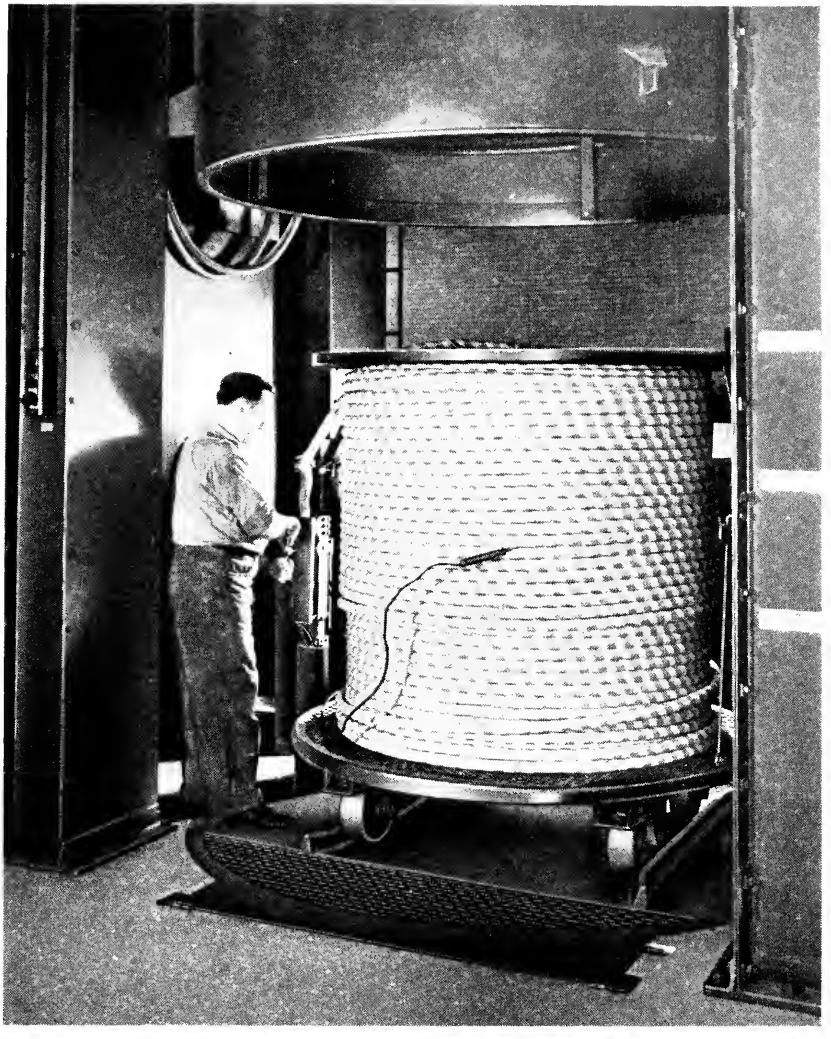


Fig. 7—Close-up of dryer with dome in raised position showing the operator tightening electrode clamp to cable. Thermocouple shown in place between layers of cable.

tween layers of the cable and is set to operate at the final desired temperature.

The drying chamber consists of a dome type shell that can be raised and lowered over a base, Figs. 3 and 7. The base is level with the floor line so that the cable core trucks can be handled readily in and

out of the dryer. The dome is fitted with a gasket to provide for sealing with the base during evacuation. Steam coils are also provided inside the dome for keeping the dryer up to temperature during operation and provide a slight amount of heat to the cable during the short electrical heating period. The bus bars, vacuum connections and thermocouple leads enter through the base. The bus bars end in large vise-like clamps for firmly holding the ends of the cable. In the units for drying the larger cables the clamps are water cooled for carrying away the heat due to contact resistance. The core truck and cable are insulated from ground in order to minimize the danger

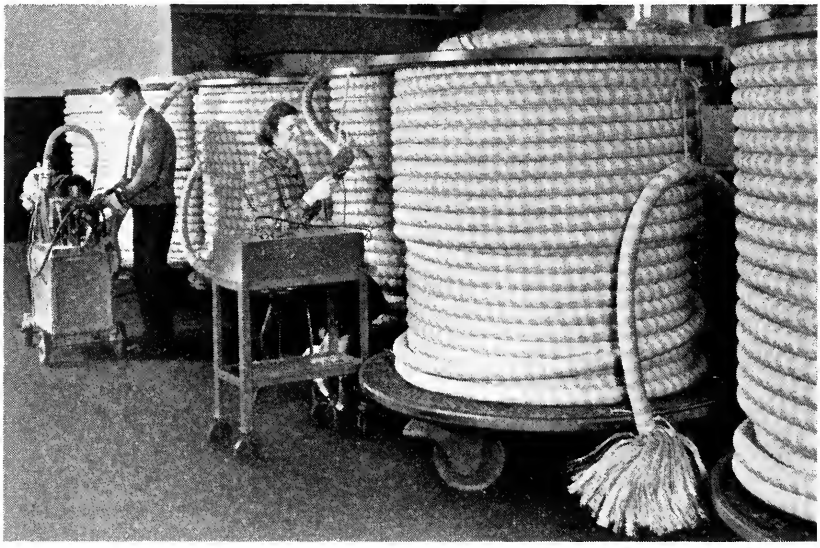


Fig. 8—Removal of paper from cable end and testing of circuits preparatory to drying.

from short circuit should the cable ground on the core truck. The push-button control for raising and lowering the domes is interlocked with the control for the heating cycles so that voltage cannot be applied at the clamp when the domes are up.

For the sizes and lengths of cables to be dried it was found advisable to have three dome dryers to one motor generator set, since all cables need some additional vacuum drying after the heating period is over. This arrangement also permits the removal of the dried cable and the loading of the next cable to be dried while the other two units are in operation. When the current is automatically disconnected at the

completion of the heating cycle in progress, the operator turns the rheostat to zero voltage and throws a three-way switch, which connects the set to the next cable to be heated.

In preparing the cable for the regular testing process the insulation is removed from one end for a distance of four or five inches, Fig. 8. After the test shows that the cable is satisfactory for lead covering, the insulation is removed from the other end. At each end the wires are bunched together and tied with cotton string so that the individual wires act together in parallel the same as one large wire. The weight of insulation is sufficiently near enough to the ratio of the wire sizes so that a cable made up of several sizes of individual conductors and their particular insulation will heat satisfactorily by applying heat in proportion to the size of wire.

The electrical drying of cables requires a total of approximately 1½ hours as compared to the process it is replacing, which requires 12 hours or more and necessitates three-shift operation of the vacuum dryers. The short period permits a more rapid turnover of process stock as well as better planning of manufacture. By coordinating the drying and lead sheathing operations cables are lead covered immediately following their removal from the drying tank. The regain of moisture is slight under this condition and therefore the expensive storage oven is unnecessary. As a consequence, the total cost of equipment for the drying operation has been reduced to one-half the former investment value and the floor space from 19,000 square feet to 9500 square feet. Another advantage follows from replacing the continuous three-shift operation by one that operates only in sequence with other operations that may be only on a one- or at most a two-shift basis.

By applying heat to each conductor in proportion to the size of wire, each cable is given an individual treatment which insures a uniformity of drying not possible in the old vacuum tank process. In the replaced system several truckloads of varying amounts of cable were placed in the same vacuum tank and all dried for the same period. It was an averaging process leading to variations in dryness of individual cables and followed of necessity from the fact that a large number of different designs and lengths must be handled each day. To approximate individual handling under such a condition, where the drying period was 12 hours or more, would have involved a large number of dryers and increased floor space and operator-time. Also in the replaced system the layers of cable in the center position on

the core truck were not fully up to the temperature of the outer layers, leading to variations in drying throughout the length. Thus, with the tendency in manufacture toward longer and longer lengths on the core truck, the variations in the degree of dryness under the vacuum tank process and the advantages of the electrical drying become more pronounced.

Electrical Wave Filters Employing Crystals with Normal and Divided Electrodes

By W. P. MASON and R. A. SYKES

I. Introduction

IN SEVERAL previous papers [1, 2, 3, 4] the application of piezoelectric crystals to electric wave filters has been discussed. The underlying principles and some of the design procedures were given. These filters have received wide application in carrier telephone systems and radio systems both in the United States and abroad.[5] It is the purpose of the present paper to discuss more completely all the standard types of filters with crystals, and methods for determining their constants and attenuation characteristics. In addition some of the newer results for simplifying such filters are given.

The use of a divided plate crystal for filters resulted in cutting the number of crystals in half as was pointed out in three former papers.[2, 3, 4] The theory of the use of such crystals is discussed in this paper and an equivalent circuit is given for a crystal with two sets of plates. The application of this circuit to unbalanced filters allows the results for balanced lattice filters to be realized for unbalanced filters. For one connection of the two plates the resonance of the crystal can be made to appear in one arm of the equivalent lattice, while for the reverse connection the resonance appears in the other arm of the lattice.

II. Crystal Filter Sections Which Can Be Realized in Lattice Networks

As pointed out in a previous paper [1] the most general filter characteristics for networks employing crystals can be realized in a lattice network, since every known form of a network can be reduced to a

[1] "Electrical Wave Filters Employing Quartz Crystals as Elements," W. P. Mason, *B. S. T. J.*, July 1934, pp. 405–452.

[2] "Resistance Compensated Band Pass Crystal Filters for Unbalanced Circuits," W. P. Mason, *B. S. T. J.*, Oct. 1937, pp. 423–436.

[3] "The Evolution of the Crystal Wave Filter," O. E. Buckley, *Jour. of Applied Physics*, Oct. 1936.

[4] "Crystal Channel Filters for the Cable Carrier Systems," C. E. Lane, *B. S. T. J.*, Vol. XVII, Jan. 1938, p. 125.

[5] "Channel Filters Employing Crystal Resonators," H. Stanesby and E. R. Broad, *P. O. E. E. Jour.*, 31, pp. 254–264, Jan. 1939.

[1] Loc. cit.

lattice network with realizable constants, whereas the converse is not necessarily true.

Let us consider first what types of filter characteristics can be obtained by using a crystal in one arm of a lattice network, and electrical or crystal elements in the other arm. As is well known the equivalent electrical network of a crystal is as shown in Fig. 1. The

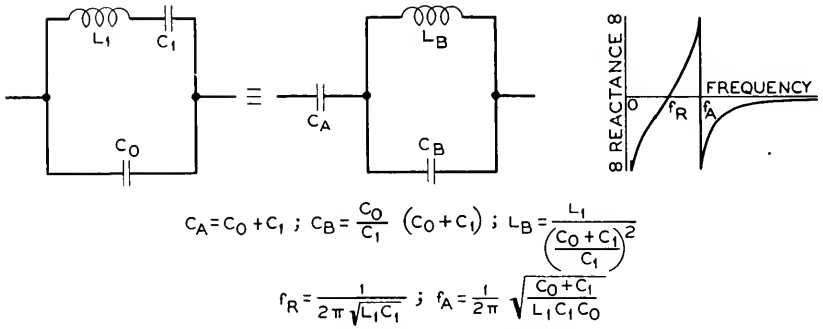


Fig. 1—Equivalent electrical circuit and reactance frequency characteristic of piezo-electric crystal.

element values, as calculated in a recent paper, for a plated crystal vibrating longitudinally are [6]

$$C_0 = \frac{Kl_w l_y}{4\pi l_t} \times \frac{1}{9 \times 10^{11}} \text{ farads}; \qquad C_2 = \frac{S}{\pi^2} \frac{d'^2_{12}}{S'_{22}} \frac{l_w l_y}{l_t} \times \frac{1}{9 \times 10^{11}} \text{ farads};$$

$$L_1 = \frac{\rho S_{22}^2 l_t l_y}{8 d'^2_{12} l_w} \times 9 \times 10^{11} \text{ henries}, \qquad (1)$$

where l_y, l_w, l_t are respectively the length, width, and thickness of the crystal expressed in centimeters, K = specific inductive capacity, S_{22}' = inverse of Young's modulus along the direction of vibration, d_{12}' is the value of the piezo-electric constant along the direction of vibration, and ρ is the density of the crystal. The resistance depends on the clamping resistance, acoustic radiation from the ends of the crystal, internal damping losses, etc. In general the ratio of the reactance of the inductance L_1 to the resistance R at the resonant frequency f_R is from 20,000 to 300,000, depending on how the crystal is mounted, whether it is evacuated, etc. In general this resistance is so small that it can be neglected for design purposes, and only the ideal reactance characteristic need be considered.

[6] "A Dynamic Measurement of the Elastic, Electric and Piezoelectric Constants of Rochelle Salt," W. P. Mason, *Phys. Rev.*, Vol. 55, April 15, 1939, p. 775.

The reactance characteristic of the crystal, as shown in Fig. 1, is a negative reactance at low frequencies up to a resonant frequency f_R. For frequencies greater than f_R, the reactance becomer positive up to the anti-resonant frequency f_A, above which the reactance is again negative. The ratio of the anti-resonant frequency to the resonant frequency is determined directly by the ratio r of C_0 to C_1 existing in the crystal. As shown by Fig. 1,

$$\frac{f_A}{f_R} = \sqrt{1 + \frac{1}{r}}. \tag{2}$$

This ratio is usually greater than 125 for a quartz crystal and hence the anti-resonant frequency is less than .4 per cent higher than the resonant frequency.

The previous papers considered mainly band-pass filters and discussed briefly low and high-pass crystal filters. It is also possible to obtain band elimination and all-pass crystal filters by combining electrical elements with the crystals in the proper manner. We consider, first, all the types of filters which can be obtained by using a single crystal in one arm of a lattice filter and electrical elements in the other arms. Figure 2 shows all the possible single-band characteristics which can be obtained by using a crystal in one arm and an electrical impedance, or crystal impedance, in the other lattice arm. For example, the first filter of the table shows a filter with a crystal in one arm and a capacitance in the other arm. Column B shows the reactance characteristic of each arm. A lattice filter will have a pass band when the reactances are of opposite sign and will attenuate when the reactances are the same sign. When the two reactances are equal the filter will have an infinite attenuation. This result follows from the expressions for the propagation constant and characteristic impedance of a balanced lattice network which are

$$\tanh \frac{P}{2} = \sqrt{\frac{Z_1}{Z_2}}; \qquad Z_0 = \sqrt{Z_1 Z_2}, \tag{3}$$

where Z_1 is the impedance of the series arm of the lattice and Z_2 that of the shunt arm. The third column shows the attenuation characteristic of this filter. It is a narrow band filter having a pass band between the resonant and anti-resonant frequencies of the filter. There is a peak of attenuation either above or below the band depending on the value of the capacitance C_1 in the lattice arm. The last column shows the value of the characteristic impedance of the filter as a function of the frequency. The dotted line indicates a

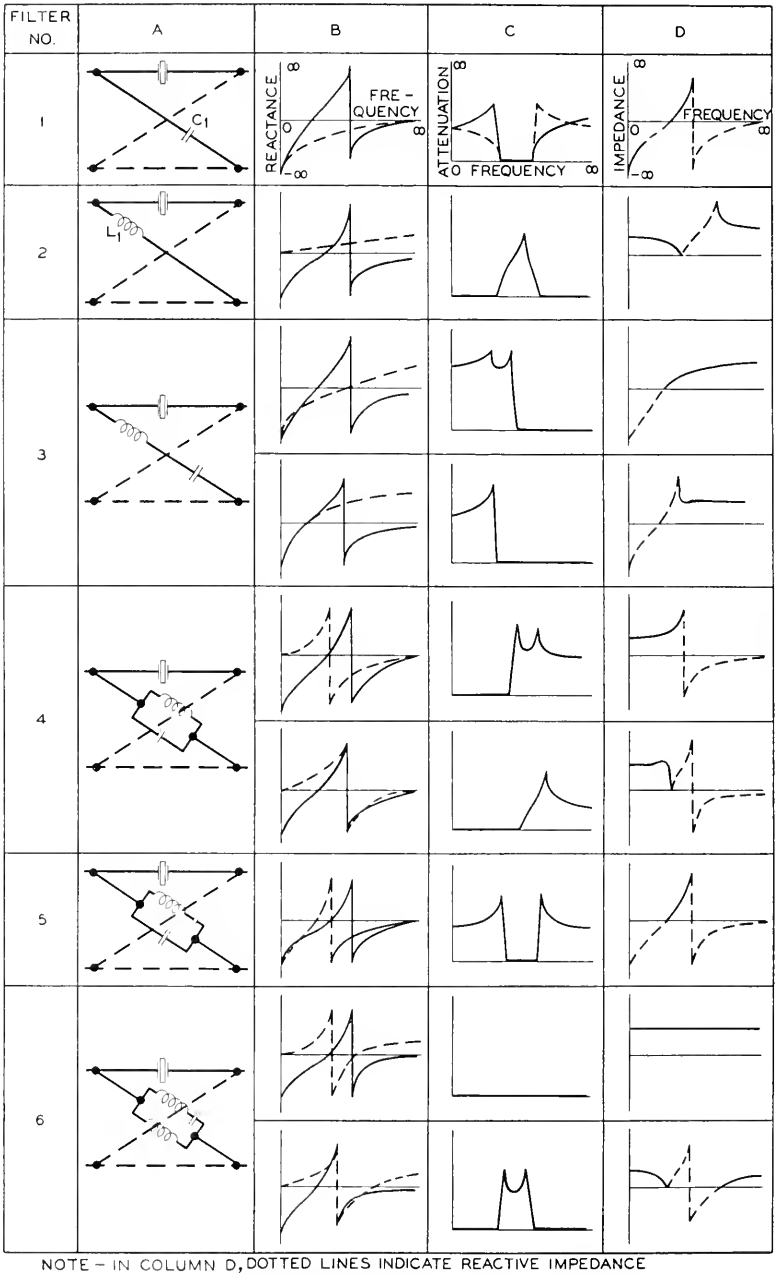


Fig. 2—Single band lattice filters employing a crystal in one arm.

reactance while a solid line indicates a resistance. In the pass band the filter has a resistive characteristic indicating a transmission of energy, while in the attenuating band the characteristic impedance is reactive indicating a reflection of energy.

Filter No. 2 shows what characteristic will be obtained if an ideal inductance is used in the lattice arm. As can be seen a band elimination filter will result with one attenuation peak. The width of the suppression band will be the separation between the resonant and anti-resonant frequencies.

The use of a series resonant circuit results in a high-pass filter as can be seen from filter No. 3. It is possible to obtain two dispositions of the resonant frequencies which will give a single pass band as shown by the two sets of curves. The first set gives a high-pass filter with two attenuation peaks and a simple characteristic impedance. The other arrangement gives one attenuation peak and a more complicated type of characteristic impedance. The theory of this balancing of characteristics obtainable with a lattice filter is well known,[7] and is useful, when it is necessary on account of reflection effects, to make the characteristic impedance constant nearly to the cutoff.

The use of an anti-resonant circuit results in a low-pass filter as shown by filter No. 4. Two characteristics are possible. Filters No. 5 and 6 show the characteristics obtainable by using series resonant circuits shunted by a capacitance or an inductance. In one case a band-pass filter with two peaks results, and in the other either a band suppression filter with two attenuation peaks or an all pass filter. It will be noted that the configurations used in the lattice arm of filter 5 is the equivalent circuit of the crystal and hence a crystal can be used in this arm. In fact the circuit is similar to one discussed in the former paper.[1]

Since the crystal positive reactance region is very narrow ($< .4\%$), all of the band pass and band elimination filters obtained by using a crystal in one arm will of necessity have very narrow band pass or band suppression regions. For high and low-pass filters the attenuation peaks will of necessity come close to the cutoff frequencies. In the all-pass structure the phase shift will be very sharp in the neighborhood of the crystal resonance. It was shown in the first paper,[1] that wider pass bands and more general characteristics can be obtained by employing inductance coils in series or parallel with the crystal. Figures 3 and 4 show the possible types of filters obtained by

[7] Cauer, Siebschattungen VDI, Verlag Berlin, 1931. H. W. Bode, "A General Theory of Electric Wave Filters," *Jour. of Math. and Physics*, Vol. XIII, pp. 275–362, Nov. 1934.

[1] Loc. cit.

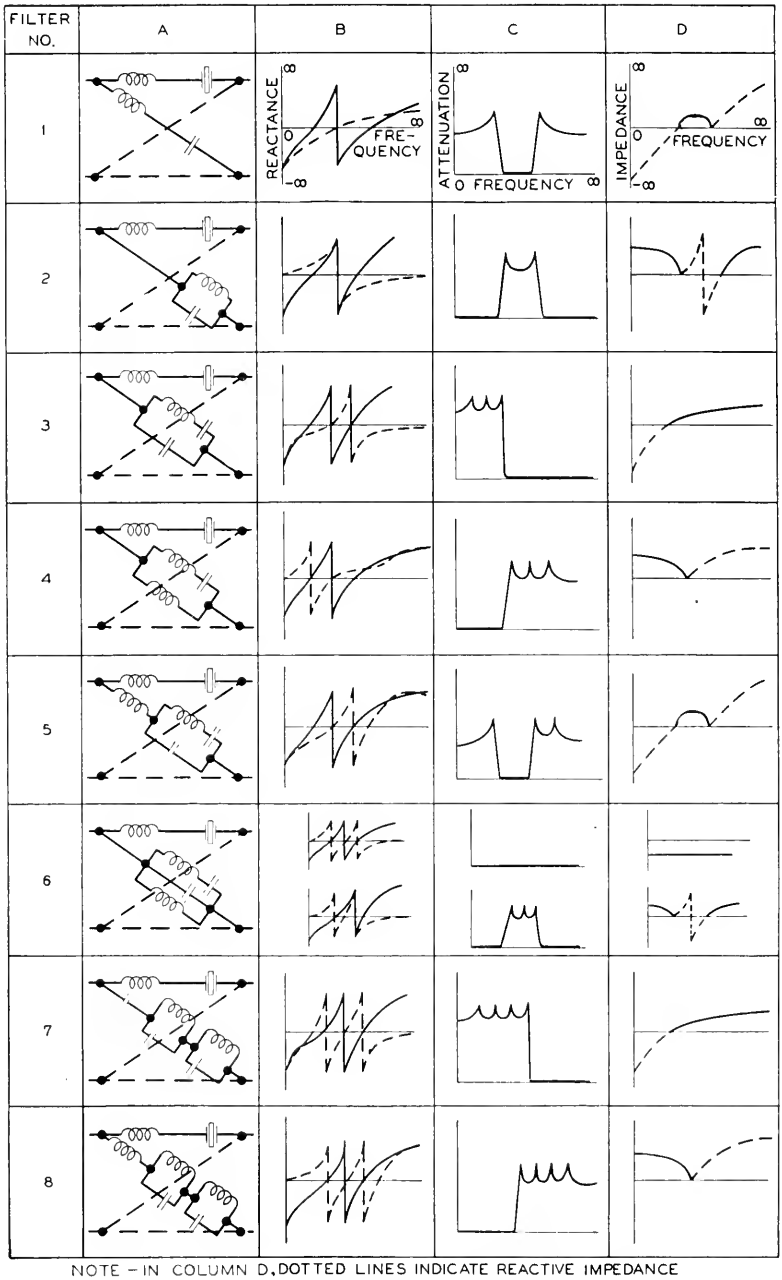


Fig. 3—Single band lattice filters employing a crystal and coil in series in one arm.

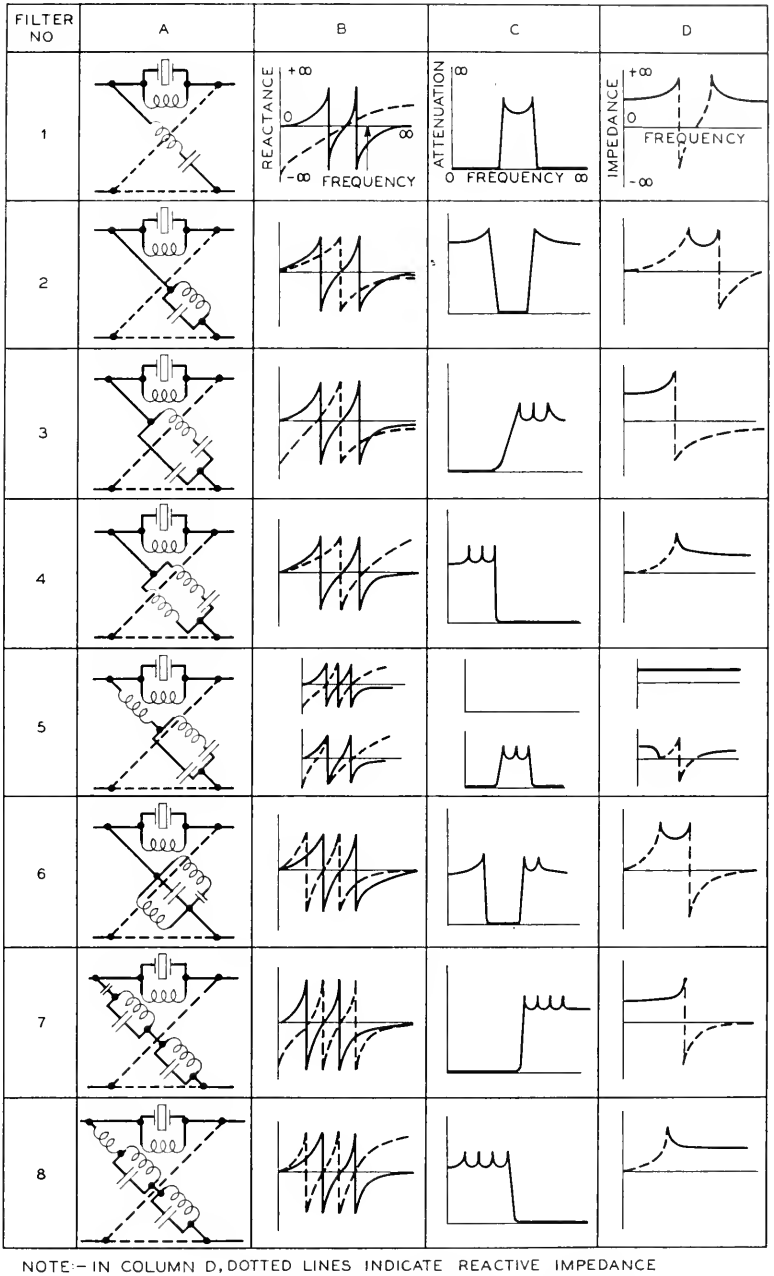


Fig. 4—Single band lattice filters employing a crystal and coil in parallel in one arm.

using a crystal and coil in one arm of the lattice and electrical elements in the other. Band-pass, band elimination, high and low-pass, and all-pass filters result. Only the simplest combinations of resonant frequencies giving the highest amount of attenuation are shown. As in filters 4 and 6 of Fig. 2, some of the anti-resonances and resonances of the two arms may be made to coincide giving filters with less attenuation but more flexibility in the impedance characteristics. Condensers can be incorporated in parallel or series with the crystals without affecting the type of characteristic obtained. This procedure is useful in controlling the widths of the pass or attenuation bands and in controlling the position of the peak values. In a number of the filters of Figs. 3 and 4, the equivalent circuit of the crystal occurs in the electrical circuit in the lattice arms. In these filters, crystals can also be used in the lattice arms. Filters 1 and 5 of Fig. 3 and filters 2 and 6 of Fig. 4 are band-pass filters which have been discussed in detail in former papers.[1, 2]

Crystals may also be used in more complicated electrical circuits, for example with transformers as shown in Fig. 5. This figure shows high, low, band-pass, band elimination and all-pass filters which can be constructed by employing transformers and crystals in each arm. More complicated structures still using single crystals can also be constructed but they tend to be of less importance since the dissipation introduced by the electrical elements neutralizes any benefit of using crystals.

It is possible, however, to use more crystals than one in one arm of a lattice and obtain filters having higher insertion losses outside the band without introducing more loss due to the electrical elements in the band. Figure 6 shows a number of such combinations with and without coils. The result of adding an additional crystal in one arm of a lattice is to add another elementary section of the type discussed in Appendix I. An example [8] of the characteristic obtainable with a band-pass filter with two crystals in each arm is shown on Fig. 7.

All of the filters discussed above were assumed to be constructed from dissipationless elements. When coils are used, however, a certain amount of resistance is associated with them which may alter the characteristics obtainable. As has been pointed out previously,[1] if the dissipation associated with the coils can be brought out to the ends of the arm either in series or parallel with the complete arm the effect of these resistances will be to add a constant loss independent of

[1, 2] Loc. cit.
[8] This filter was constructed and tested by Mr. H. J. McSkimin.
[1] Loc. cit.

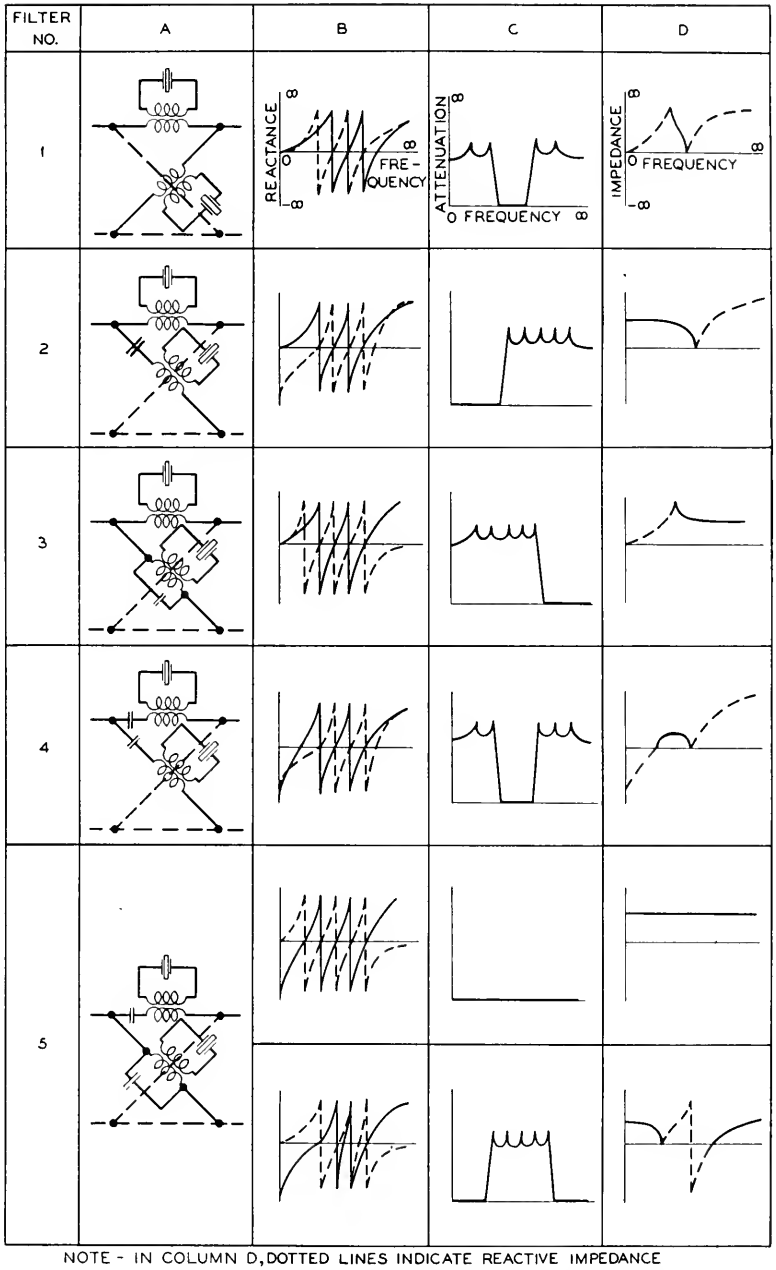


Fig. 5—Single band lattice filters employing transformers and crystals.

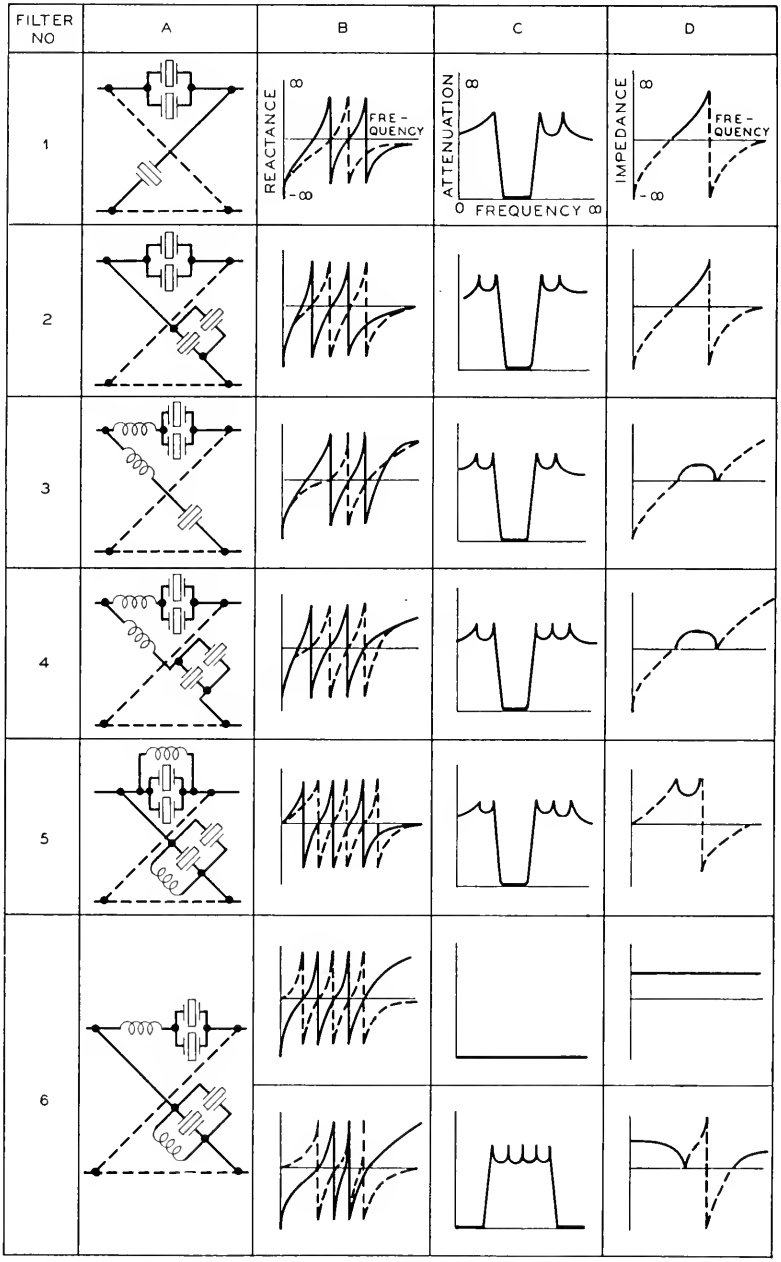


Fig. 6—Single band lattice filters employing more than one crystal in the impedance arms.

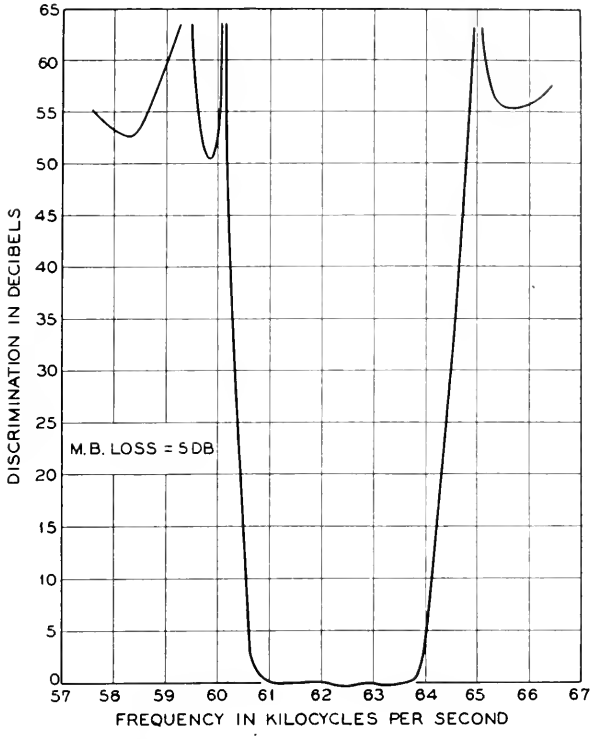


Fig. 7—Insertion loss characteristic of a single section band pass filter employing two crystals in each arm.

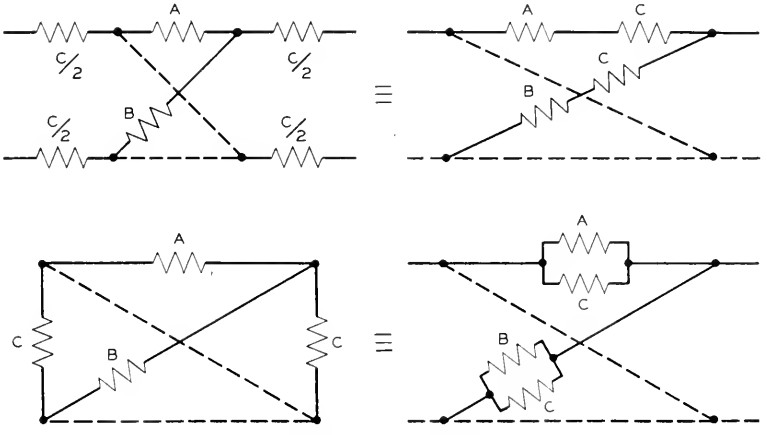


Fig. 8—Equivalences between lattice networks.

the frequency, and hence the discrimination obtainable will not be affected. This follows from the network equivalences shown in Fig. 8 which were first proved in a previous paper.[1] Even if this resistance compensation cannot be completely obtained it can often be obtained over a limited range near the cutoff and the peak frequencies by adding resistances to some of the crystal or electrical elements of such a value that the resistive components of the two arms are nearly equal over a limited frequency range. This results in cutting down the distortion near the cutoff and increasing the loss in the attenuated regions.

The lattice filters of Figs. 2 to 6 can be realized in ladder or bridge T forms in certain cases. If the two arms have two common series elements, then by the first equivalence of Fig. 8 they can be taken outside the lattice. Similarly, if two common shunt elements can be found in the two arms, then, by the second theorem of Fig. 8, the

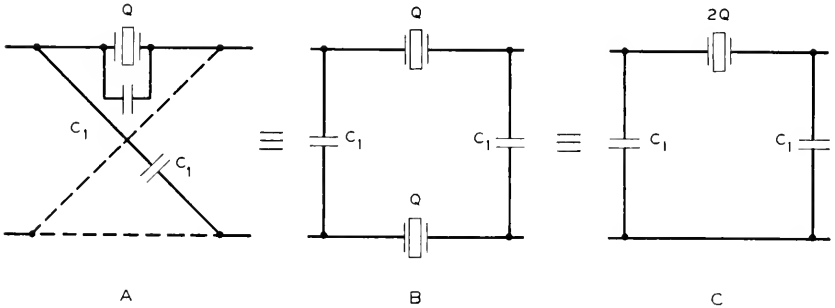


Fig. 9—Method for reducing a lattice filter to a π network filter.

elements can be placed in shunt on the ends of the filter. For example, suppose that we consider filter No. 1 of Fig. 2 and shunt the crystal by a capacitance C_1 which is equal to the capacitance of the lattice arm as shown in Fig. 9A. Then the two capacitances can be taken out in shunt leaving a crystal in the series arm of a π network as shown in Fig. 9C. This has the same type of characteristic as the lattice but considerably greater limitations.

A somewhat more general transformation can be made to a bridge T network of the type shown in Fig. 10A. This network is equivalent to the lattice network shown in Fig. 10B. As is evident, if we have an impedance in parallel with one arm and in series with the other, the resulting lattice can be transformed into a bridge T network. For example, in Fig. 3, filter No. 2, if we reverse the lattice and series arms, which can be done without changing the characteristics except for a 180° phase reversal, the filter can easily be reduced to a bridge

T network as shown in Fig. 11. The shunt coil L_0 is usually considerably smaller than the series coil L_1 so that L_1 can be divided into two coils, L_0 and $L_1 - L_0$. The transformation then becomes as

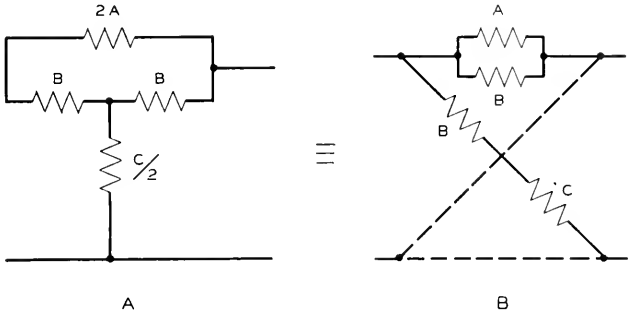


Fig. 10—Equivalence between bridge T and lattice networks.

shown in Fig. 11 with the element values shown. $Q/2$ indicates that the impedance of the crystal in the shunt arm is half that in the lattice arm. This transformation is applicable particularly to low and high-pass filters and band elimination filters.

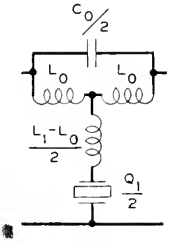


Fig. 11—A bridge T band elimination crystal filter.

Another transformation which can be employed is that for a three-winding transformer, for, as shown in a previous paper,[2] a three-winding transformer connected to two impedance arms as shown in Fig. 12 is equivalent to a transformer and a lattice filter with small

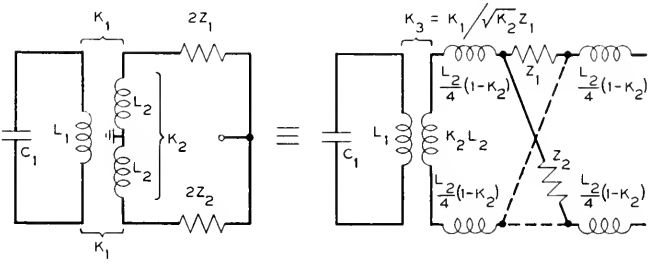


Fig. 12—Lattice equivalent of a three-winding transformer.

[2] Loc. cit.

coils on the ends. By making the coupling in the secondary high these coils can be made very small and can usually be neglected.

Another method for reducing balanced lattice filters to unbalanced circuits is to employ crystals with two sets of plates as described in section IV.

III. Method for Calculating the Element Values of the Filter

The curves in Figs. 2 to 6 give a qualitative picture of what type of characteristics can be obtained by the use of crystals in filter networks. In order to determine what band widths and dispositions of attenuation peaks are realizable with crystals it is necessary to calculate the element values, since a crystal cannot be made with a ratio of capacitances under 125.

The actual process of calculation can be divided into two parts. The first part consists in a determination of the critical frequencies of the arms of the network in terms of the desired attenuation characteristic. The second part consists in calculating the element values from the critical frequencies by means of Foster's theorem.

The attenuation characteristics obtainable with filters are discussed in Appendix I, and it is there shown that the attenuation characteristic of a complicated filter structure can be regarded as the sum of the attenuation characteristics of a number of elementary filters. The critical resonant frequencies of the filter are evaluated in terms of the cutoff frequencies and the position of the attenuation peaks with respect to the cutoff frequencies. The ratios of the impedances of the two arms at zero or infinite frequencies are evaluated in terms of the network parameters. With the aid of these equations, and Foster's theorem discussed below, the element values can be evaluated for any desired attenuation characteristic. Whether the characteristic is realizable or not depends on whether the element values of the equivalent circuit of the crystal calculated have a low enough ratio of capacitances to be realized in practice. The actual value of the series capacitance C_1 of the equivalent circuit of the crystal shown in Fig. 1 may also be too large to be physically realizable.

Having obtained the critical frequencies by the calculations given in the appendix, the element values can be calculated by using Foster's theorem. Foster's theorem [9] deals with impedances in the form of a number of series resonant circuits in parallel as shown on Fig. 13A or a number of antiresonant circuits in series as shown on Fig. 13B.

[9] See "A Reactance Theorem," *B. S. T. J.*, April 1924, page 259.

In either case the impedance of the network can be written in the form

$$Z = -jH\left[\frac{\left(1-\frac{\omega^2}{\omega_1^2}\right)\left(1-\frac{\omega^2}{\omega_3^2}\right)x\cdots x\left(1-\frac{\omega^2}{\omega_{2n-1}^2}\right)}{\omega\left(1-\frac{\omega^2}{\omega_2^2}\right)x\cdots x\left(1-\frac{\omega^2}{\omega_{2n-2}^2}\right)}\right], \qquad (4)$$

where $H \geq 0$ and $0 = \omega_0 \leq \omega_1 \leq \cdots \leq \omega_{2n-1} \leq \omega_{2n} = \infty$. For the series resonant circuits of Fig. 13A, the element values are given by

$$L_i = \frac{1}{C_i\omega_i^2} = \left(\lim_{\omega \to \omega_i}\right)\left(\frac{j\omega Z}{\omega_i^2 - \omega^2}\right); \qquad i = 1, 3, \cdots 2n-1. \qquad (5)$$

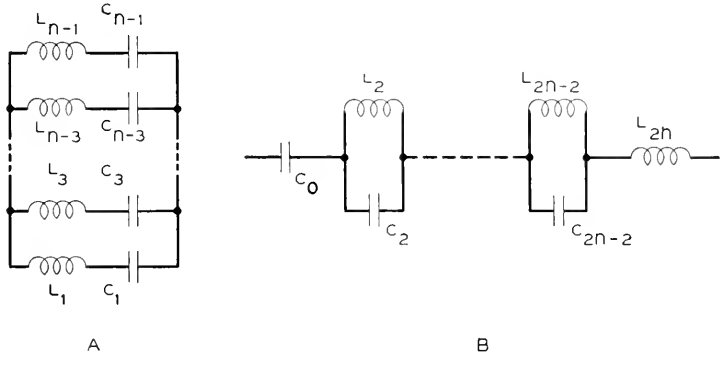


Fig. 13—Impedances arranged in form for application of Foster's Theorem.

For the antiresonant circuits in series the values become

$$C_i = \frac{1}{L_i\omega_i^2} = \left(\lim_{\omega \to \omega_i}\right)\left(\frac{j\omega}{Z(\omega_i^2 - \omega^2)}\right); \qquad i = 0, 2, 4, \cdots 2n. \qquad (6)$$

These values include the limiting values for the series case of Fig. 13B.

$$C_0 = \frac{1}{H}; \qquad L_0 = \infty, \quad C_{2n} = 0; \qquad L_{2n} = \frac{H(\omega_2^2\omega_4^2 \cdots \omega_{2n-2}^2)}{(\omega_1^2\omega_3^2 \cdots \omega_{2n-1}^2)}. \qquad (7)$$

Hence if the elements of one arm of the lattice are arranged in either of the forms shown on Figs. 13A or B, the element values can be calculated from equations (5) and (6).

If they are not in this form, they can be transformed into one of these two forms by well known network transformations. For example, all the filters of Figs. 2, 3 and 4 are of this form or can be put in this form by employing the simple network transformation of Fig. 1. For the two crystal sections shown on Fig. 6, the series

inductance can be evaluated by equation (7) and subtracted from the impedance Z. This leaves an impedance

$$Z' = Z - j\omega L, \tag{8}$$

from which can be evaluated the constants of the two crystals in parallel by employing equation (6). In this way the constants of any filter can be evaluated when the desired attenuation and impedance characteristic are specified. Several of the band-pass filters are discussed in detail in a former paper.[2]

IV. Application of Divided Plate Crystals to Balanced and Unbalanced Filters

The use of a divided plate crystal to cut the number of crystals in half in a balanced lattice filter has been mentioned previously.[2, 3, 4, 10] The theory of this use has not been previously discussed and since it results in further applications it seems worth while to present it here.

In order to use the divided plate crystal in filters it is necessary to find an equivalent circuit for such a crystal which will hold for measurements between any pair of the four terminals. It was shown in a previous paper [6] that an electro-mechanical equivalent of a fully plated crystal free to vibrate on both ends could be represented as shown in Fig. 14A. In this figure the capacitance C_0 is the static capacitance of the crystal, the condenser C_M represents the effective compliance of the crystal at the resonant frequency, and the inductance L_M represents the effective mass. A perfect transformer of impedance ratio 1 to φ^2, where

$$\varphi^2 = \frac{(d_{12}l_w)^2}{(s_{22'})} \tag{9}$$

represents the coupling from electrical to mechanical energy. φ in effect is the ratio of the force exerted by the crystal when it is clamped, to the applied voltage or it is the force factor of the system. If now we use only half the plating on the crystal, for example the plates 1, 2 of Fig. 14B, the same representation will hold. The static capacitance C_0 will be divided by 2, and the force applied by a given voltage will also be divided by 2 or the transformer ratio will be $\varphi/2$. The same compliance and mass will be operative. Hence the equivalent circuit of a crystal with plates covering half the crystal will be as shown on Fig. 14C. For a crystal with two sets of plates, the repre-

[2, 3, 4] Loc. cit.
[10] See patent 2,094,044, W. P. Mason, issued Sept. 28, 1937.
[6] Loc. cit.

sentation shown on Fig. 14D can be used if we are interested only in the transmission from one set of plates to the other. The numbering on the terminals agrees with that shown on Fig. 14B and is necessary in order that a given voltage E will produce the same displacement in the crystal when the voltage is applied between 1 and 2 and 3 and 4.

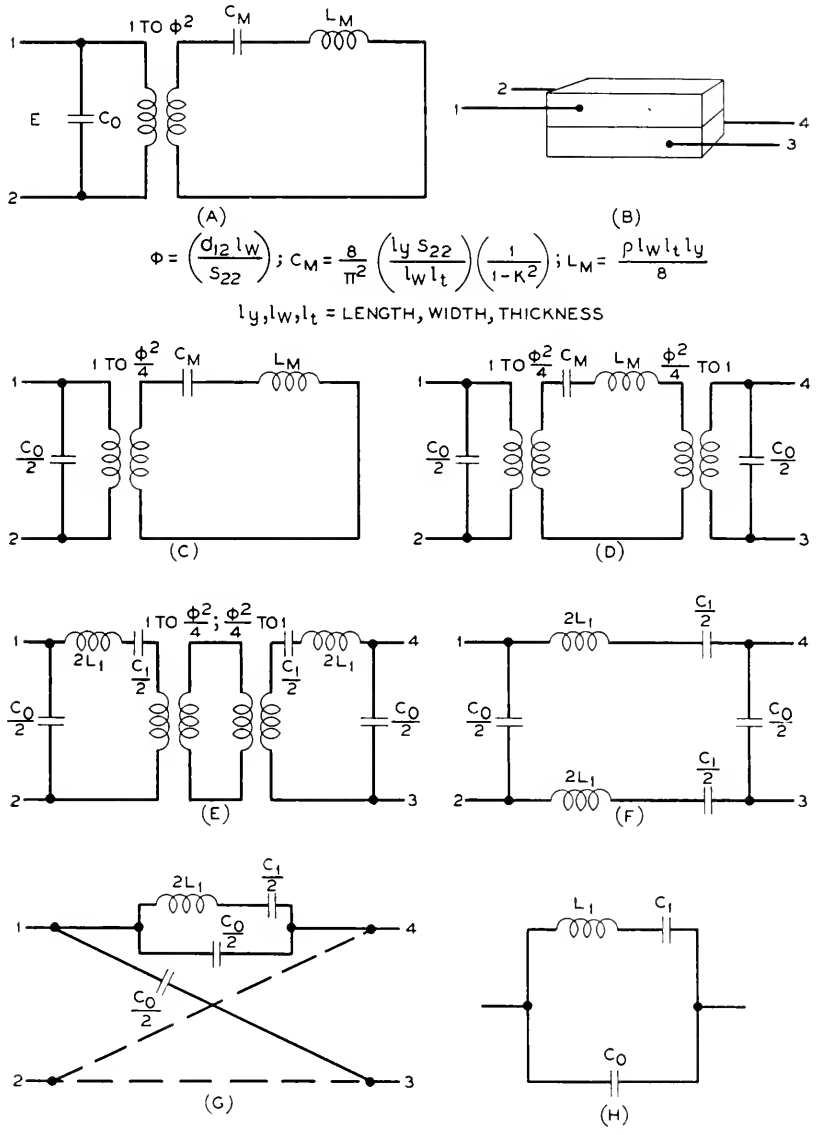


Fig. 14—Equivalent networks to represent transmission through divided plate crystal.

For purely electrical measurement, we can get rid of the two ideal transformers by taking half the impedance of C_M and L_M through each transformer as shown in Fig. 14E. Since we have left two opposing transformers of equal ratio they can be eliminated and the network of Fig. 14F results. This is shown in balanced form. Figure 14G shows the same network expressed in lattice form which is easily done by using the equivalences of Fig. 8. This represents a crystal of twice the impedance of the fully plated crystal in each series arm of the lattice with the static capacitances C_0 in the other arm. If we connect terminal 1 to 3 and 4 to 2, or in other words we use a completely plated crystal, the equivalent circuit reduces to that for a fully plated crystal as shown in Fig. 14H.

The networks of Fig. 14, F and G, represent the two plate crystals for transmission through the crystal, but do not give a four-terminal equivalence. For example, if we measure the crystal between terminals 1 and 3 we should not expect any impedance due to the vibration of the crystal since there is no field applied perpendicular to the thickness. The representation of Fig. 14, F or G, would not indicate this. The same sort of problem arises when it is desirable to obtain a four-terminal representation of a transformer and can be solved by using a lattice network representation with positive and negative inductance elements. The same procedure can be employed for a crystal and the steps are shown in Fig. 15.

We start with the lattice representation of Fig. 14G but employ the series form of the impedance of a crystal shown in Fig. 1. The series capacitance is divided into two parts, $C_0/2$ and a negative capacitance necessary to make the total series capacitance equal to C_0 plus C_1. This negative capacitance and the antiresonant circuit are lumped as one impedance $2Z$ in Fig. 15B. Now by the network equivalence of Fig. 8, we can take the series capacitances $C_0/2$ outside the network. We can also add an impedance $Z/2$ on the ends of the network provided we add a negative Z in series with all arms of the network as shown in Fig. 15C. The network of Fig. 15C is equivalent to that of Fig. 15A as far as transmission through it is concerned, but is different if we measure impedances between any of the four terminals. For example, if we measure the impedance between the terminals 1 and 4, the impedance of the network reduces to that shown in Fig. 15D. The impedance of the parallel circuit reduces to a plus Z shunted by a minus Z which introduces an infinite impedance. Similarly between 1 and 4, 2 and 3, and 2 and 4 the impedance becomes infinite as it should be if we neglect the small static capacitances existing in the crystal. If we take account of these the complete

four-terminal representation of a crystal becomes that shown in Fig. 15E. Ordinarily the capacitances C_{13}, C_{14}, C_{23}, C_{24} are small enough to be neglected. Figure 15E then is a complete equivalent circuit for a two-plate piezo-electric crystal which is valid for any kind of impedance or transmission measurements.

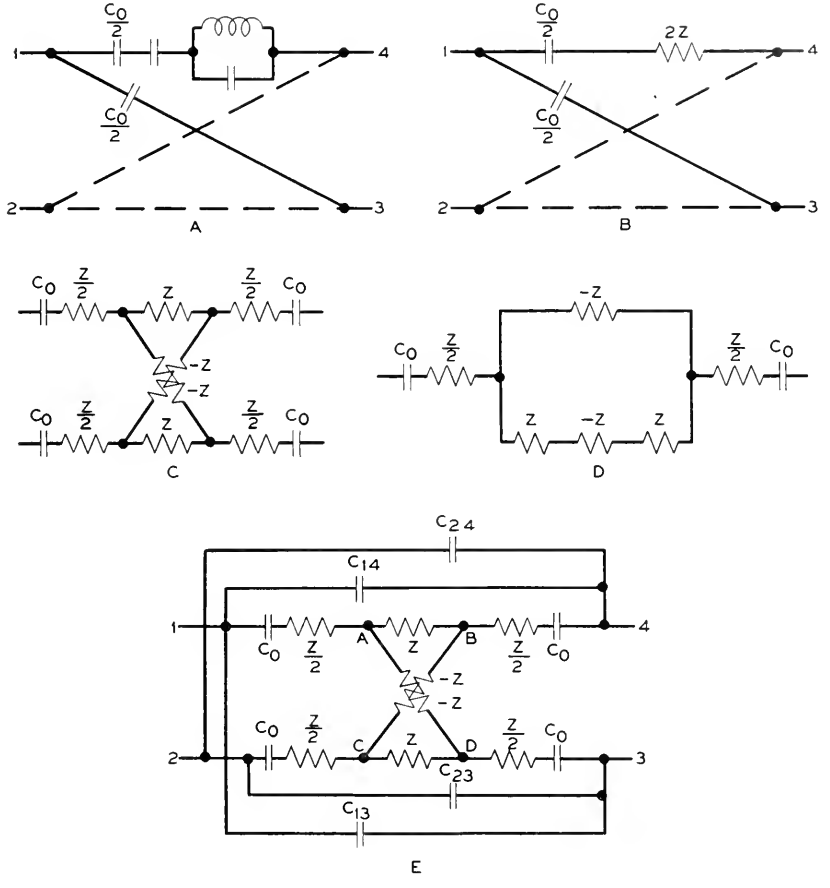


Fig. 15—Four-terminal equivalent network for divided plate crystal.

There are four possible connections for a crystal with two sets of plates used in a balanced filter. These connections and their equivalent circuits for transmission through as used in the filter are shown in Fig. 16. In order to prove these equivalences let us consider the equivalence shown on Fig. 16A. The four-terminal network representation for this case is shown in Fig. 17, which is obtained from

Fig. 15E. The capacitances C_{13} and C_{24} will be equal due to the symmetry in the crystal, while C_{14} will equal C_{23} for the same reason. These capacitances are directly connected to the outside terminals

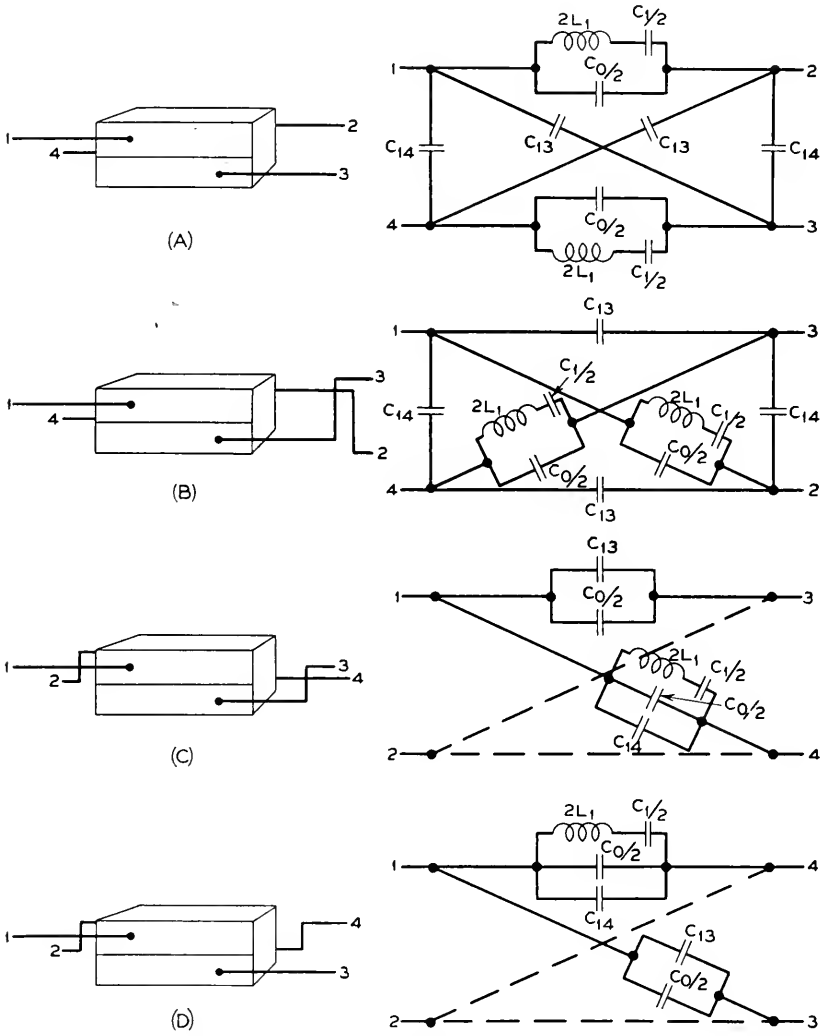


Fig. 16—Balanced divided plate crystal connections and their equivalent lattice electrical circuits.

and hence in obtaining the equivalent lattice they can be connected in directly. The remainder of the circuit can be reduced to its equivalent lattice by employing the equivalence shown in Fig. 8. Taking in the parallel impedance Z and the series impedances $Z/2 + (-j/\omega C_0)$, the

network of Fig. 17B results. On account of the paralleling of the $-Z$ and $+Z$ the lattice arm vanishes and the network reduces to that shown in Fig. 16A. In a similar manner the other equivalences result.

The use of divided plating crystals to obtain wide band filters by using series coils to widen the band is obvious. If we connect two

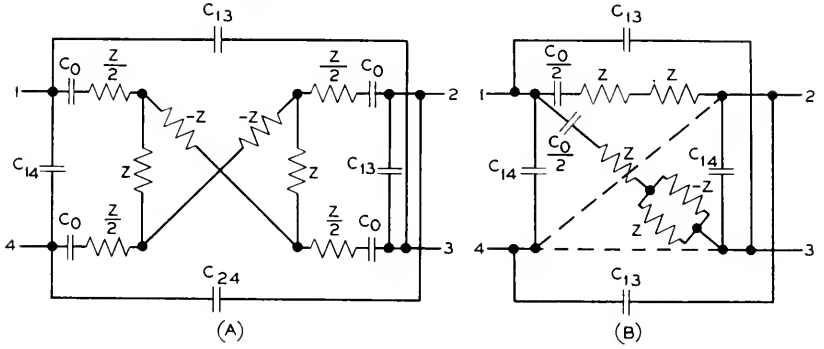


Fig. 17—Method for proving equivalence of Fig. 16A.

crystals as shown in Fig. 18A, one crystal being connected as shown in Fig. 16A and the other in Fig. 16B, a lattice filter equivalent to that is shown in Fig. 18B. In the series arms we have a crystal of twice the impedance of the fully plated crystal Q_1 shunted by the

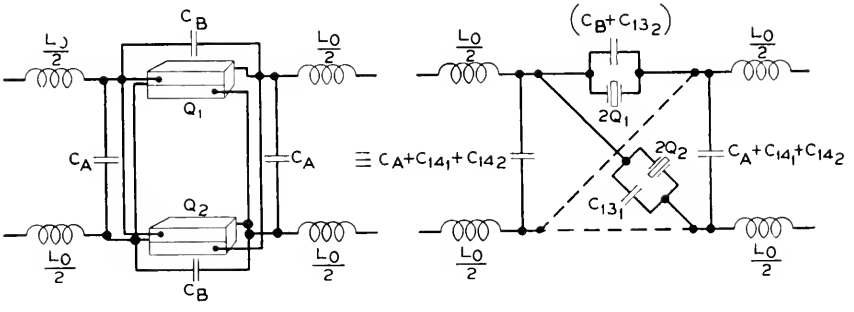


Fig. 18—Band pass crystal filter employing connections of Figs. 16A and B.

capacitance C_B and the capacitance C_{13} of the second crystal Q_2. In the lattice arms we have a crystal of twice the impedance of the fully plated crystal Q_2 in parallel with the capacitance C_{13} of the crystal Q_1. On the ends of the lattice we have capacitances $C_A + C_{14_1} + C_{14_2}$. It is obvious, then, that by using divided plate crystals we can replace two identical crystals in the two arms with crystals having twice the impedance of the fully plated crystals. This result can be

utilized in any type of filter where two crystals occur in series or lattice arms of a balanced lattice filter.

The connections *C* and *D* of Fig. 16 can also be used to give wide band filters but on account of the extra capacitance shunting each crystal, as wide bands cannot be obtained. This is shown in Fig. 19 which shows two crystals connected as shown in Figs. 16C and D used in a filter. Since each crystal is shunted by half the static capacitance of the other, the ratio of capacitances will be about twice that in the connection shown in Fig. 18 and the band width possible will be about 70 per cent of that shown in Fig. 18. Hence the connections of Fig. 18 are usually desirable.

The connections of 16C and D can be duplicated in unbalanced form as shown in Fig. 20. These equivalents are easily worked out from the network of Fig. 15E by employing Bartlett's theorem. An

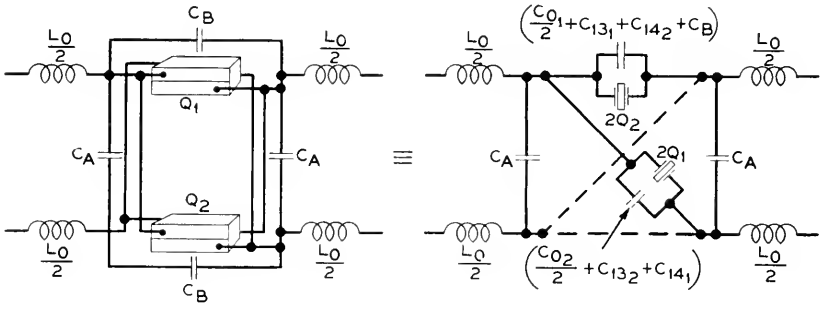


Fig. 19—Band pass crystal filter employing connections of Figs. 16C and D.

unbalanced filter of the type shown in Fig. 19 can be obtained by combining the two connections shown in Fig. 20 as shown in Fig. 21. It will be noted that across the series arm we have a capacitance $C_B + 2C_{14_1} + 2C_{13_2} + C_{0_2}/2$ while the lattice arm has only the capacitance $C_{0_1}/2$. To get attenuation peaks which are separated from the pass band by a large frequency range it is necessary to keep the capacitances C_{14_1} and C_{13_2} small. This can be accomplished by using shielding strips on the plating as shown in Fig. 22 for the two types of connection. In the *B* connection, the grounding strip is effectively obtained by making the grounded plates 2 and 3 slightly larger than 1 and 4. These grounding strips then act like a guard ring and reduce the stray capacitances.

The same process can be applied to any of the filters of Figs. 2 to 6 to obtain in unbalanced form the characteristics obtained in lattice form. In general the characteristics are somewhat more limited since

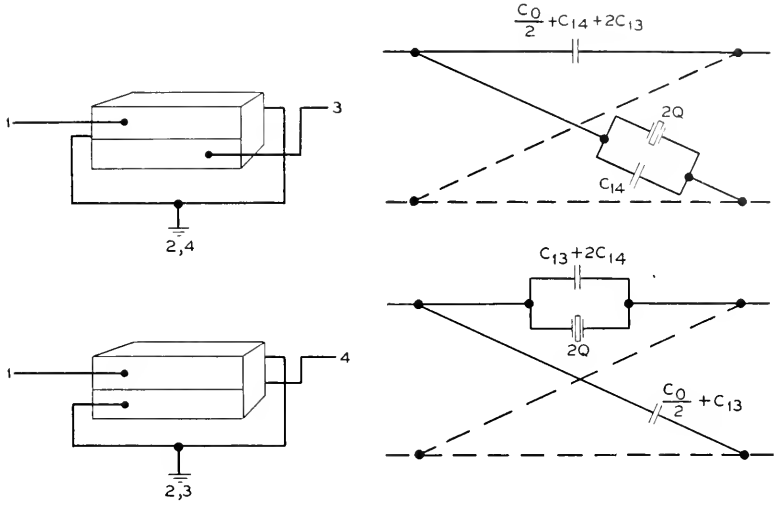


Fig. 20—Unbalanced divided plate crystal connections and their equivalent electrical circuits.

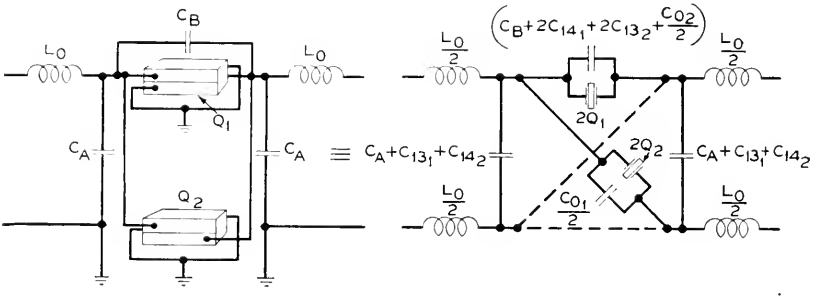


Fig. 21—Unbalanced band pass filter employing the connections of Fig. 20.

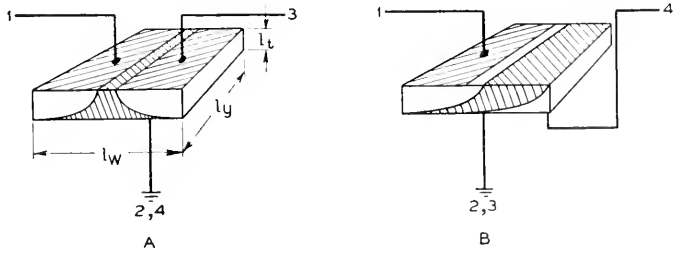


Fig. 22—Method for reducing stray capacitances in unbalanced filter connections.

in effect we have to use crystals with twice the ratio of capacitances than can be used in the balanced case.

APPENDIX

A Determination of the Resonant Frequencies of Lattice Filters

In order to obtain the element values of the filters shown in this paper it is necessary to determine the resonant frequencies of the elements in terms of the desired characteristics of the filter. It is the purpose of this appendix to show how these resonant frequencies may be derived.

The simplest type of band-pass filter section—referred to as the elementary section—is one in which there is one resonance in each arm of a lattice filter as shown in Fig. 23A. The impedance of the

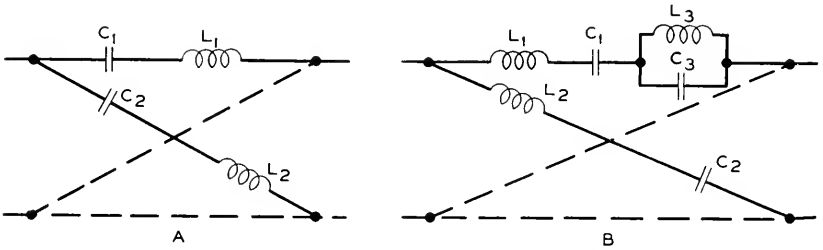


Fig. 23—Lattice filter configuration for elementary band pass sections.

series and lattice arms takes the form

$$Z_1 = \frac{-j}{\omega C_1}\left[1 - \frac{\omega^2}{\omega_A^2}\right]; \qquad Z_2 = \frac{-j}{\omega C_2}\left[1 - \frac{\omega^2}{\omega_B^2}\right], \tag{10}$$

where ω is 2π times the frequency f, f_A the resonant frequency of the series arm which also is the lower cutoff of the filter, and f_B the resonant frequency of the lattice arm which is also the upper cutoff.

The characteristic impedance and propagation constant are from equation (3)

$$Z_0 = \sqrt{Z_1 Z_2} = \sqrt{-\frac{1}{\omega^2 C_1 C_2}\left[1 - \frac{\omega^2}{\omega_A^2}\right]\left(1 - \frac{\omega^2}{\omega_B^2}\right)},$$

$$\tanh \frac{P}{2} = \sqrt{\frac{Z_1}{Z_2}} = \sqrt{\frac{C_2}{C_1}\left(\frac{1 - \omega^2/\omega_A^2}{1 - \omega^2/\omega_B^2}\right)} = m\sqrt{\frac{1 - \omega^2/\omega_A^2}{1 - \omega^2/\omega_B^2}}. \tag{11}$$

It is desirable to correlate the value of m with the frequency of infinite attenuation in the filter. Since the filter will have an infinite attenua-

tion when $\tanh P/2 = 1$, we have

$$m = \sqrt{\frac{1 - \omega_\infty^2/\omega_B^2}{1 - \omega_\infty^2/\omega_A^2}}, \tag{12}$$

where ω_∞ is $2\pi f_\infty$, where f_∞ is the frequency of infinite attenuation. For a single section since $m = \sqrt{C_2/C_1}$, m must be real and lie between 0 and infinity. The possible attenuation characteristics obtainable with the simple section can be calculated from equation (11). It will be noted that when $\omega = 0$

$$m = \tanh \frac{P_0}{2}. \tag{13}$$

Similar equations for low-pass, high-pass and all-pass filters can be derived from these equations by letting f_A go to zero or f_B to ∞, or both. These equations are:

For Low-Pass Filters

$$\tanh \frac{P}{2} = \lim \omega_A \to 0 \left(\sqrt{\frac{1 - \omega_\infty^2/\omega_B^2}{1 - \omega_\infty^2/\omega_A^2}} \sqrt{\frac{1 - \omega^2/\omega_A^2}{1 - \omega^2/\omega_B^2}} \right)$$

$$= \sqrt{1 - \omega_B^2/\omega_\infty^2} \sqrt{\frac{-\omega^2/\omega_B^2}{1 - \omega^2/\omega_B^2}} = m_l \sqrt{\frac{-\omega^2/\omega_B^2}{1 - \omega^2/\omega_B^2}}. \tag{14}$$

For High-Pass Filters

$$\tanh \frac{P}{2} = \lim \omega_B \to \infty \left(\sqrt{\frac{1 - \omega_\infty^2/\omega_B^2}{1 - \omega_\infty^2/\omega_A^2}} \sqrt{\frac{1 - \omega^2/\omega_A^2}{1 - \omega^2/\omega_B^2}} \right)$$

$$= \sqrt{\frac{1}{1 - \omega_\infty^2/\omega_A^2}} \sqrt{1 - \omega^2/\omega_A^2} = m_h \sqrt{1 - \omega^2/\omega_A^2}. \tag{15}$$

For All-Pass Filters

$$\tanh \frac{P}{2} = \begin{matrix} \lim \omega_A \to 0 \\ \lim \omega_B \to \infty \end{matrix} \left(\sqrt{\frac{1 - \omega_\infty^2/\omega_B^2}{1 - \omega_\infty^2/\omega_A^2}} \sqrt{\frac{1 - \omega^2/\omega_A^2}{1 - \omega^2/\omega_B^2}} \right)$$

$$= \sqrt{\frac{1}{-\omega_\infty^2}} \sqrt{-\omega^2}. \tag{16}$$

For this case, since there is no peak in the real frequency range, we must let ω_∞ be imaginary or $i\omega_a$. Then

$$\tanh \frac{P}{2} = \frac{1}{\omega_a} \sqrt{-\omega^2} = m_a \sqrt{-\omega^2}. \tag{17}$$

The band elimination filter cannot be obtained from the band-pass filter by a limiting process. For the simplest band elimination filter

with a single peak as shown on Fig. 2, filter 2, the equations are

$$\begin{aligned} Z_0 &= \sqrt{\frac{L_1}{C_2}\left(\frac{1-\omega^2/\omega_A^2}{1-\omega^2/\omega_B^2}\right)}; \\ \tanh\frac{P}{2} &= \sqrt{\frac{-\omega^2 L_1 C_2(1-\omega^2/\omega_B^2)}{(1-\omega^2/\omega_A^2)}} = \frac{1}{\omega_\alpha}\sqrt{\frac{-\omega^2(1-\omega^2/\omega_B^2)}{(1-\omega^2/\omega_A^2)}}; \\ \omega_\alpha &= \frac{1}{\sqrt{L_1C_2}} = \sqrt{\frac{-\omega_\infty^2(1-\omega_\infty^2/\omega_B^2)}{(1-\omega_\infty^2/\omega_A^2)}}. \end{aligned} \tag{18}$$

Hence when the position of the peak of infinite attenuation and the characteristic impedance Z_0 at zero frequency are specified L_1 and C_2 can be determined.

We next consider the case of a filter with a total of three resonances rather than two. For a band-pass filter this will be represented by the impedance arms shown on Fig. 23B. The impedance of the series and lattice arms will be

$$Z_1 = \frac{-j}{\omega C_1}\left[\frac{(1-\omega^2/\omega_A^2)(1-\omega^2/\omega_B^2)}{1-\omega^2/\omega_2^2}\right]; \quad Z_2 = \frac{-j}{\omega C_2}\left(1-\frac{\omega^2}{\omega_2^2}\right). \tag{19}$$

Combining these to form the propagation constant and characteristic impedance we find

$$\begin{aligned} Z_0 &= \sqrt{\frac{-1}{\omega^2 C_1 C_2}(1-\omega^2/\omega_A^2)(1-\omega^2/\omega_B^2)}, \\ \tanh\frac{P}{2} &= \sqrt{\frac{C_2}{C_1}\frac{(1-\omega^2/\omega_A^2)(1-\omega^2/\omega_B^2)}{(1-\omega^2/\omega_2^2)^2}} \\ &= B\sqrt{\frac{(1-\omega^2/\omega_A^2)(1-\omega^2/\omega_B^2)}{(1-\omega^2/\omega_2^2)^2}}. \end{aligned} \tag{20}$$

We wish to show now that this type of section has an attenuation characteristic equal to that obtained by two sections of the kind shown in Fig. 23A. To show this we write

$$\tanh\frac{P_1+P_2}{2} = \frac{\tanh\frac{P_1}{2}+\tanh\frac{P_2}{2}}{1+\tanh\frac{P_1}{2}\tanh\frac{P_2}{2}}. \tag{21}$$

Substituting the value of $\tanh P/2$ given by equation (14) in (21) and letting the two cutoffs ω_A and ω_B coincide for the two sections, we have

$$\tanh\frac{P_1+P_2}{2} = \frac{m_1+m_2}{1+m_1m_2}\sqrt{\frac{(1-\omega^2/\omega_A^2)(1-\omega^2/\omega_B^2)}{(1-\omega^2/\omega_2^2)^2}}, \tag{22}$$

where

$$\omega_2^2 = \frac{\omega_A^2\omega_B^2(1+m_1m_2)}{\omega_A^2+\omega_B^2m_1m_2}. \tag{23}$$

Comparing (22) with (20) we see that

$$P = P_1 + P_2 \quad \text{and} \quad B = \frac{m_1 + m_2}{1 + m_1 m_2} = \tanh\left(\frac{P_{0_1} + P_{0_2}}{2}\right). \tag{24}$$

We see then that a section with three resonant frequencies can be made to have the same attenuation characteristic as the sum of two simple sections. It is, however, more general since in equations (23) and (24) real values of ω_2^2 and B can be obtained by taking

$$m_1 = m_{1_r} + i m_{1_i}; \qquad m_2 = m_{1_r} - i m_{1_i}; \tag{25}$$

that is, the parameter m_1 can be made complex if the second parameter m_2 is made its conjugate. Such complex sections can be made to have attenuation peaks which are finite even in the absence of dissipation.[7]

By letting $\omega_A \to 0$ or $\omega_B \to \infty$, the equivalent relations for low-pass, high-pass and all-pass filters can be obtained. These are

Low-Pass Filter

$$\tanh\frac{P}{2} = (m_1 + m_2)\sqrt{\frac{-(\omega^2/\omega_B^2)(1 - \omega^2/\omega_B^2)}{(1 - \omega^2/\omega_2^2)^2}}; \qquad \omega_2^2 = \frac{\omega_B^2}{1 + m_1 m_2}; \qquad m_1 = \sqrt{1 - \frac{\omega_{\infty 1}^2}{\omega_B^2}}. \tag{26}$$

High-Pass Filter

$$\tanh\frac{P}{2} = \frac{m_1 + m_2}{1 + m_1 m_2}\sqrt{\frac{(1 - \omega^2/\omega_A^2)}{(1 - \omega^2/\omega_2^2)^2}}; \quad \omega_2^2 = \frac{\omega_A^2\omega_B^2[1 + m_1 m_2]}{\omega_A^2 + \omega_B^2 m_1 m_2}. \tag{27}$$

All-Pass Filter

$$\tanh\frac{P}{2} = (m_1 + m_2)\sqrt{\frac{-\omega^2}{(1 - \omega^2/\omega_2^2)^2}}; \qquad \omega_2^2 = \frac{1}{m_1 m_2}. \tag{28}$$

Band Elimination Filter

For a two-peak band elimination filter such as shown in Fig. 3, filter 2, the equations are:

$$Z_0 = \sqrt{\frac{L_1}{C_2}}\frac{(1 - \omega^2/\omega_A^2)(1 - \omega^2/\omega_B^2)}{(1 - \omega^2/\omega_2^2)^2}; \qquad \tanh\frac{P}{2} = \frac{1}{\omega_\alpha}\sqrt{\frac{-\omega^2}{(1 - \omega^2/\omega_A^2)(1 - \omega_2^2/\omega_B^2)}}; \qquad \omega_\alpha = \frac{1}{\sqrt{L_1 C_2}} = \sqrt{\frac{-\omega_{\infty 1}^2}{(1 - \omega_{\infty 1}^2/\omega_A^2)(1 - \omega_{\infty 1}^2/\omega_B^2)}}, \tag{29}$$

$\omega_{\infty 1}\omega_{\infty 2} = \omega_A\omega_B$.

In a similar manner more sections can be added and the resonant frequencies determined in terms of the cutoff frequencies and the position of the attenuation peaks. The most general section considered in this paper has a maximum of five equivalent sections. For this case by applying the process described above the propagation constant and critical frequencies are given by the equations

$$\tanh \frac{P}{2} = \frac{A + C + E}{1 + B + D} \sqrt{\frac{(1 - \omega^2/\omega_A{}^2)(1 - \omega^2/\omega_3{}^2)^2(1 - \omega^2/\omega_5{}^2)^2}{(1 - \omega^2/\omega_2{}^2)^2(1 - \omega^2/\omega_4{}^2)^2(1 - \omega^2/\omega_B{}^2)}}, \quad (30)$$

where

$$A = \sum_1^5 m = m_1 + m_2 + m_3 + m_4 + m_5;$$

$$B = \sum_{m=1}^{5} \sum_{n=1}^{5} m_m m_n; \qquad n \neq m;$$

$$C = \sum_{m=1}^{5} \sum_{n=1}^{5} \sum_{o=1}^{5} m_m m_n m_o; \qquad n \neq m; \qquad n \neq o; \qquad m \neq o;$$

$$D = \sum_{m=1}^{5} \sum_{n=1}^{5} \sum_{o=1}^{5} \sum_{p=1}^{5} m_m m_n m_o m_p; \qquad m \neq n; \qquad m \neq o;$$

$$m \neq p; \qquad n \neq p; \qquad n \neq o; \qquad o \neq p;$$

$$E = m_1 m_2 m_3 m_4 m_5. \quad (31)$$

The resonant frequencies are given by the equations

$$f_2{}^2 = \frac{2 f_A{}^2 f_B{}^2 (1 + B + D)}{f_A{}^2(2 + B - \sqrt{B^2 - 4D}) + f_B{}^2(B + 2D + \sqrt{B^2 - 4D})}, \quad (32)$$

$$f_4{}^2 = \frac{2 f_A{}^2 f_B{}^2 (1 + B + D)}{f_A{}^2(2 + B + \sqrt{B^2 - 4D}) + f_B{}^2(B + 2D - \sqrt{B^2 - 4D})}, \quad (33)$$

$$f_3{}^2 = \frac{2 f_A{}^2 f_B{}^2 (A + C + E)}{f_A{}^2(2A + C - \sqrt{C^2 - 4AE}) + f_B{}^2(C + 2E + \sqrt{C^2 - 4AE})}, \quad (34)$$

$$f_5{}^2 = \frac{2 f_A{}^2 f_B{}^2 (A + C + E)}{f_A{}^2(2A + C + \sqrt{C^2 - 4AE}) + f_B{}^2(C + 2E - \sqrt{C^2 - 4AE})}. \quad (35)$$

For any smaller number of sections the values can be obtained by letting some of the m's go to zero. For example, for a three section filter $m_4 = m_5 = 0$. For low, high, and all-pass networks the values can be obtained by letting $f_A{}^2 \to 0$, $f_B{}^2 \to \infty$ or $f_A{}^2 \to 0$; $f_B{}^2 \to \infty$.

The Coronaviser, an Instrument for Observing the Solar Corona in Full Sunlight*

By A. M. SKELLETT

Introduction

BECAUSE of the rarity of solar eclipses, their short duration, and their occurrence usually at inconvenient places on the earth's surface, the problem of observing the solar corona in full sunlight is an important one for astronomers. It is also of considerable interest to those telephone engineers who are concerned with radio transmission over long distances. The major disturbances of such transmission have their origin in the sun and studies to date have indicated that a day-to-day knowledge of the activity of the corona might prove useful in predicting the transmission conditions.

The first attempt to solve this problem was made by Huggins in 1878 and since that time every conceivable optical means to accomplish the desired result has been tried. The problem is to observe the corona, not in itself a faint object, through the blinding glare of the sky in the region around the sun. If one holds his hand at arm's length so that it blots out the sun, he will find the glare in the sky around it so intense as to be painful. It is generally at least a thousand times brighter than the corona. The trials have usually been made at very high altitudes where the atmospheric glare is greatly reduced but since the scattered light from the telescope itself, particularly the objective, is some hundreds of times brighter than the corona no success was obtained until M. Lyot[1] invented his *coronographe*, a telescope in which this latter kind of glare is greatly reduced. With this instrument at the top of Mt. Pic du Midi in the Pyrenees mountains he has obtained several photographs which show some of the features of the inner corona. At best he has to work with a glare that is nearly as bright as the brightest parts of the corona. There are only a few days in the year when the intensity of the glare is low enough to enable him to observe coronal features through it.

It is obvious that a method of greater discrimination is needed if day-to-day observations are to be made. Such a method was proposed several years ago.[2] It is based on the use of television technique;

* Presented at mtg. of Nat. Acad. Sciences, Providence, R. I., October 1939, and before Amer. Philos. Soc. in Philadelphia, Pa., November 1939.

[1] Lyot, B., *M. N. R. A. S.*, **99**, 8, 580, 1939.

[2] Skellett, A. M., *Proc. Nat'l. Acad. Sci.*, **20**, 461, 1934.

the corona is separated from the glare by electrical filters while the image of the sky around the sun is temporarily represented by an electric current.

Apparatus and Technique

Dr. G. W. Cook, Director of the Cook Observatory, kindly offered the use of his 15-inch horizontal telescope for a trial of the method. Special television apparatus was developed at the Bell Telephone Laboratories for use in conjunction with this telescope. This apparatus has been called the coronaviser. Figure 1 shows a layout of the complete apparatus. Starting at the left, the plane mirror M is coupled to driving mechanism to form a siderostat so that sunlight may be continuously held on the axis of the telescope. The solar disc, about 2″ in diameter, is focussed by the objective of approximately 18 feet focal length to fall on the tilted mirror which reflects the direct sunlight out through a hole in the side of the telescope and into a light trap consisting of a black walled tube with a black velvet end. Figure 2 shows the scanning mechanism in greater detail. Immediately in back of the mirror R, which is the one just referred to for throwing sunlight out of the telescope, a black masking disc D further prevents sunlight from getting into the scanning apparatus. Several different sizes of this disc are used to take care of the different diameters of the solar image which occur throughout the year and to shield the scanner from sunlight which spills over with bad seeing. This mirror and disc are supported by means of the plate glass P so that there is no obstruction in the field around the sun.

To the right of the plate glass lies the scanning apparatus. This is a mechanical device which scans the region of the sky around the sun in a spiral path. The scanning motion thus consists of a circular and a radial motion.

The simple plano-convex lens L which is silvered on the back is equivalent to a concave mirror and forms an image of a portion of the sky image that lies in the plane of D on the scanning hole H which is on the axis of the telescope and scanner. The light that enters the scanning hole passes through the lens, the prism, and the light tunnel U into the photo-cell E. When the lens L is rotated about the axis by the motor, the effect is the same as moving the scanning hole around in the image plane at D. This takes care of the circular component of the scanning motion.

The radial component is obtained by changing the angle of tilt of the lens L while it is rotating. A worm W is mounted on the motor shaft but held stationary so that as the gear G revolves as a whole

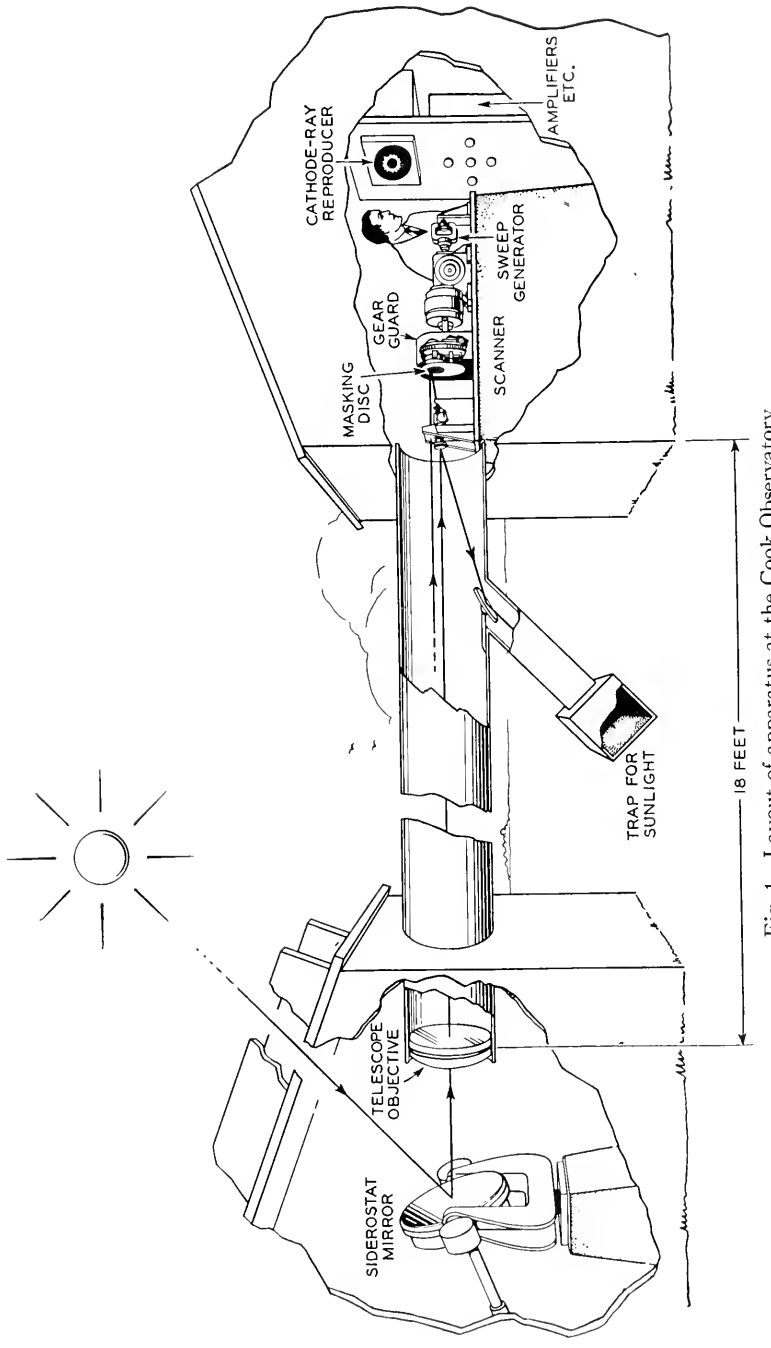


Fig. 1—Layout of apparatus at the Cook Observatory.

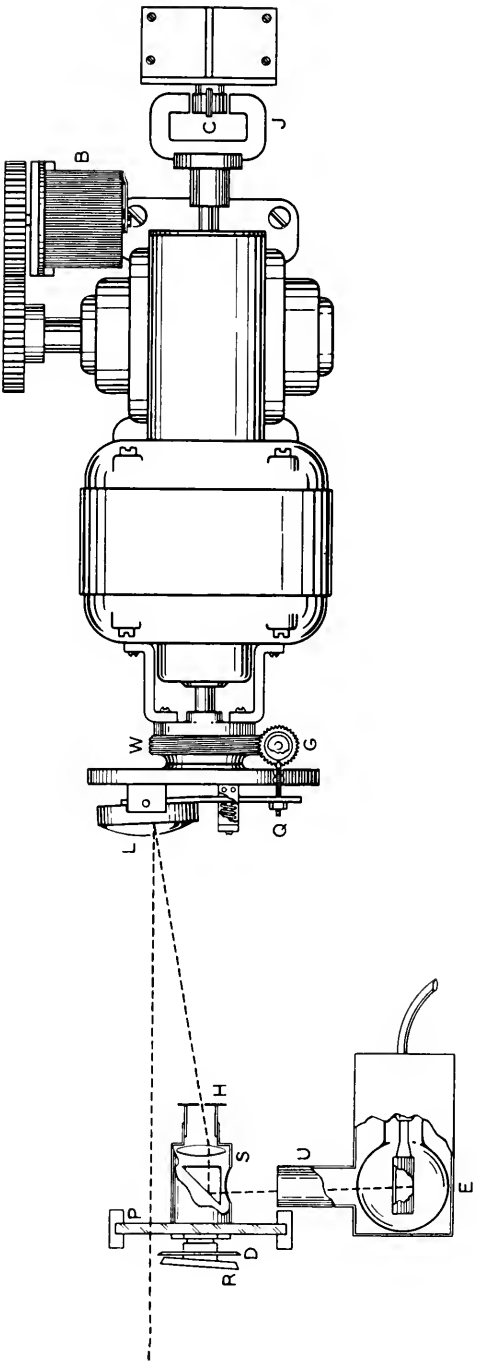


Fig. 2—The mechanical scanner and sweep frequency generator.

about the axis of the scanner it is turned more slowly about its own axis. A hardened pin Q attached to the arm of the lens mounting rides on a cam which is fixed to this gear and thus imparts a cyclic tilting motion to the lens. There are another similar gear and cam (not shown) mounted opposite this one which serve only to compensate for the weight of the working gear so that the rotating mechanism may be balanced to a high degree of precision.

The lens behind the scanning hole forms a stationary enlarged image of the scanning hole on the photosensitive surface of the photoelectric cell.

The reproduction of the image is done entirely by electrical means. At the opposite end of the motor a C-shaped permanent magnet rotates about a cylindrical iron core around which there is wound a coil of wire. The electric current thus generated in the coil is smoothed into a true sine wave by tuning the circuit of the coil so that it resonates at the frequency of circular scan which is approximately 30 cycles per second. The resulting wave is split into two components 90° apart in phase and these, after amplification, are impressed on the deflector plates of the reproducing cathode ray tube so that the spot may move in a circular path.

At the generator end of the motor there is also a reduction gear box, the slow speed shaft of which turns once for a complete cycle of the radial scanning motion, i.e. once a second. This shaft is geared to that of the potentiometer unit B so that its sliding contact also revolves at this rate. The potentiometer winding is continuous around the circle and connections are made at the opposite ends of a diameter. The scanning voltage from the generator circuit is fed into this unit and as the arm revolves the amplitude of the sine wave is made to vary uniformly between its minimum and maximum values. Thus the scanning spot spirals outward, scanning a complete image of the field, and then inward, giving another complete image, so that there are two images per cycle of the radial scanning motion. The spot does not follow the identical path on the two scans; the outward scan crosses over the lines of the inward scan along one diameter and interlaces along the diameter at 90 degrees to this. The radial resolution along the latter diameter is therefore double that along the former. Since the frequency of the radial component of the scanning motion is approximately one cycle per second, these two resolutions are 15 and 30 lines respectively.

The glare of a clear sky is uniform around the solar image and therefore as the scanning spot travels around the field it gives rise mainly to a direct current in the photo-cell. The coronal features,

however, that is the streamers, arches, etc., give rise to alternating components and only these components are amplified, the direct current being eliminated by resistance capacity coupling of the cell to the amplifier. Inaccuracies of alignment and any non-uniformity of the intensity of the glare across the field give rise to strong low-frequency components and because of the high glare levels that were occasioned by the location and by the siderostat mirror it was found necessary to filter these components out of the amplified currents. A high-pass filter is inserted between the first and second stages of the amplifier for this purpose. A low-pass filter with a cut-off at 3750 cycles is also included to cut out the noise at frequencies above the desired band. The top frequency is 3600 cycles for the 0.04 inch diameter scanning hole that was generally used. After amplification the signal current is made to modulate the intensity of the cathode-ray beam.

The electrical frequency spectrum of the television image consists of the fundamental scan frequency (about 30 cycles) plus a large number of its harmonics. By varying the characteristics of the high-pass filter it was possible to eliminate the fundamental and several of the lower harmonics as desired. This became advantageous in studying the prominences and smaller coronal details which give rise almost entirely to higher harmonics and are reproducible therefrom with a good degree of accuracy. In addition, for these smaller details, the coupling capacity between the first and second stages of the amplifier was greatly reduced so that the gain at the upper end of the band was considerably enhanced in relation to that at the lower end.

The light of the corona is practically identical with that of the sun in its spectral characteristics and a caesium sulphide photo-cell which has a maximum sensitivity in the green was used. See Fig. 3. It was found that by using gas amplification in the photo-cell adequate sensitivity was obtained. The inner corona has a surface brightness of about the same magnitude as the full moon and the sensitivity of the apparatus was checked by obtaining images of the moon in its various phases.

It is convenient to measure light intensity levels in millionths of the brightness of the sun's surface. The brightness of the full moon is about 2 millionths and it is known that the inner corona falls off fairly rapidly with distance from a brightness of a little more than one millionth measured within one minute of the solar limb. The level of the glare was measured from time to time by means of a photronic cell behind an aperture placed at the focus of the objective. On days when the haze in the sky was very noticeable but yet not in the form of clouds its brightness at 2 minutes from the limb was as high as 6000

millionths or approximately 6000 times as bright as the corona. On very clear days the brightness was as low as 1250 millionths and this limit may have been set by the scattered light from the telescope parts, particularly the siderostat mirror, rather than the sky. The

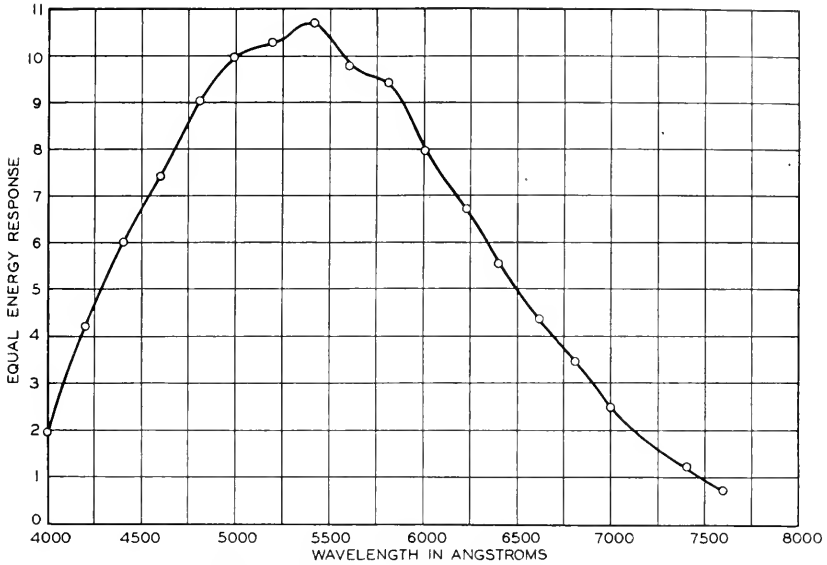


Fig. 3—Spectral characteristic of photo-cell.

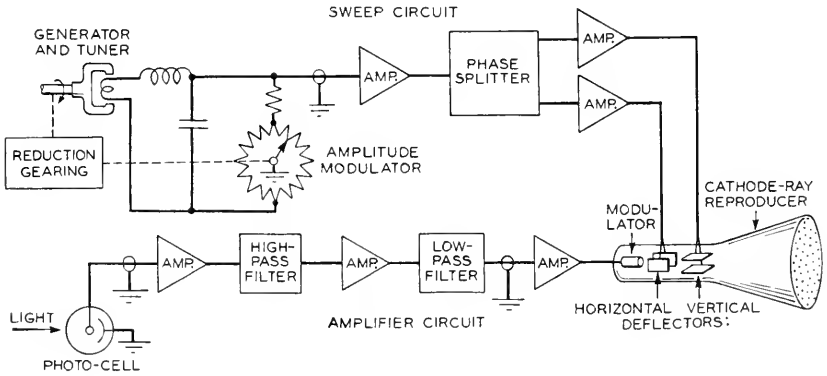


Fig. 4—The circuit diagram of the coronaviser.

siderostat mirror had a surface of evaporated aluminum but by the time the development of the equipment had progressed to the point where useful images could be obtained, its surface had begun to show blemishes.

The scanning rate was so slow (2 images per second) that even with a cathode ray screen having an appreciable time lag there was difficulty in studying the complete image visually. The expedient was therefore adopted of photographing the images and this proved advantageous in comparing images as well as in greatly reducing the effects of noise. The latter advantage was realized by taking exposures of from 20 to 30 seconds during which there were reproduced some 40 to 60 images and although the noise patterns might be very noticeable in a single scan the fluctuations balanced out in a statistical manner, leaving a uniform field.

Theoretically the limiting amount of glare through which it is possible to work with this method is determined by the shot noise in the photoelectric current but there are practical limits, set mainly by the cleanness of certain of the optical parts, which were the important factors in this case. Although the whole scanner is designed to reduce, as much as possible, the scanning of the optical parts of the instrument by the beams of light, certain parts are of necessity scanned, particularly the plate glass used to support the scanning hole unit. This plate is also near the focal plane and the slightest smudge or speck of dust on its surfaces gives rise to an overloaded image on the cathode ray screen. The glass itself was specially selected to be free from bubbles or blemishes of any kind and in addition it was carefully washed at frequent intervals. Great pains had also to be taken in keeping the other glass surfaces in the optical train clean, for the essence of the method is the amplification of minute variations in the intensity of the illumination from point to point in the field.

Occasionally tiny specks of brilliant light would float across the screen, the sources of which were very puzzling at first. They were finally traced to insects or wind-borne seeds which drifted across the sky in the path of the shaft of light. Being illuminated by direct sunlight, they scattered enough light in the direction of the telescope to give a bright diffraction pattern. They ruined quite a few plates.

Since the glare decreased in the direction outward from the sun (though not so rapidly as the coronal light), patterns that were caused by instrumental defects took on at times appearances which might easily have been confused with that of a coronal image. It was necessary, therefore, to have an absolute criterion by which one could distinguish between these spurious images and those which were associated with the sun. The siderostat mounting of the telescope furnished such a test. With this type of mounting the celestial field rotates about the optic axis of the telescope with time. This rate varies with the declination of the object and other factors and in our case it ranged

within a degree or two of 7° per hour. Thus by taking a series of photographs over a period of several hours it was possible to determine definitely whether or not the image in question was associated with the sun or with the apparatus. In addition to this test, for the prominences, there was the spectrohelioscope at hand by which a direct comparison could be made. Another test applied to the prominence images was furnished by their color. A red glass filter, such as the Schott RG 2 which has a cut-off just below the H_α line, reduced the general glare level by about 30 times whereas its reduction of the light of the prominences which is a maximum at this wave-length was not nearly so great.

Results

The prominences shown in Fig. 5 are among the first of which good images were obtained. Seven photographs were taken of them be-

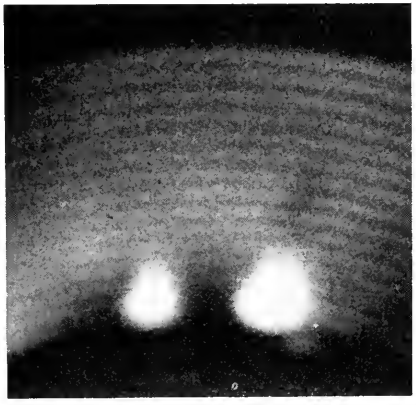


Fig. 5—Prominences taken with red filter on Feb. 21, 1938.

tween $16^h\ 58^m$ and $19^h\ 11^m$ G.C.T. on February 21, 1938, some in white light and the others with the red filter in front of the photo-cell. This particular photograph was taken in red light; those taken in white light were of considerably less contrast.

Figure 6 is another one of the many prominence photographs that have been taken with the apparatus. These are the prominences that were present around the sun at $18^h\ 30^m$ G.C.T. on October 31, 1938. This was also taken with the red filter.

Figure 7 shows a pair of bright prominences photographed in white light on October 3, 1938.

Figure 8 shows a jet or flare in the corona that was photographed on October 18, 1938. It is one of 11 photographs that were taken

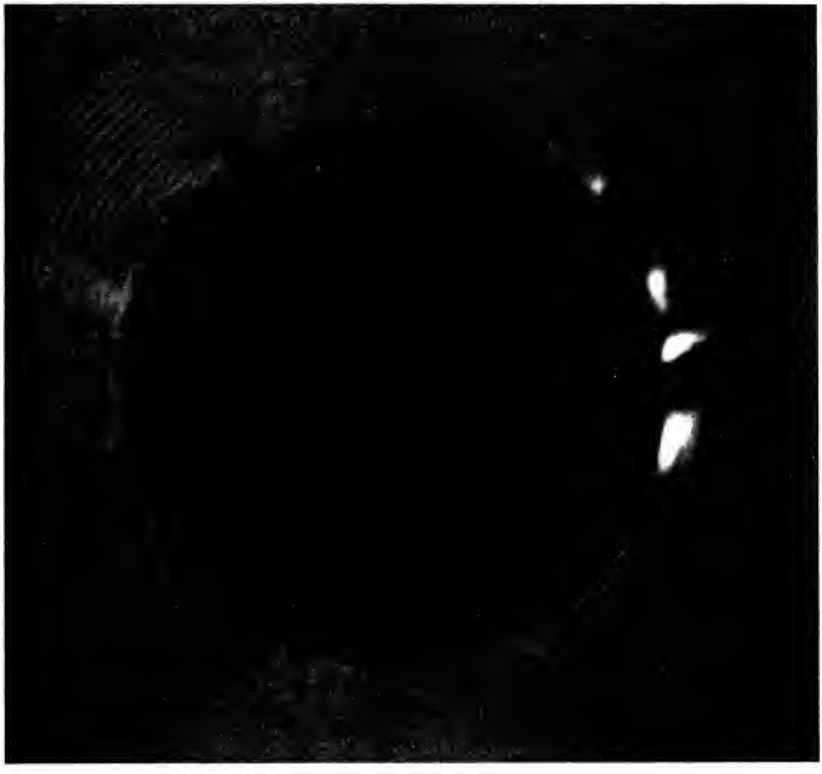


Fig. 6—Prominences around the sun on Oct. 31, 1938. This image was obtained with a scanning hole 0.013 inch in diameter which had 1/10 the area of that used for the other photographs.

Fig. 7—A pair of prominences taken without optical filter (i.e., in white light) on Oct. 3, 1938.

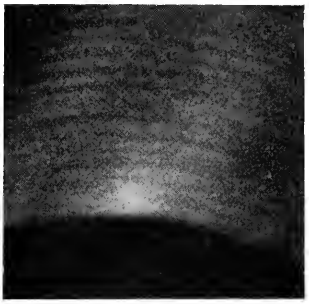


Fig. 8—A jet or flare in the corona photographed on October 18, 1938.

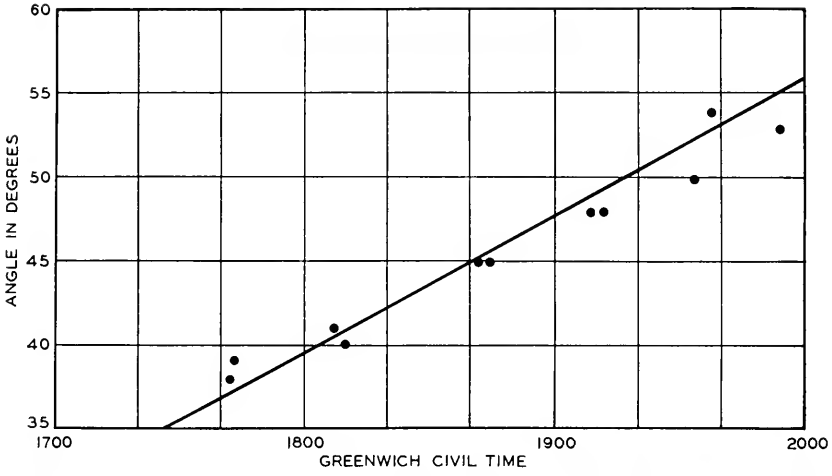


Fig. 9—Positions of the flare of Fig. 8 at the times shown. The turning of the image is due to the horizontal mounting of the telescope.

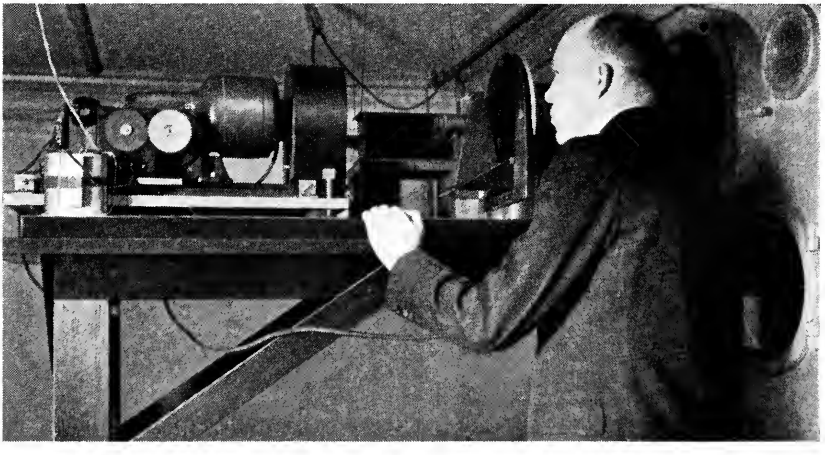


Fig. 10—The input scanner.

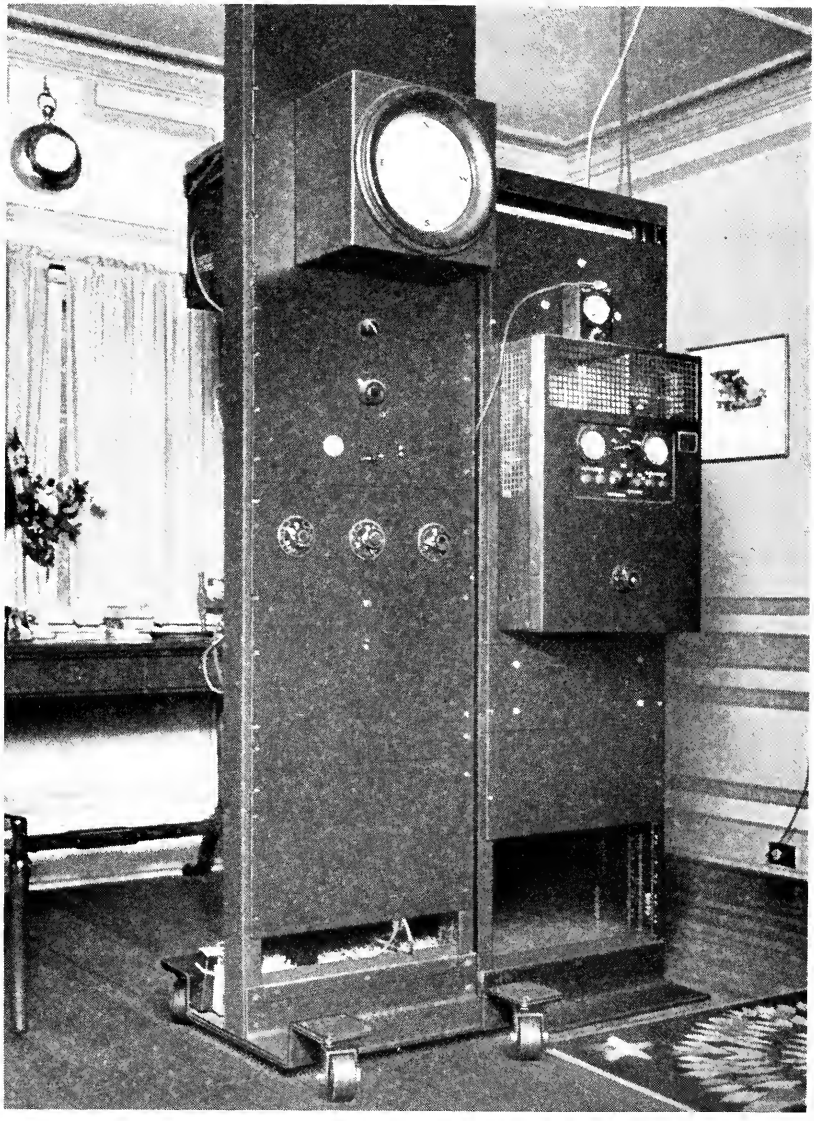


Fig. 11—The amplifying and sweep circuits and reproducing tube.

over a period of more than 2 hours. All these show this feature, and its position on each plate is plotted against the time of exposure on Fig. 9. The slope of the line is the correct rate for the turning of the celestial field in the neighborhood of the sun at that time. It was not brighter on the plates that were taken with the red filter than on those taken with white light, and it is concluded, therefore, that it was white in color.

There was no prominence at its position on the limb of the sun, but the next day the observatory at Huancayo reported that an eruptive prominence blew off from the limb at this position (28° N. Lat.). It seems quite probable that the unusually bright jet in the corona was lying over and probably associated with this active region.

A number of other plates have shown details which appear to have their origin in the corona, but generally they have been partly obscured by other patterns of instrumental or other origin.

The major objectives of this phase of the work which were the development of an adequate instrument and the proving in of the method have been achieved. The next phase of the investigation should be carried out under the most favorable conditions possible and this means a location on a mountain top with a telescope preferably pointing directly at the sun.

I wish to acknowledge the helpful cooperation of Dr. Cook and his associates at the Cook Observatory and of many of my colleagues in the Bell Telephone Laboratories. In particular, Dr. J. B. Johnson has greatly contributed to the investigation by his counsel and aid.

Lead-Tin-Arsenic Wiping Solder *

By EARLE E. SCHUMACHER and G. S. PHIPPS

SOME fourteen or more wiped joints occur in every mile of lead-sheathed telephone cable, and in making these joints from one to two million pounds of solder are used per year. To join cables a lead sleeve of sufficient diameter to accommodate the bundle of spliced wires is slid in place at the junction, the ends of the sleeve are beaten to conform to the circumference of the cable, and an air-tight and mechanically strong joint formed at each end of the sleeve by molding a solidifying mass of solder into the desired shape. This last step is called the wiping operation.

The making of a successful wiped joint depends upon a satisfactory composition in the solder and considerable skill on the part of the splicer. The two factors are inter-related in that the more dextrous operators can produce satisfactory joints with solder compositions which could not be shaped by the average operator. The most satisfactory composition for a wiping solder from practical tests has been found to be about 38 per cent tin, 62 per cent lead. A solder containing 40 per cent tin also possesses satisfactory handling qualities and is used to some extent. If the tin content is much above 40 per cent the workable temperature range in which the solder is plastic becomes too limited for practical handling. The plastic range can be increased by increasing the lead content above 62 per cent but then it is found that the joint becomes coarse-grained and porous. The highest practicable lead content is of course advantageous from an economic standpoint.

The impurities allowable in a wiping solder are also closely controlled since in general small percentages of most impurities have been found to have a harmful effect upon the handling character of the solder or the properties of the joints. One exception, which has hitherto not been recognized, is arsenic whose beneficial effects in small quantities are discussed in this paper.

An engineer would prefer to interpret the handling of a wiping solder in terms of basic properties which can be measured in the laboratory. Such attempts [1] have been made but with only limited success since

* *Metals and Alloys*, Vol. 11, pp. 75–76, March 1940.

[1] "Some Physical Properties of Wiping Solders," D. A. McLean, R. L. Peek, Jr., and E. E. Schumacher, *Journal of Rheology*, Vol. 3, January 1932, p. 53.

not only does the wiping process itself not admit of scientific measurement but also the basic properties related to the process are difficult to determine and are of restricted practical bearing. These difficulties arise because a complex solid-liquid system is involved, and during most of the time of wiping the joint the system is far from being in an equilibrium condition.

Experience has shown, however, that a wiping solder should possess certain general characteristics which are enumerated below, although in many instances the characterization cannot be extended beyond a qualitative statement.

1. The temperature at which the solder begins solidification should be lower than the temperature of beginning melting of the cable sheath and sleeve. The temperature of beginning solidification for the 38–62 tin-lead solder is 240° C. while the lead alloy sheath begins to melt at approximately 310° C. Since with this solder no trouble is encountered with melting sheath it appears that a 70° differential is satisfactory.

2. The solidification range of the solder should be such as to provide, during cooling, an ample forming period between the time when enough primary lead has precipitated to give sufficient body to permit forming to begin until the mass is too solid to manipulate. In the 38–62 wiping solders the solidification range is approximately 60° C. while the forming range is about 40° C.

3. The tendency for the joint to drain or slip and break apart during wiping should be a minimum. These properties are associated with surface tension and plasticity. The desired condition is sometimes referred to as a "buttery" texture.

4. The solder should readily wet and alloy with the parts to be joined. This implies a freedom from non-reducible oxides in the solder and a minimum in the tendency to form reducible oxides. Suitable non-corrosive fluxes are used to clean the surfaces being joined and reduce the oxides which cannot be entirely excluded from the melted solder. To facilitate handling, the solder should not adhere to the splicers' wiping cloths.

5. The solder should be such that the strength of the joint formed should be equal to or greater than that of the parts being joined. The joint must also be gas-tight. This property is secured through a fine texture in the solder and freedom from draining of the lower melting constituents. The test for porosity of a joint is simple. A positive gas pressure of from six to nine pounds per square inch is applied inside the sleeve and soap-suds are then painted on the joint. Observation of the soap-suds will then show whether or not the joint is porous.

6. The health hazard under normal conditions of use should be negligible.

An investigation was made of modifications of the lead-tin type wiping solder to produce a solder which more fully meets the practical requirements than does the present compositions. Through the addition of 0.1 per cent of arsenic substantial improvements have been

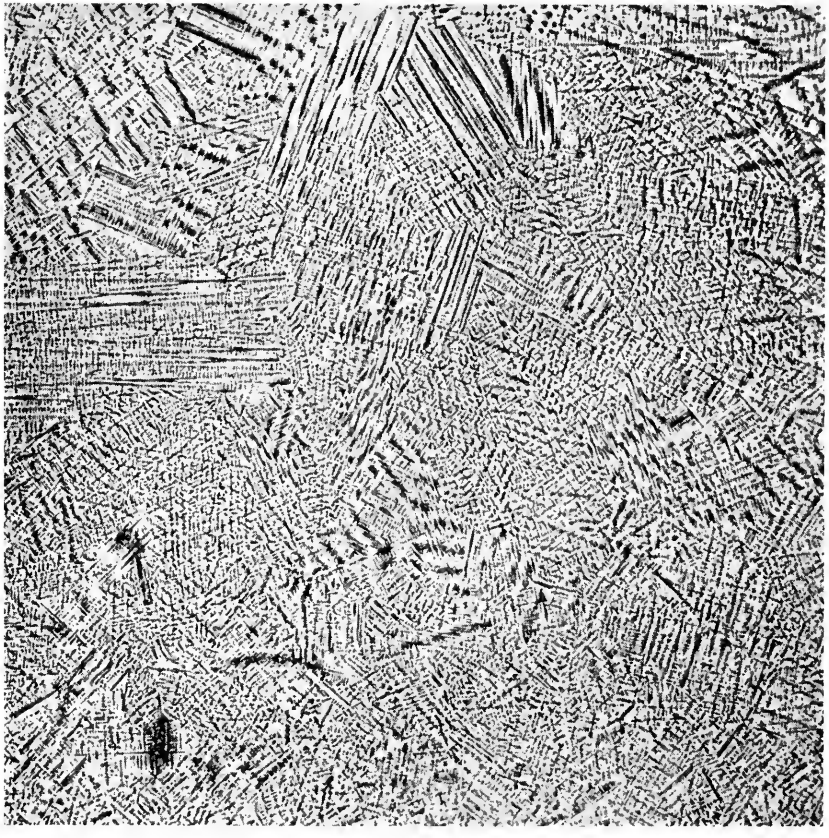


Fig. 1—Photomacrograph of a nominal 38% tin–62% lead wiping solder. The sample was slowly cooled from 300° C. Specimen etched with a mixture of 4 parts of glacial acetic acid and 1 part of 30% hydrogen peroxide.

achieved in certain of the characteristics. The nominal composition of an alloy which shows this improvement is tin—37.25 per cent, arsenic—0.1 per cent, and balance lead. The effect of arsenic manifests itself at percentages as low as 0.04 per cent. Amounts much over 0.1 per cent should be avoided because of a tendency to segregation at the higher percentages. As the tin is reduced below the recom-

mended percentage the difficulties in producing a satisfactory joint increase. There is the possibility, however, that a lower tin content can be tolerated but definite conclusions await further substantiation in subsequent investigation.

Although the percentage of arsenic in the new solder is relatively small it has two important beneficial effects. The amount of dross formed on the arsenic-bearing solder is but a fraction of that ex-

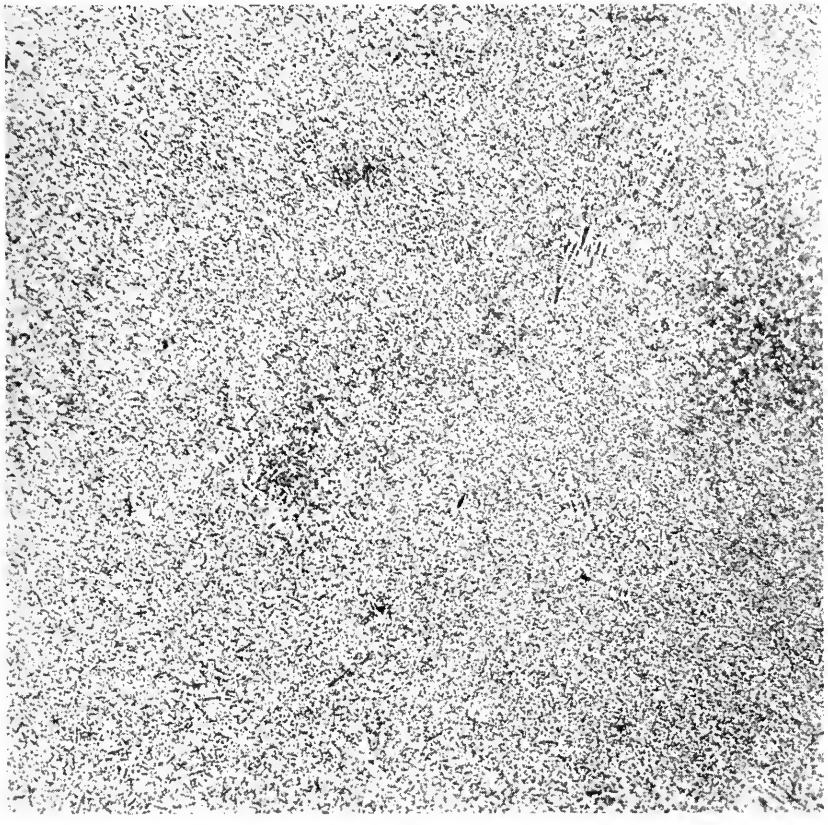


Fig. 2—Same as Fig. 1, except for an addition of 0.1% arsenic. Note the finer grain.

perienced in the ordinary lead-tin solders. The practical advantage is that less time need be spent in skimming the molten solder in the melting pot, and there is less possibility for the inclusion of dross in the finished joints. The presence of dross in the wiped joints is to be avoided because of its possible contribution to porosity.

The second beneficial effect of arsenic is related to the grain size in the solidified alloy. This is illustrated by the photomacrographs shown in Figs. 1 and 2 which compare the grain structure of a lead-tin solder

and an arsenic-modified solder which have undergone similar handling and cooling treatments. The arsenic-bearing solder exhibits a finer and more uniform texture. The finer texture is associated with improved handling characteristics and freedom from porosity in the finished joint. The texture of the solder is more buttery in the wiping process and there is less tendency for the lower freezing constituents to drain from the partially finished joints than with the lead-tin compositions. Practically, this provides for the splicer a longer forming range, although the solidification range is materially the same as for the corresponding unmodified alloy.

Although the mechanism by which arsenic reduces the size of the dendrites and refines the grain of wiping solder has not been definitely established, it seems probable that it either provides new and more numerous nuclei of crystallization around which the primary lead precipitates, or imposes barriers against the growth of the primary lead crystals or both. Arsenic forms a compound, Sn_3As_2, with tin and it is probable that this constituent is responsible for the mechanism postulated. In slowly cooled solders this compound is discernible beginning at approximately 0.1 per cent arsenic. With a greater number of precipitated crystals present a greater surface is made available to which the molten eutectic may cling, producing a readily formable mass.

Laboratory observations on the solder have been verified by field tests in the Bell System. The arsenic-bearing solder is handled in the same way as the ordinary lead-tin compositions. Joints have been wiped on large and small aerial and underground cables and in difficult situations involving large branched joints. The consensus of splicers from several different localities is that the solder possesses handling properties superior to those of the lead-tin compositions. These joints in all cases were pressure tested after wiping and found to be sound.

Regarding the possible health hazard involved in using the new solder, tests have been made to determine whether arsenic or arsenic compounds would be volatilized from this alloy under the conditions encountered in practice. These tests gave entirely negative results and showed that no additional hazards would be introduced by substituting arsenic-bearing solder for standard solder.

Nuclear Fission

By KARL K. DARROW

This article pertains to the most newly-discovered and most sensational mode of transmutation, in which the entry of a neutron into a massive atom-nucleus brings about an internal explosion in which the nucleus is "fissured" or divided into two fragments which share the total mass and charge between them in nearly equal proportions. (In all other modes of transmutation except those affecting the very lightest elements, the division is into fragments of very unequal mass and charge.) The conversion of rest-mass into kinetic energy, or (as is more commonly said) the release of energy, is unprecedented in scale. A multitude of radioactive bodies, many hitherto unknown, is formed; and there is spontaneous emission of fresh neutrons in great quantities, possibly sufficient to convert the process once initiated into a self-perpetuating one under realizable conditions.

EVERY now and then a physicist is liable to receive a letter from some yearbook or other, in which he is invited to write x thousand words on the "most important developments in physics during the year just ending." The only safe reply is of course that for ten years at least and perhaps for fifty it will be impossible to tell which is the most important development in physics during the year just ending. This year, however, it looks as though one need not be so cautious; for ever since the first few weeks of the year many have felt pretty sure that one particular discovery would long be recognized as the most important to be made, or at any rate to be revealed, in 1939. It came early—the first publication was on the sixth of January, and there was a rain or perhaps I should say a deluge of others before the end of February. Inasmuch as these others proceeded from laboratories sprinkled all of the way from Copenhagen to Berkeley, it is literally true for once that a discovery commanded immediate attention. Nor is attention even yet diverted, though the pace of publication has grown less.

The phenomena of fission are as yet confined to the last three elements of the periodic table: thorium, protactinium, uranium. I show their chemical symbols, their atomic numbers or nuclear charges, and the mass-numbers—to wit, the nearest integers to the actual values of the masses—of their several isotopes (charges expressed of course as multiples of e, masses as multiples of one-sixteenth the mass of the

commonest kind of oxygen atom); charge appears as a subscript before the symbol, mass as a superscript after it:

$$_{90}Th^{232}; \qquad _{91}Pa^{231}; \qquad _{92}U^{234}; \qquad _{92}U^{235}; \qquad _{92}U^{238}$$

From this list I omit several very unstable isotopes of which we shall probably never be able to assemble enough to observe their fission. Protactinium 231 is itself so rare that only one man in the world (he is von Grosse, of Chicago) ever got enough of it together for this experiment. He brought his precious sample—less than 9 mg.—to Dunning at Columbia for the test, and the three of them found fission. I make this allusion at the start, because there will be little further occasion to refer to protactinium, and yet it should not be forgotten. There is danger of forgetting even thorium, since so disproportionately great an amount of study has been lavished on uranium. Neither thorium nor uranium is a very rare element, but more than 99 per cent of any sample of uranium consists of the isotope 238, so that the two other isotopes must also be classed as rare; yet it is believed at present that 235 is responsible for some of the most remarkable of the phenomena.

Now let me indicate two qualities shared by all five of these nucleus-types. First: all are radioactive, that is to say, they are unstable. I must not be too emphatic with this word; the average lifetime of nuclei of either Th^{232} or U^{238} is hundreds of millions of years, and there are not many organizations which would be considered unstable if they could bank on a probable lifetime of that scale. Still they are, in the physicist's sense, unstable; and this suggests that it might be relatively easy to disorganize, to disrupt, to explode them by a fitting agency coming from without.

Now to any physicist the term "fitting agency coming from without" suggests at once the bombarding particles by which transmutation was first effected: alpha-particles, protons, deuterons—the positively-charged nuclei of the elements helium and hydrogen at the other end of the periodic table from uranium. Should one not project these nuclei against uranium nuclei, and see what happens? Well, it has often been done, and nothing has happened;[1] and an adequate reason is supplied by the second important quality of these five nucleus-types, their greatness of atomic number. All of them are so highly charged with positive electricity, that the proton, the deuteron and the alpha-particle, however fast they are when they start, cannot approach them closely enough to do them any harm. (What with the current progress in cyclotrons that statement may soon be out of date!) Even with our

[1] Until in October of this year Gant reported strong indications of fission of uranium produced by very energetic (8-Mev) deuterons.

present resources of energy we could not tamper with any of these nuclei, had we not at our disposal those chargeless particles the "neutrons" with which to assail them.

One might of course foretell that mighty powers of transmutation would be possessed by a particle which is *not* repelled as it approaches a nucleus. Actually the transmuting powers of the neutron are greater than, I should think, anyone can have expected; nor can many people, if any, have foreseen that the *slow* neutron—the neutron having no more speed and kinetic energy than a molecule of air at room-temperature—would prove to be more potent than the fast one. Yet so it is. When the other agents of transmutation were first applied—alpha-particle, proton, deuteron, photon, fast neutron—it took years to get proof of the transmutation of even a few elements; but when the slow neutron was first applied, Fermi and his half-a-dozen colleagues at Rome managed to do something to almost every element in a very few months! Let me recall that neutrons mostly are what we call fast—i.e., they have energies of millions of electron-volts—when they start their careers. Slow neutrons are initially-fast ones which have been sent through layers of paraffin or water, and have lost nearly the whole of their initial energy by making elastic impacts with hydrogen nuclei. We shall later have to distinguish between the fast neutrons and the slow as agents of nuclear fission.

Now to supply a fitting historical background to the discovery of fission, I must draw attention to a theorem which until the end of 1938 was believed to govern the whole of transmutation, and which still governs nearly the whole of the field. It is this: with the exceptions presently to be related, *no transmutation ever produces a change in atomic number greater than 2 or a change in mass-number greater than 4.* I am going to illustrate this theorem by writing in symbolic form three of the reactions of transmutation produced by neutrons and recognized before the end of 1938. Here I use E as the general symbol for element; Z and A as the general symbols for atomic number and mass-number; and α, p, d, n, and ϕ for alpha-particle, proton, deuteron, neutron and photon; and I recall that the mass-numbers of these five particles are 4, 1, 2, 1, 0 respectively.

$$ {}_{Z}E^{A}\ (n, \alpha)\ {}_{Z-2}E^{A-3}; \qquad {}_{Z}E^{A}\ (n, p)\ {}_{Z-1}E^{A}; \qquad {}_{Z}E^{A}\ (n, \varphi)\ {}_{Z}E^{A+1}. $$

The first of these (for example) is to be read: a neutron enters a nucleus (Z, A) and an alpha-particle comes out, leaving behind a nucleus $(Z - 2, A - 3)$. There was a similar (not identical) rule setting a limit on the changes of atomic number and mass-number suffered by radioactive bodies. Every radioactive nucleus emits either a positive

electron or a negative electron, or an alpha-particle; the corresponding changes in Z are -1, $+1$, -2, those in A are 0, 0, 4.

Now it is high time that we get on to uranium. This, like thorium, was one of the elements that Fermi exposed to slow neutrons, and to which he observed that something was happening. As he continued these researches, and as the great institutes of nuclear physics of Hahn in Berlin and the Joliots in Paris followed suit, it became evident that a great deal was happening. With nearly all of the other elements, there occurred just one of the reactions which I symbolized above, or maybe one reaction with some of the nuclei and another with others. Sometimes the reaction would lead to a stable nucleus-type; in such a case, when the neutron-bombardment ceased all the excitement was instantly over. Sometimes it would lead to a radioactive nucleus-type; in such a case, the radioactivity would continue after the bombardment ceased, but it would steadily die away to nothing, in accordance with the well-known law. But with uranium and thorium there was a swarm of radioactive products, so many that to this day they have not all been separated and identified. Moreover, some of these were descendants of others, for after the bombardment ceased they grew in strength for a while before declining. All of them were emitters of electrons, and the electrons (in every case in which their sign is known) were *negative*.

Owing to the theorem which has just been stated, it was taken for granted that the immediate effect of the neutron entering the nucleus of uranium was to provoke one of the three reactions which I lately listed. Of these the one most commonly assumed was the reaction,

$$_{92}\mathrm{U}^{238}\ (n, \phi)\ _{92}\mathrm{U}^{239}$$

the so-called "reaction of pure neutron-capture"—called pure, because no massive particle goes forth. I mention it here because it is still believed in, and we shall meet it later. Uranium 239 is radioactive, and some of the other radioactive products were believed to be direct or indirect descendants of it. Well, every one of the radioactive substances resulting from the reaction or reactions $U(n)$—for so I will symbolize them in general—is an emitter of *negative* electrons. Therefore each has a greater positive charge on its nucleus than does its predecessor; therefore by this theory, each descendant from U^{239} must have a greater positive charge than the $92e$ of the uranium nucleus. But uranium is the final element of the periodic table; therefore by this theory the radioactive bodies in question had to be isotopes of new elements beyond uranium.

These so-called "trans-uranic elements" were for several years the principal study of Hahn and Meitner and their colleagues at the great institute in Berlin-Dahlem. The groups at Paris and at Rome contributed also—not very much, but enough to signify their full adhesion to these concepts. Other physicists scarcely entered the field, but had the fullest reliance in views sustained by such authorities. Yet now, the trans-uranic elements are gone! This is regrettable, because it was pleasant to think that human artifice had succeeded even in lengthening the list of the elements. It is regrettable for the chemists especially, because they were looking forward to getting information about the chemical properties of elements beyond 92. Whether on balance there is regret among physicists I doubt, because the knowledge that has replaced the trans-uranic elements seems even more spectacular than they did. Let us see how this knowledge was attained.

Some time in 1938, Hahn observed that three of the radioactive substances resulting from the exposure of uranium to neutrons had some of the chemical properties of barium—enough to follow barium in certain of the distinctive precipitations which are known to chemists. Now this is a statement which is true of radium. Hahn assumed that he had three new isotopes of radium, and this was entirely natural, for two reasons. First, radium and its isotopes already known are all radioactive, suggesting that any which remained to be discovered should also be so; and second, the atomic number of radium is 88, so that radium isotopes could conceivably come into being through the reaction $U(n, \alpha)$ followed by the spontaneous emission of an alpha-particle from the residue. Yet (and this is the fact which came out on the 6th of January 1939) these substances were much too much like barium! When Hahn and Strassmann used some of the procedures which separate radium from barium, the novel substances declined to be separated. In a typical experiment, one of them together with some well-known isotope of radium would be introduced into a solution of some salt of barium. Fractional crystallization being performed, it was found as usual that the relative concentration of the radium isotope was greatly changed in the first-to-be-formed of the crystals; but not the relative concentration of the new substance, which entered into the crystals in just the same proportion as the barium itself.[2]

There are people who have revolutionary and false ideas about questions of science, and who irritate the scientists by their overconfident, their often arrogant ways of offering those ideas to the world.

[2] The salts were the bromide and the chromate (perhaps also the chloride and carbonate) of barium; the isotopes of radium were ThX and $MsTh_1$.

Listen now to men of science having a revolutionary and true idea, and expressing it in a befitting way:

"Now we must speak of some more recent investigations, which we publish only with hesitation because of the strange results. . . . We come to the conclusion: our 'radium isotopes' have the properties of barium; as chemists, we really ought to say that these new substances are barium, not radium. . . . As chemists, we ought to use the symbols Ba and La and Ce where we have been using Ra and Ac and Th. But as 'nuclear chemists' closely associated with physics, we cannot yet bring ourselves to make this leap, in contradiction to all previous lessons of nuclear physics. Perhaps, after all, our results have been rendered deceptive by a series of strange accidents. . . ."

Here I ought to mention another famous group of nuclear physicists who at an earlier date might have taken the leap, but recoiled before it so vehemently that they could not even bring themselves to mention in print the possibility of making it. These were the physicists of the Institut du Radium at Paris: Irene Curie and Savitch discovered in early 1938 that one of the products of $U(n)$ was indistinguishable, by all the tests that they applied, from the rare-earth element lanthanum ($Z = 57$). Later on they said that at a certain moment they had envisaged what we now call the fission of the uranium nucleus, but had preferred to believe that they had before them one of the trans-uranic elements resembling lanthanum more closely than anyone as yet had foreseen.

Now we come on to the middle of January 1939, and I must introduce the grand idea which with the force and suddenness of revelation burst upon several people far apart in the world, as soon as they heard of the experiments of Hahn and Strassmann and of the leap which these two had dared to envisage and publish if not quite to take.

Let us be audacious enough to take the leap, and let us further imagine that after the entry of the neutron the nucleus divides itself into two pieces or "fragments" of which one shall be barium. I must say directly that this assumption is more specific than need be, and that the same conclusions would be reached if we assumed that one of the fragments is some other element close to barium in the Periodic Table. It will be simpler, however, to be definite: let us assume barium, and for still greater definiteness let us suppose that the isotopes concerned are 238 of uranium and 139 of barium. The neutron, then supposedly enters a nucleus ${}_{92}U^{238}$ and with it forms the transitory "compound nucleus" ${}_{92}U^{239}$, and from this there splits off a nucleus ${}_{56}Ba^{139}$. What is left behind must be (if in a single piece) the nucleus

of which the charge added to 56*e* makes 92*e*, and of which the mass-number added to 139 makes 239—that is to say, the nucleus $_{36}Kr^{100}$.

This is an example of the type of process which has been named by borrowing the word "fission" from biology. The biologists seem not to have found a specific verb to correspond (I am told that they use "divide") and the physicists have had no better inspiration. The dictionaries, however, authorize the use of "fissure" as a verb both transitive and intransitive, and I will henceforth so use it.[3]

Now a difficulty looms up, or rather what seems to be a difficulty but is really a great advantage, for the grandeur of the idea depends on it. Mass-numbers are only approximations to true masses, and the true mass of the nucleus U^{239} is greater than the sum of the masses of Ba^{139} and Kr^{100}. There is a superfluity of mass, and by classical ideas this superfluity might have to vanish, which would indeed be a stumbling-block. However, that stumbling block does not exist, because of something I have now to introduce. *It is the rest-mass, in the sense of relativity, of* U^{239} *which exceeds the sum of the rest-masses of* Ba^{139} *and* Kr^{100}. Now U^{239} before the explosion is practically at rest, but we are not obliged to make the same assumption about the fragments, and in fact we can assume that *the fragments fly apart at just such speeds that their relativistic increase of mass with speed brings up the sum of their masses to exactly the right value.* If so, their kinetic energies must be 50 to 100 Mev apiece. These on the nuclear scale are immense amounts of kinetic energy, and particles possessing it must be easy to isolate and easy to detect. This is why the idea is a grand one.

As it might occur to some reader to go to the tables of constants and look up the mass-values of U^{239} and Ba^{139} and Kr^{100}, I must say at once that he will not find them. Generally speaking, the mass-spectrograph cannot be used on radioactive and unstable atoms because one cannot get enough of them together for the experiment (exception being made for very long-lived ones like U^{238} and Th^{232}). All those three belong in that category, and therefore we have to estimate their masses by extrapolation from those of stable isotopes. The extrapolations for Ba^{139} and U^{239} are so small that the uncertainty is trivial, but Kr^{100} is no less than fourteen units heavier than the heaviest stable isotope of krypton, and this is serious. However, one is not so much concerned about conceivable defects in grand ideas when the ideas have already done their work by leading with success to grand experiments. I lay emphasis again, for a reason later to appear, on the extent to which Kr^{100} is out of line with the stable krypton isotopes; and now we pass to the experiments.

[3] I am indebted to Dr. Elizabeth Patterson of Bryn Mawr College for this solution.

There are actually two grand experiments, which I tried to distinguish above in a sentence by saying that the energetic particles must be *easy to isolate* and *easy to detect.* "Isolate" is not a very happy word: the fact is, that if so energetic they must be able to fly right out of the bombarded sheet of uranium (unless they start too deep beneath its surface)—thus, if some sort of a collector is placed across from the uranium and not too far away, they must assemble on it and there they should be found together with all their descendants. Joliot published this experiment before the end of January. He found radioactive substances on his collector even when more than two centimetres of air [4] had separated it throughout from the uranium.

The experiment has been performed by many, some introducing new refinements. Meitner and Frisch for instance used a bowl of water for collector, and then could concentrate the radioactive bodies by letting the water evaporate, or by precipitating various salts which in advance they had dissolved in it. This last is the chief technique for finding out the chemical nature of the radioactive products, to wit, the elements of which they are unstable isotopes; but we have not space for entering into the details of the technique, already practiced these five years. Glasoe, McMillan and others modified the method by piling very thin foils of very light substances—aluminium, filter-paper, cigarette-paper—on the uranium. Some of the radioactive matter is found embedded in each of the first few foils, and one may study thus their "distribution-in-range," an almost self-explanatory term. In McMillan's experiment the utmost perceptible range was slightly above 2.2 cm of air.

Already in the first experiment Joliot observed that in respect of its decay in time, the radioactivity on the collector was very like that remaining on the uranium. Later more accurate work has merely strengthened that conclusion, and Segré in particular affirms that out of many radioactive bodies there are only two which are found in the bombarded uranium itself and not on the distant collector also. On the distant collector there are found, in particular, the substances once classed as "trans-uranic elements." This is very important, for in the theory of the trans-uranic elements there occurs no stage in which the fragments of the uranium nucleus (or any other) are thrown apart with so tremendous energies. Were these elements trans-uranic, they should not be able at all to escape from the bombarded uranium

[4] To give the thickness of air (of the density corresponding to 15° C. and one atmosphere of pressure) which can just be traversed by a charged particle is the ordinary way of stating the "penetrating power" of the particle. Often some other substance than air is used in the tests; it is then not the actual thickness of the substance, but the "air-equivalent" thereof, which is ordinarily stated. Joliot appears to have used actual air in the experiments.

target. When in defiance of this the radioactivity crossed over to the collector, the trans-uranic elements were doomed.

In these experiments, then, the fragments of the initial explosion are found *en masse* together with their descendants upon the distant plate. In those of the other grand type, they are detected each by itself *en route*. Being charged particles of great momentum, they cleave through any gas in nearly linear paths, along which quantities of ions stay behind.[5] The Wilson chamber may be used to make these visible, and has indeed already been so used (by Joliot, and by Corson and Thornton); but another device gave the first and as yet most instructive results. This is the ionization-chamber equipped with linear amplifier and oscillograph. In the first, the ions due to the passage of a single particle are drawn to a collector and their charges united; in the second, the united charge is multiplied by a large fixed factor; in the third, the multiplied charge produces a sharp sidewise motion of the oscillograph-beam and the spot which this last produces. On the photographic plate the moving spot produces a line, the length of which is measured. Instances of these lines appear in Fig. 1.

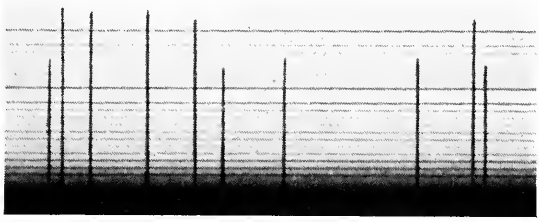


Fig. 1—Records of fission-fragments obtained with ionization chamber, linear amplifier, and oscillograph. The short lines due to alpha-particles are lost in the hazy dark band beneath. (Courtesy of J. R. Dunning)

Uranium is a spontaneous emitter of alpha-rays (this is how the radioactivity of U^{238} and U^{234} is manifest) and so the apparatus will show "kicks" even when neutrons are absent. This is an advantage really, since when the neutrons are admitted and the kicks due to the fragments appear they are so much the larger that there is no danger of confusing them with alpha-particle kicks, while these last may be pressed into service for calibrating the device. The calibration reposes upon a theorem of very great value in physics: viz., the (average) amount of energy expended by a fast charged particle in producing an ion-pair is fixed and constant, whatever the charge and mass and

[5] This is correct whether they travel as isolated nuclei, or are attended by some though not the full quota of orbital electrons which would environ them were they already the nuclei of completed atoms. Capture of electrons along the course is almost certain (it has been proved to occur with alpha-particles).

speed of the particle. So, the ratio of the kicks caused respectively by a fragment and an α-particle is the ratio of their initial energies, provided the chamber is so "deep" that they both run their courses completely to the end in the gas thereof. If on the other hand the chamber is so shallow or "thin" that fragment and α-particle shoot across it and only a small part of the total course of each is comprised within it, then the ratio of the kicks may be the ratio of the densities-of-ionization along the two tracks. Both the initial energy and the density-of-ionization are known for the α-particles, permitting the calibration. Also the constant value of the energy-expended-per-ion-pair is known (it is about 30 ev.) so that if the experimenter can measure the actual amount of charge set free in his ionization-chamber he need not bother with the α-particles.[6] In Fig. 1, by the way, the alpha-particle tracks are quite lost in the black band of the "background."

The second grand experiment, then, consisted in showing that when the neutrons were falling upon the uranium, there instantly appeared among the smallish kicks due to the α-particles others which were much greater—ten- and twenty-fold greater. This was done in four places [7] at least in America in the closing days of January 1939; in Copenhagen, however, a fortnight earlier.

The greatness of the kicks when the ionization-chamber is deep signifies the greatness of the initial energies of the fragments: I shall presently quote the latest data of these. But when the chamber is thin, the kicks due to the fragments again stand out very much over those due to the α-particles; and this signifies that the ionization-density along the fragment-tracks is great. (Take note, by the way, that one and the same chamber may be thin or thick, according as the density of the gas within is low or high—a very convenient fact.) The fragments, then, not only have remarkably great energy to start with, but also spend it at a remarkable rate in ionization along their courses. The course or "range" of a fragment must therefore be much shorter than would be that of an α-particle of the same energy. This is a verifiable fact, the ranges being easily measured by this method. We have just seen how Joliot was able to estimate them earlier, wiping out by this observation the possibility that each of the great kicks may be due to many α-particles starting off together. From the ion-

[6] All this is contingent upon the ions being completely gathered in by the collector of the ionization-chamber before any serious fraction of them is annulled by recombination, or (failing that) upon the loss by recombination being the same in proportion for α-particles and for fragments. Owing to the (unprecedently) great density of ion-pairs along the tracks of the fragments, this is by no means sure.

[7] New York (Columbia), Baltimore (Johns Hopkins), Washington (Carnegie Institution), Berkeley (University of California). In these cases the suggestion originated with Fermi.

ization-density and the range and the energy all taken together, it may be inferred that in both charge and mass these particles much exceed the α-particles; but here, better data and fuller theory are urgently required.[8]

Now we will consider the energies of the particles according to the data of Kanner and Barschall of Princeton.

If the immediate products of the fission are really just a pair of fragments nearly but not quite identical, we may expect a distribution-

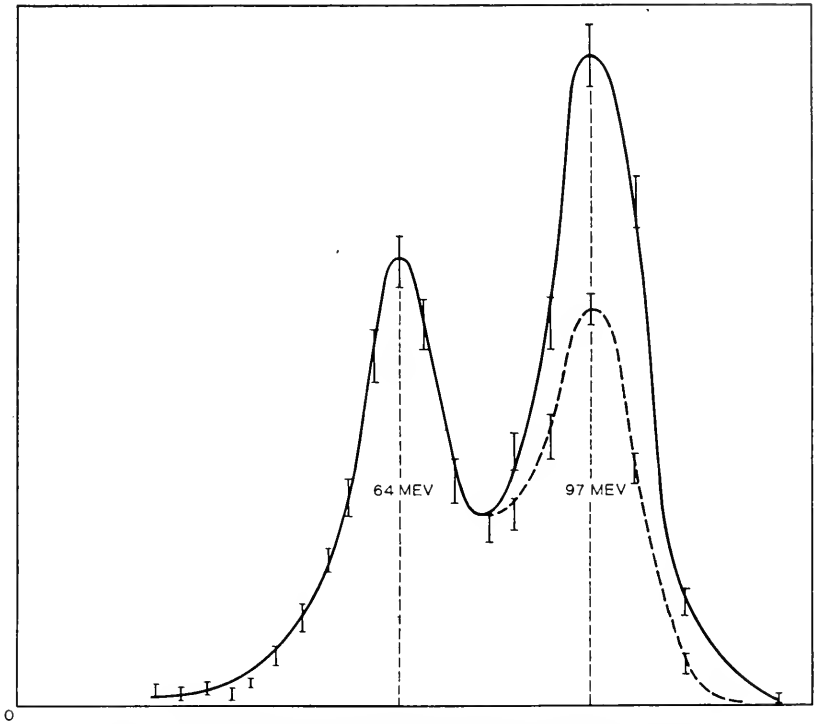


Fig. 2—Distribution-in-energy of fission-fragments of uranium. (Kanner and Barschall; *Physical Review*)

in-energy curve with two sharp peaks. If different fissions result in different fragment-pairs, the peaks must be broadened. If three or more particles are formed at a fission, there should be a broad continuous distribution of energies. This third of the possibilities is well excluded by the curve of Fig. 2; it remains to be seen whether the breadths of the humps speak for the second over the first.

[8] The inferring of charge and mass from energy, range and ionization-density is much practiced in the field of cosmic-ray research, in which, however, the particles usually have charge e and masses between the proton-mass and the electron-mass.

In giving the data for Fig. 2, as my words have implied, only a single fragment from each fission escaped into the gas of the ionization-chamber; this was arranged by laying down a very thin film of uranium upon a thick sheet of another metal. Figure 3 was obtained by laying down the uranium film upon a foil so very thin, that from most of the fissions both of the fragments entered into the gas. The great peak of Fig. 3 therefore indicates the sum of the energies of the fission-fragments, the peaks of Fig. 2 the components of that sum. Making al-

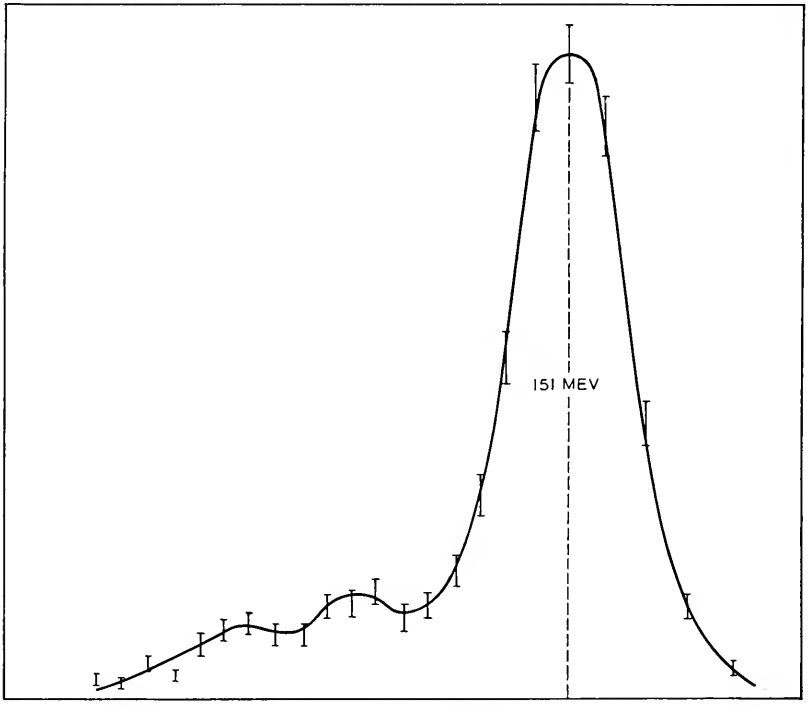


Fig. 3—Distribution-in-energy of pairs of fission-fragments from uranium. (Kanner and Barschall)

lowance for the average energy-loss suffered by the fragments in passing through solid matter before they escape to the gas, Kanner and Barschall decide on 159 Mev. for the sum, 98 and 65 Mev. for the components: the discrepancy between 159 and (98 + 65) lies within the uncertainty of experiment. By the law of conservation of momentum, the ratio of the component energies is the ratio of the fragment-masses; if one of these is about 100, the other is therefore about 150—and the uncertainty implied by "about" is broad enough to permit the hypotheses which we have made and are to make about the nature

of the fragments. The Columbia school has done the same experiment, with like results.

Another way of ascertaining the energy released by fission was adopted by Henderson of Princeton; it is the oldest and most unimpeachable of all the methods of measuring energy, for he determined the rate at which heat was being developed in a uranium target and a container surrounding it while the fissuring was going on. His value was 175 Mev. per fission, with an uncertainty of some ten per cent. As some of this energy belonged not to the fragments but to the electrons emitted after the fission, the agreement is better than passable.

Now we come back to the question of the masses of the initial fragment-pair; and I will develop a second consequence of these masses, entirely different from the first. I revert to the use of mass-numbers, since the corrections needed for converting these into actual masses have not the slightest bearing on the point which is now to occupy us.

If the members of the initial fragment-pair are Ba^{139} and Kr^{100}, then the second of these two is fourteen units heavier than the heaviest stable isotope of $_{36}Kr$. It is therefore much too massive for its charge. This suggests that it may be able to shed neutrons, and so bring down its weight to the highest value compatible with its charge. But one may also say that Kr^{100} is too feebly charged for its mass. One has to go no fewer than six steps along the periodic table—to $_{42}Mo$—to find an element with a stable isotope of mass-number 100. Yet there is nothing to prevent us from assuming that the nucleus $_{36}Kr^{100}$ may shoot out six negative electrons, and so increase its charge to the minimum value compatible with its mass. The six might come out *seriatim*, in which case there would be a chain of six radioactive substances comprising all the elements from $_{36}Kr$ to $_{41}Cb$. Again, the nucleus might conceivably eject any number of neutrons under fourteen and some number of negative electrons under six, arriving at a sort of compromise pair of values of mass and charge compatible one with the other. One guesses already a mighty number of possible radioactive bodies resulting from the fission!

But now let us discard the assumption that Ba^{139} and Kr^{100} are the actual fragments of the fission, replacing these with any two nuclei which (*a*) lie in the middle region of the Periodic Table and (*b*) have atomic numbers adding up to 92 and mass-numbers adding up to 239. What, then, will happen to our two inferences from the masses? Essentially, *nothing*. Whichever such pair we take, one at least of its members must be too heavy for its charge and too feebly charged for its mass. (With most conceivable pairs, this will be true of both the members!) This derives from one of the fundamental facts of nuclear

physics: the fact that when mass-number is plotted against atomic number, the points representing the stable nuclei cluster about a concave-upward curve. Moreover, whichever pair of fragments we assume, there will be a superfluity of rest-mass which will manifest itself in a high kinetic energy of the two fragments. This derives from another fundamental fact: when the percentage of excess of mass-number over true mass is plotted against the mass-number (or for that matter the atomic number) the points representing the nuclei cluster along an upward-trending curve.

And so, the kinetic energy of the fragments and the facts of the emerging negative electrons tell us neither which is the initial fragment-pair, nor even whether the initial pair is in all cases the same! Can these questions be answered out of the study of radioactive substances? Some of these we can indeed exclude by observing that they grow out of others; but as to these others, we shall never be able to exclude the possibility that they grow out of still others so short-lived, as to be quite unidentifiable. The half-period of a radioactive substance must be appreciable, if the substance is to be detected and its chemical character recognized; and "appreciable" thus far has signified, among the products of fission, "several seconds or more." It is true that certain *tours de force*, whereby much shorter half-periods have been measured among the natural radioactive bodies, have not yet been applied to the fission-products (so far as publications tell); they might prove workable.[10]

Thus it may be necessary for the nonce to lay aside the problem of deciding which is the true initial fragment-pair (or pairs) and be contented with identifying as many as possible among the radioactive substances and tracing their interrelations. Of these—hereafter to be called "the fission-products"—there is indeed a multitude. Among them, chemical elements have been recognized as follows: $_{34}$Se, $_{35}$Br, $_{36}$Kr, $_{37}$Rb, $_{38}$Sr, $_{39}$Y, $_{40}$Zr, $_{41}$Cb, $_{42}$Mo, $_{52}$Te, $_{53}$I, $_{54}$Xe, $_{55}$Cs, $_{56}$Ba, and $_{57}$La. Yet by counting these one does not count all of the distinguishable products; experimenters say that they can tell apart three isotopes of barium, three of strontium, four of iodine and no fewer than seven of tellurium! Some of these agree in their half-periods with radioactive isotopes of those same elements already formed by the older ways of transformation, and frequently we can thus identify their mass-numbers with a fair degree of certainty. (Ba^{139}, which I introduced into the hypothetical reaction of page 272, is such a one; but we shall see

[10] I refer particularly to the use of a rapidly-turning wheel to carry a target swiftly from a place where it is under bombardment (or receiving a deposit of radioactive nuclei) to another place where it is opposite a detector; and the measurement of the radioactivity of a beam of fast-flying nuclei at various points along the beam (the method applied by Jacobsen to RaC').

that certainly it is not always and possibly it is never an initial fragment.) Others were unknown till 1939.

So numerous are these and the other fission-products still unrecognized, that the "decay-curve" for a piece of bombarded uranium or for the deposit on a nearby collector, due to all of them conjointly, looks like the resultant of contributions practically limitless in number and with a random distribution of half-periods. Not only is this also true when neutrons impinge on thorium, but the curves for the two elements cannot be told apart! Only after chemical separations have been made can individual half-periods be sorted out from among the welter; and if there are some characteristic differences between the results of the fission of uranium and those of the fission of thorium, they have not yet been proved.

Special interest attaches to the fission-products which are gaseous. They can be separated physically from the rest: the fission-products are received or dissolved into water (or indeed the uranium may be exposed to neutrons while in aqueous solution) and through the water a stream of air is bubbled, which takes along these particular ones to distant points in the system of tubing where they and their descendants can be studied. They cannot themselves be identified, but among their descendants are found (radioactive) isotopes of $_{37}$Rb and $_{55}$Cs; therefore the gases comprise unstable isotopes of krypton ($_{36}$Kr) and xenon ($_{54}$Xe). Could these be initial fragments of various types of fission? If so, their mates are $_{56}$Ba and $_{38}$Sr. Now, barium and strontium are found indeed among the fission-products, which seems to sustain this idea. But barium and strontium may also be the immediate descendants of the caesium and the rubidium aforesaid. This alternative idea is testable; and according to Hahn and Strassmann, two among the three barium isotopes (Ba^{139} being one of the two) are surely descendants of caesium, while the third may be an initial fragment.

Many other such "genetic" relationships have been published, but it would be lengthy and might be premature to quote them. I will mention at least that several sequences have been traced by Abelson in greater or less detail among the many fission-products which are isotopes of the three consecutive elements $_{51}$Sb, $_{52}$Te, and $_{53}$I. A special interest attaches to one of these bodies, the "77-hour tellurium"; for it has been identified as tellurium not only by its chemical properties but also by its X-rays. Let us pause to consider this.

The ordinary way of evoking an X-ray spectrum is to use the element in question as the target, or a constituent of the target, of an X-ray tube. This means that the atoms are excited by projecting electrons

against them. They may also be excited by projecting photons against them, and this is sometimes done. Both of these ways are completely out of the question as yet with any artificial radioactive substance, for the greatest amount yet produced of any of these is so small that if its atoms were placed in a target, the hits made upon them by electrons or photons projected in streams of any feasible strength would not be numerous enough to produce detectable X-rays. If, however, the necessary photons proceed from the nuclei of the atoms themselves, then the whole situation is changed, because now the efficiency of excitation is so great. Such is the case with many of the natural radioactive substances, and now also (it appears) with the "77-hour tellurium." Excited presumably by photons proceeding from their nuclei,[11] the atoms emit X-rays, and these have been found (by Abelson in Berkeley, by Feather and Bretscher in England) to be the characteristic rays of the K-series of iodine. "Iodine" here is *not* a misprint for tellurium! When the nucleus is radioactive by virtue of the emission of an electron, the photon (if any) leaves after the electron is gone, by which time the atom is already an atom of the daughter-substance.[12]

Now we take up the yield of the fission-process: how does it depend on the energy of the incident neutrons?

Here uranium sets itself apart from the two other fissurable elements. Thorium and protactinium respond to fast neutrons only, uranium both to slow and to fast (but not to intermediate) neutrons. It is, however, believed that with uranium, one isotope is sensitive only to fast and another both to fast and to slow (or possibly only to slow). There are good theoretical grounds for this belief, and also for choosing the respective isotopes; but as yet there is not the certainty to be expected from some future and probably imminent experiment on separated isotopes.[13] Accepting nevertheless the current belief, we sup-

[11] Another mode of excitation is now known: an electron may fall into a nucleus, and by quitting its place in the orbital electron-family create the condition for the emission of an X-ray photon. Whether this or the other or both be the mode of excitation of the 77-hour tellurium is not yet certainly known.

[12] As this is likely to cause confusion, I emphasize that when the chemical separation is made, the atoms which have not yet emitted nuclear electrons are still tellurium atoms, and when they manifest themselves by that emission it appears among the tellurium; from then on they are iodine atoms among the tellurium, but no longer manifest themselves except through these X-rays. One may wonder whether the "transuranic elements," to which the 77-hour body was formerly thought to belong, would have been discredited if this measurement on the X-rays had been made earlier. Well, the measurement *was* earlier made (though not so precisely) and the rays were interpreted as characteristic X-rays of the L-series of a transuranic element. It is hard to make a guess as to whether further and better measurements would have destroyed this possibility.

[13] As these pages start for the press I am authorized to say that the separation has been achieved by Nier and the experiment performed by Dunning, Booth and von Grosse. The "light fraction," consisting of U^{235} with a small proportion of the very rare isotope U^{234} is definitely sensitive to slow neutrons; U^{238} is definitely *not*.

pose that of four known types of nuclei three (Th^{232}, Pa^{231}, U^{238}) are fissurable by fast neutrons only, one (U^{235}) by slow neutrons and probably also by fast. The mass-numbers just given are those of the nuclei awaiting the invading neutrons. If one prefers (as many do) to think of the transient composite nuclei formed by the neutron-invasions, one must write Th^{233}, Pa^{232}, U^{239} and U^{236}.

To speak of "fast" neutrons is vague, but not much vaguer than the state of knowledge, which as yet is rudimentary. Fission has been detected of thorium at neutron-energy of about 2 Mev, of protactinium at about one Mev, of uranium at about 0.5 Mev. It thus appears that the "threshold," or least contribution of energy demanded for fission, declines as the end of the Periodic Table is approached; and this seems natural. (Remember always that even with the fastest neutrons ever used, the contribution is very small compared with the energy released.) Values given in the literature for the "cross-section for fission by fast neutrons" include: $0.5 \cdot 10^{-24} cm^2$ and $0.1 \cdot 10^{-24}$ for uranium and for thorium bombarded by 2.4-Mev neutrons (Princeton) and $0.1 \cdot 10^{-24}$ for uranium bombarded by the "RnBe" neutrons.[14]

If between the source of neutrons and a target of uranium a screen of paraffin or water is inserted, the fissions become more abundant; but if now between the paraffin or water and the uranium a shield of cadmium is placed, the fissions become very rare. Now, paraffin and water convert fast neutrons into slow or "thermal" ones,[15] and cadmium is a very efficient absorbent for slow neutrons. We recognize, therefore, a specific effect of slow neutrons, peculiar to uranium. "Slow" or "thermal" signify in this usage: having kinetic energies of the very modest magnitudes possessed by molecules of air (or anything else) at ordinary temperatures: fractions of one electron-volt, and rarely more. Clearly then it is not the energy of motion of the neutron which is the insignificant spark setting off the mighty explosion; it is the mere presence of the neutron within the nuclear system.

A beam of slow neutrons falling upon a thin uranium layer produces many more fissions than does a fast-neutron beam of identical strength. Twenty-to-one was the ratio of yields found by the Columbia school, in the same experiment as gave them the value $0.1 \cdot 10^{-24}$ for the cross-section for fission by the fast "RnBe" neutrons. If, however, we put $2 \cdot 10^{-24}$ for the cross-section appropriate to the thermal neutrons, we

[14] Cross-section for fission, σ_f, is so defined that if N neutrons strike a thin layer comprising M nuclei per unit area, $MN\sigma_f$ fissions occur.—The "RnBe" neutrons, *viz.*, those released when α-particles from radon and its descendants impinge on beryllium, have a very broad energy-range extending at least from 14 Mev indefinitely downward (cf. Dunning, *Phys. Rev.* **45**, 586; 1934).

[15] The neutrons lose their great kinetic energies in repeated elastic impacts with hydrogen nuclei.

are in effect assuming that all the nuclei in the layer are equally liable to being fissured by these. To remain faithful to the well-grounded assumption that only the nuclei U^{235} are liable thus, we must multiply by 140, since only one nucleus in one hundred and forty is of this isotope.[16] The resulting value is large-sized for the nuclear scale, though not unprecedented: there are elements which absorb thermal neutrons so voraciously (without however suffering fission) that the cross-section for absorption is found to be hundreds of times more extensive.

Now in conclusion we turn to the particles other than nuclei, which go forth into space when or after the fission occurs. These comprise photons, electrons, and newborn or "secondary" neutrons; and the last are by far the most sensational.

Of the electrons, almost all has been said that should find place in this account. I recall that by virtue of the second argument from the masses (page 279) the nuclei of the fission-products should go from instability over to stability by emitting electrons which are negative. Observation shows that the emitted electrons are negative indeed (and yet there must be many among the products for which the sign has not been ascertained). Unstable nuclei emitting positive electrons are not at all unknown; indeed they are formed in many transmutations; their absence from among the fission-products is therefore significant. Many of the electrons coming forth are of "secondary origin," i.e. released by photons from the electron-families of the atoms. When classified with the many radioactive bodies formed by other modes of transmutation, some of the fission-products are found to be identical with some of those others, and the rest are in no wise peculiar.

Of the photons, some are X-ray photons engendered as I have recently described (page 282). Others are of the gamma-ray type, *i.e.*, they spring from unstable nuclei among the fission-products. Their existence not being in the least surprising, they have in the main been left for future study.

Coming now to the secondary neutrons, I will begin by dividing them into the "delayed" and the "instantaneous." The former come forth and are detected during an appreciable time—a few seconds up to a few minutes—after the fissions cease. Here then are radioactive bodies, of which the radioactivity consists in the emission of neutrons! Nothing of the sort had ever been known, and the discovery (made at the Carnegie Institution of Washington) created a sensation. In number they are much fewer than the "instantaneous" neutrons, define

[16] The figure is from Nier, *Phys. Rev.* **55**, 150 (1939), who gives 139 ± 1 as the abundance-ratio of U^{238} and U^{235}.

(for the present) as those which come out within a few thousandths of a second of the moment of fission. The ratio, according to the Columbia school, is about one to sixty. Delayed electrons and delayed photons have also been observed. Most observations on secondary neutrons are made while the target is being bombarded, and therefore relate to a mixture of the instantaneous with a small proportion of the delayed.

In energy the secondary neutrons differ greatly from the primary, a remarkable contrast! This is shown by several neat and pretty experiments, in which the secondaries manifest themselves by acting on detectors which cannot perceive the primaries at all. Thus if the primaries are thermal neutrons, an expansion-chamber or an ionization-chamber full of gas can be set among them without showing any sign of them,[17] since they cannot strike hard blows against the molecules therein. Let, however, a piece of uranium be set nearby, and the chamber will show dense trains of ions, produced by nuclei struck very hard and driven out of the molecules by neutrons which are fast. In this manner Halban and Joliot and Kowarski in Paris detected secondaries running in energy up to 11 Mev and beyond, while Zinn and Szilard of Columbia mapped the energy-spectrum up to 3.5 Mev. But also there are detectors able to discriminate between fast and faster neutrons: e.g. phosphorus, which becomes radioactive when bombarded by neutrons if, but only if, these have energy greater than 2 Mev. Dode and others in Paris prepared a source producing neutrons of energy one Mev; placed it next to a uranium target; surrounded source and target with a tank of liquid carbon disulphide, in which phosphorus was dissolved; and the liquid grew radioactive.

But how many neutrons are released per fission? This is a question of singular and perhaps of devastating importance, as will presently appear.

The obvious way to answer it seems to be that elected by Zinn and Szilard, who measured the number of fissions and also estimated, from the number of recoiling nuclei observed in their expansion-chamber, that of the secondary neutrons. Most of the trials have been made by a different method, in which all of the secondary neutrons are reduced to thermal energies before they are detected. (Incidentally there is the advantage, that if neutrons are released with low initial energies they will be counted by this way but not by the other.)

In this more customary method, the neutron-source and target are close together in the midst of a great tank of water, as large as can

[17] Except that if nitrogen is contained in the chamber, the thermal neutrons will react with the nitrogen nuclei so as to release protons (Zinn and Szilard).

conveniently be made. Paraffin may surround [18] the target and the source, to slow down the primary neutrons; or the water itself may perform this office. Again, the target may be diffused throughout the entire water-mass, in the form of a soluble salt of uranium. The detector is a substance becoming radioactive when exposed to slow neutrons. It may itself be spread throughout the water in the form of a soluble salt, or it may be in the form of a thin foil which can be moved from place to place in the water. In the former case, the water is thoroughly stirred after the exposure is over, and then a sample is taken, the activity of which is a measure of the average density—and therefore of the total quantity—of thermal neutrons in the entire tank during the exposure. In the latter case, the foil is used for mapping out the density of thermal neutrons in the water as function of the distance r from the target in the middle, and what is usually plotted is the "Ir^2 curve," I standing for the strength of the activity of the foil.

The total quantity of thermal neutrons, existing at any moment dispersed throughout the water, is greater in the presence of the uranium than in the absence thereof [19] (Anderson, Fermi and Hanstein); this is the simplest proof of the fundamental result. When the Ir^2 curves are compared, it is found that the presence of the uranium lowers the curve in the close neighborhood (within 13 or 14 cm) of the neutron-source, but raises it further out. Presumably this is because the uranium swallows up the slow primary neutrons, and those which it gives out in exchange are themselves not slow until they have gone a long way onward in the water. In tanks of sufficient size, the increase farther out more than balances the diminution nearer in, and the total quantity of thermal neutrons is augmented by the presence of uranium (Halban, Joliot and Kowarski); this agrees with the other result. It is therefore established what when the primaries are slow, the fission-process delivers more neutrons than it consumes. The same holds true when the primaries are fast, for when a beam of RnBe neutrons is sent through a plate of uranium oxide the detector beyond reveals a greater quantity of rapid neutrons than when the plate is absent (Haenny and Rosenberg, experimenting with a plate 8 cm thick).

How many neutrons then emerge, for every one which is spent in producing a fission? This is a remarkably difficult question to put to

[18] It is not necessary that the "slowing-down" substance be actually between the target and the source, since slowed-down neutrons come out of it in all directions.

[19] To make the situations strictly comparable, the uranium is replaced in the control experiment by some substance possessing an equal absorbing-power for neutrons, but not liable to fission.

the test of experiment. About all that the several answers have in common is, that *more neutrons emerge than are spent.* Zinn and Szilard say, two or three times as many; Anderson, Fermi and Szilard say, between one and two; the Paris school, three or four. A yet higher value (eight) published from Paris seems to comprise some "tertiary" neutrons produced by the secondaries.

But if every fission produces a fresh neutron to replace the one which caused it, and then some extras in addition, must we not anticipate a self-sustaining, nay even a self-amplifying effect? Must we not fear, in fact, a cataclysmic explosion?

Were anything of the sort to happen, we may take it for granted that the world would know of it, though in all probability the experimenter would not himself survive to report it. Evidently then it has not happened, and there must be a brake or brakes in Nature which impede the slide toward the catastrophe, and have thus far averted it. In other words, there must be ways in which neutrons are made harmless by some innocuous type of capture, before they ever produce a fission.

Some of these other ways are known already. If the uranium is mixed with other elements—as, in Nature, it invariably is—the nuclei of these can take up some of the neutrons. Whether the composite nuclei so formed are stable or radioactive is in this connection not important; they give no neutrons out in exchange for the ones absorbed, and so the chain is broken. But if all other elements are carefully extracted, do any brakes remain?

Two surely do, and one is the fact that the newborn neutrons are rapid, and cannot be efficacious as agents of fission until they are slowed down to thermal energies. In pure uranium the slowing-down can only be extremely gradual, so unfavorable is the huge mass-ratio—238 to 1—for the energy-transfer in the elastic impacts. Yet if the volume of purified metal were great enough, this brake would relax. Thus the durability of small-size pieces of uranium made chemically pure, well attested as it is, is not by itself a proof that much larger pieces would be safe. Those who are trying to approach the catastrophe, while hoping not to provoke it, are engaged in piling up uranium in greater and greater masses.

The other brake is supplied by the "reaction of pure neutron-capture," which I mentioned on an early page (p. 270). Every now and then, when a neutron enters a nucleus of uranium, the composite nucleus finds itself able to live on without fissure. It survives for a time, then emits a negative electron of energy trivial compared with

fission-energies, and relapses into permanent stability. This is especially likely to happen when the neutron-energy is about 10 ev. Suppose then a volume of pure uranium so great, that the rapid neutrons released within it can make collisions numerous enough to bring their energies down to the thermal range where they are dangerous. Before they reach this range they must pass successfully through that other where they are liable to be disarmed—or put away in prison, rather. This second brake does not diminish in its strength as the volume of uranium is raised.

Perhaps the second by itself is powerful enough to avert the explosion. In this case there is no danger of incurring the cataclysm by piling up uranium, however pure. There remains however the chance of separating the two isotopes 235 and 238, verifying that the fission by thermal neutrons occurs only in the one and the reaction of pure neutron-capture only in the other, and then accumulating the dangerous one by itself. Enough has just now been separated, as I said in an earlier footnote, for the verification to begin. To separate enough for the dangerous trial will take a good deal longer in the doing. After it is done, there is yet another brake which may avert catastrophe. When the cumulative processes begin, the heating of the metal may and probably will so affect the energies of the neutrons, that their efficiency for fissuring the nuclei will be greatly abated and so the processes find a natural limit. Otherwise it is to be hoped that those who build up great masses of sensitive uranium will recognize preliminary signs that the danger-point is close, before they actually attain it.

Literature

(Year always 1939 unless otherwise stated)

Early Chemical Identifications of Fission-Products: Hahn and Strassmann, *Naturwiss.* **27**, *passim;* also Curie and Savitch, *Jour. de Phys.* **9**, 355 (1938).

Isolation of Fission-Products by their Spontaneous Egress from Bombarded Uranium or Thorium: Joliot, *C.R.* **208**, 341; Meitner and Frisch, *Nature* **143**, 471; McMillan, *Phys. Rev.* **55**, 610; Segre, *Phys. Rev.* **55**, 1104.

Detection of Fission-Products in Ionization-Chamber: Meitner and Frisch, *Nature* **143**, 239, 276; Columbia school, *Phys. Rev.* **55**, 511; Roberts, Meyer and Hafstad, *Phys. Rev.* **55**, 416; Green and Alvarez, *Phys. Rev.* **55**, 417; Fowler and Dodson, *ibid.*

Fission-Products Detected in Expansion-Chamber: Joliot, *C.R.* **208**, 647; Corson and Thornton, *Phys. Rev.* **55**, 509.

Distribution-in-Energy of Fission-Products: Booth, Dunning and Slack, *Phys. Rev.* **55**, 981, 982, 1273; Kanner and Barschall, *Phys. Rev.* **57**, 372 (1940).

Further Chemical Studies of Fission-Products: Hahn and Strassmann, *Naturwiss.* **27**, 451; *Preuss. Akad.* 1939, no. 12; Savitch, *C.R.* **208**, 646; Thibaud and Moussa, *C.R.* **208**, 652; Heyn, Aten and Bakker, *Nature* **143**, 517; Bretscher and Cook, *Nature* **143**, 559; Dodson and Fowler, *Phys. Rev.* **55**, 880; Glasoe and Steigman, *Phys. Rev.* **55**, 982; Abelson, *Phys. Rev.* **56**, 1.

Decay-Curves for Totality of Fission-Products Unseparated: Bjerge, Brostrom and Koch, *Nature* **143**, 794.

Identification of Iodine among Fission-Products by X-rays: Abelson, *Phys. Rev.* **55**, 418; Feather and Bretscher, *Nature* **143**, 516.

Delayed Radiations: Roberts, Meyer and Wang, *Phys. Rev.* **55**, 510; Booth, Dunning and Slack, *Phys. Rev.* **55**, 876; Barschall *et al.*, *Phys. Rev.* **55**, 989; Gibbs and Thomson, *Nature* **144**, 202.

Yield of Fission-Process: Columbia School, *Phys. Rev.* **55**, 511; Ladenburg *et al.*, *Phys. Rev.* **56**, 168.

Energy and Yield of Neutrons Released at Fission: Anderson, Fermi and Hanstein, *Phys. Rev.* **55**, 797; Halban, Joliot and Kowarski, *Nature* **143**, 470, 680, 939; Haenny and Rosenberg, *C.R.* **208**, 898; Dode *et al.*, *C.R.* **208**, 995; Zinn and Szilard, *Phys. Rev.* **56**, 619; Halban *et al.*, *Jour. de Phys.* **10**, 428.

Scattering and Pure Neutron-Capture in Uranium: Whitaker *et al.*, *Phys. Rev.* **55**, 793; Anderson and Fermi, *Phys. Rev.* **55**, 1107; Halban, Kowarski and Savitch, *C.R.* **208**, 1396; Goldstein, Rogozinski and Walen, *Jour. de Phys.* **10**, 477; Anderson, *Bull. Amer. Phys. Soc.* Feb. 5, 1940.

Fission by Deuterons: Gant, *Nature* **144**, 707.

Fission of Protactinium: Grosse, Booth and Dunning, *Phys. Rev.* **56**, 382.

Fission of Separated Isotopes of Uranium: Nier, Booth, Dunning and Grosse, *Phys. Rev.* **57**, 546 (1940).

Theory of Fission: Bohr and Wheeler, *Phys. Rev.* **56**, 426, 1065.

General Review: Turner, *Rev. Mod. Phys.* **12**, 1 (1940).

A Solution for Faults at Two Locations in Three-Phase Power Systems

By E. F. VAAGE

This paper is an outgrowth of studies of double faults to ground in three-phase power systems made by the author in connection with work of the Joint Subcommittee on Development and Research, Edison Electric Institute and Bell Telephone System. The paper provides a systematic solution, based on the method of symmetrical components, by means of which currents and voltages can be determined at times of fault involving any combination of phases at one or two locations on three-phase power systems.

1. Introduction

A KNOWLEDGE of the magnitude and phase relation of power system voltages and currents for various types of faults in three-phase systems is of importance in the study of various problems, among which are relaying studies, the efficacy of current limiting devices and their reaction on the power network, and estimates of induction in paralleling communication circuits.

The method of symmetrical components as developed by Fortescue [1] and others is now extensively used in the solution for currents and voltages in three-phase power systems under fault (short-circuit) conditions. Formulas for special cases of faults, such as single and double line-to-ground faults, can be found in various text books on this subject. The solution for simultaneous faults at two locations has been treated by Miss Clark,[2] in a form particularly adaptable to the use of a calculating board.

The present development provides a complete and systematic solution for currents and voltages at times of fault on any number of phases at one or two locations in a three-phase system, in which generators may be assumed in phase and of the same internal voltage, and where load currents can be neglected. These are the usual assumptions made in computing fault currents, except for certain special problems, such as that of power system stability. The methods employed herein could be extended to cases where generators of different phase angles and voltages of more than two points of fault are involved. Formally such cases can be treated in a manner similar to that given in the paper. The number of impedances to which an n-terminal network

[1] Reference numbers refer to references appearing at the end of the article.

can be reduced is given by the expression $\frac{1}{2}n(n-1)$. For $n = 3$, the case treated in this paper, three impedances are required which necessitate six equations for the general solution of the six fault currents. For $n = 4$, which would be the case for three points of fault (or two points of fault and two generating voltages), six impedances would appear in the reduced network and this would necessitate twelve equations for the general solution of the fault currents. For larger values of n, the necessary number of equations increases rapidly, thus making the solution impractical. Such problems usually, as a practical matter, are more readily solved by the use of a-c. calculating boards.

While no departure from the general methods of symmetrical components is made in the present development, a systematic method of handling the equations is presented and means of determining the coefficients given so that numerical calculations can be directly carried out when the constants of the network are known.

2. General Solution

The equations developed in this paper are based on the sequence impedances looking into a three-phase network from two points of fault.

Consider the network shown in Fig. 1. This system can be reduced

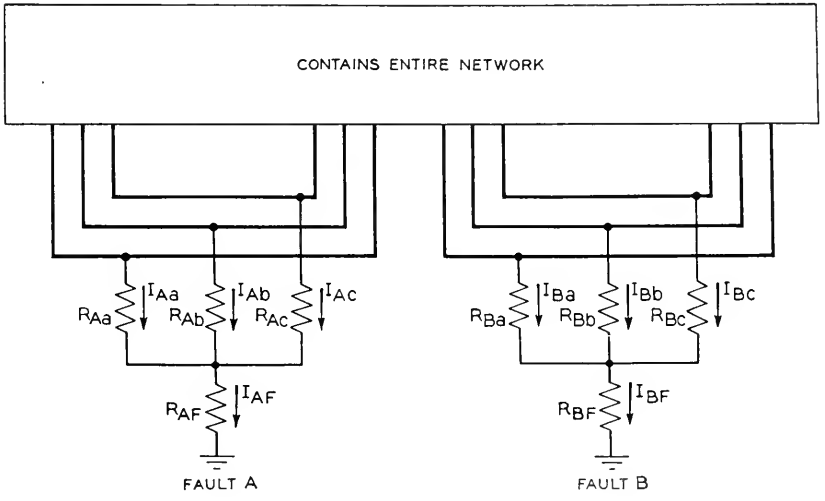


Fig. 1—General network diagram.

to an equivalent star for each of the positive, negative and zero sequence networks, with legs to the points of generation and to the faults at A and B. Figure 2 shows the reduced positive sequence network. Similar diagrams can be made for the negative and zero sequence systems except for the fact that in these cases there are no generated

voltages, and the impedances and currents are the negative and zero sequence quantities.

The reduction of a network to an equivalent star is usually a tedious and sometimes a difficult process especially in large interconnected systems. Methods of accomplishing the reduction, such as delta-star transformations, simultaneous equations or direct measurements on calculating boards can be found in the literature.[3, 5]

Having reduced the three sequence networks to equivalent stars, the equations are developed as shown in the Appendix.

The following set of equations (1) is the general solution for fault currents during simultaneous three-phase faults to ground at two different locations in a power network. The set applies directly to the calculation of ground fault currents on a system having finite

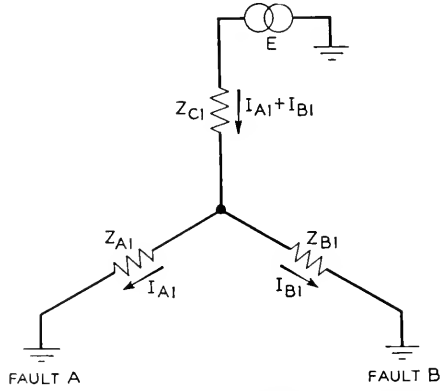


Fig. 2—Reduced positive sequence diagram.

neutral impedances or an isolated system in which the zero sequence capacitance is taken into account. For other types of faults, such as faults to ground in isolated systems in which zero sequence capacitance has been neglected, or for phase to phase faults, set (1) is not directly applicable since some of the constants become infinitely large. However, by certain transformations of set (1), more convenient sets (2) and (3) are obtained, directly applicable for solution of these latter cases.

Neutral Grounded System

I_{Aa}	I_{Ab}	I_{Ac}	I_{Ba}	I_{Bb}	I_{Bc}		
A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	$3E$	(Aa)
A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	$3a^2E$	(Ab)
A_{31}	A_{32}	A_{33}	A_{34}	A_{35}	A_{36}	$3aE$	(Ac)
A_{41}	A_{42}	A_{43}	A_{44}	A_{45}	A_{46}	$3E$	(Ba)
A_{51}	A_{52}	A_{53}	A_{64}	A_{55}	A_{56}	$3a^2E$	(Bb)
A_{61}	A_{62}	A_{63}	A_{64}	A_{65}	A_{66}	$3aE$	(Bc)

(1)

The six equations are written in matrix form with the currents and voltages outside the system matrix. For example the first row in (1) is interpreted as:

$$A_{11}I_{Aa} + A_{12}I_{Ab} + A_{13}I_{Ac} + A_{14}I_{Ba} + A_{15}I_{Bb} + A_{16}I_{Bc} = 3E$$

The values of the A's in (1) are given in Table I. It should be noted that of the 36 constants only 13 are distinct. Six of these are in the nature of self-impedances, two are transfer impedances between phases at A and two between phases at B. The remaining three are transfer impedances between the two faults at A and at B.

Considerable reductions in the constants are obtained when the positive and negative sequence impedances are assumed equal. These values are given in Table II.

Faults to ground on less than three phases at one or both locations are accounted for by assuming the corresponding fault resistances infinitely large. The currents to ground in the sound phases are zero. Striking out the columns containing these currents and the corresponding rows, indicated by the index at right in equation (1), a reduced set of equations is obtained from which the desired currents can be found. A few examples are given in subsequent sections.

In power networks with isolated neutral the zero sequence impedance Z_{C0} reduces essentially to the capacitance of the system. In this case equations (1) are still appropriate and will give a rigorous solution for the six currents. However, in many cases it is sufficiently accurate to neglect the capacitance of the system. This results in infinitely large values of all of the A's in Table I (Table II), since each depends on Z_{C0} which is infinitely large. For this condition it is desirable to transform the set of equations in (1) to a more convenient set with finite constants.

The transformation required is obtained by observing that, with Z_{C0} infinitely large, the sum of the zero sequence currents $I_{A0} + I_{B0}$ must be equal to zero. Making use of this relation the difference of the zero sequence voltages at A and B (equation (50) of appendix) reduces to:

$$V_{A0} - V_{B0} = (Z_{A0} + Z_{B0})I_{B0}$$

The last equation shows that subtraction of equations associated with phases at A from those at B removes the infinitely large element Z_{C0}. This can be done in nine ways (ignoring reversals of sign), but three of these result in the single equation:

$$I_{Aa} + I_{Ab} + I_{Ac} + I_{Ba} + I_{Bb} + I_{Bc} = 0$$

This equation with any five of the remaining six constitutes an independent set; for convenience in dealing with special cases the redundant set of seven equations is shown in the following array:

Isolated System—Capacitance Neglected

I_{Aa}	I_{Ab}	I_{Ac}	I_{Ba}	I_{Bb}	I_{Bc}		
B_{11}	B_{12}	B_{13}	B_{14}	B_{15}	B_{16}	$3(1-a^2)E$	$(Aa-Bb)$
B_{21}	B_{22}	B_{23}	B_{24}	B_{25}	B_{26}	$3(1-a)E$	$(Aa-Bc)$
B_{31}	B_{32}	B_{33}	B_{34}	B_{35}	B_{36}	$3(a^2-1)E$	$(Ab-Ba)$
B_{41}	B_{42}	B_{43}	B_{44}	B_{45}	B_{46}	$3(a^2-a)E$	$(Ab-Bc)$
B_{51}	B_{52}	B_{53}	B_{54}	B_{55}	B_{56}	$3(a-1)E$	$(Ac-Ba)$
B_{61}	B_{62}	B_{63}	B_{64}	B_{65}	B_{66}	$3(a-a^2)E$	$(Ac-Bb)$

(2)

$$I_{Aa} + I_{Ab} + I_{Ac} + I_{Ba} + I_{Bb} + I_{Bc} = 0 \qquad (2a)$$

The index to the right indicates which of the equations in (1) have been used. The values of the B's are given in Table I and Table II.

In case of faults to ground on less than three phases, as in equations (1), columns and rows associated with sound phase currents are to be deleted; with respect to the rows, however, the index is double and all rows having the index of the sound phase or phases are deleted. If, for example, the sound phase is Aa, rows 1 and 2, each of which contains Aa in its index, as well as column 1, are deleted. This leaves only four equations, which together with (2a) give the necessary five equations for the five currents. For this reason all six equations are given in (2), since any phase might be involved in special cases.

Phase-to-phase faults are obtained from the general case (1) by allowing the resistances R_{AF} and R_{BF} to become infinite. In this case phase-to-phase quantities at the same location remain finite and the appropriate set of equations is obtained by subtracting equations having the corresponding phase indexes; thus $Aa - Ab$, $Aa - Ac$ and $Ab - Ac$ indicate subtractions at A. There are six possible ways of doing this, ignoring reversals of sign. The resulting set is given in (3). The four equations obtained by taking any two of the first three and any two of the last three equations in this set together with the two equations (3a) relating to the sum of the currents at each fault location, which from physical considerations equal zero, constitute an independent set. For convenience in dealing with special cases all eight equations are given below:

Phase-to-Phase Faults

I_{Aa}	I_{Ab}	I_{Ac}	I_{Ba}	I_{Bb}	I_{Bc}		
C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	$3(1-a^2)E$	$(Aa-Ab)$
C_{21}	C_{22}	C_{23}	C_{24}	C_{25}	C_{26}	$3(1-a)E$	$(Aa-Ac)$
C_{31}	C_{32}	C_{33}	C_{34}	C_{35}	C_{36}	$3(a^2-a)E$	$(Ab-Ac)$
C_{41}	C_{42}	C_{43}	C_{44}	C_{45}	C_{46}	$3(1-a^2)E$	$(Ba-Bb)$
C_{51}	C_{52}	C_{53}	C_{54}	C_{55}	C_{56}	$3(1-a)E$	$(Ba-Bc)$
C_{61}	C_{62}	C_{63}	C_{64}	C_{65}	C_{66}	$3(a^2-a)E$	$(Bb-Bc)$

(3)

$$I_{Aa} + I_{Ab} + I_{Ac} = 0 \qquad (3a)$$
$$I_{Ba} + I_{Bb} + I_{Bc} = 0$$

The index to the right indicates which equations of (1) have been used. The values of the C's are given in Table I and Table II.

The total ground fault currents at the two fault locations are:

$$I_{AF} = I_{Aa} + I_{Ab} + I_{Ac} \tag{4}$$

$$I_{BF} = I_{Ba} + I_{Bb} + I_{Bc} \tag{5}$$

and the total residual current in the two faults is:

$$I_R = I_{AF} + I_{BF} \tag{6}$$

In an isolated system in which capacitance has been neglected this current is zero and equation (6) will be identical with (2a).

The distribution of these currents in the network can be found as follows. From equations (51) and (52) of the appendix the calculated fault currents are transformed into sequence currents. By working back into the original sequence networks the sequence currents in each branch of the system can be found and later combined from similar expressions as shown in (43) and (44) to obtain the actual branch currents.

A combination of the equations in (1) and (3) can be used for cases involving faults to ground at one location and faults between phases (not involving ground) at the other location.

For faults to ground at A and between phases at B, the three first equations in (1) and the three last in (3) together with the last in (3a) constitute the most convenient set of equations for this type of fault.

It should be noted that if all three phases are involved at B any two of the three last equations in (3) together with the last in (3a) can be used, while for less than three phases involved, the rules for striking out rows and columns automatically will result in the proper equations to be used.

Vice versa the three first equations in (3) together with the first in (3a) and the three last equations in (1) will give the solutions for phase-to-phase faults at A and ground faults at B.

This will be illustrated with an example in a later section.

The voltages to ground at the two locations of faults can be obtained directly from equations (53) and (54) of the Appendix, after the currents have been evaluated. At any other point in the system the voltages to ground are found by adding the voltage drops of the lines in question to these voltages, treating each sequence network separately, then adding the sequence voltages together according to equations (46) or (47).

TABLE I

CONSTANTS FOR EQUATIONS (1), (2) AND (3)

Equation (1)	Equation (2)	Equation (3)
$A_{11} = S_a + 3R_{Aa} + 3R_{AF}$	$B_{11} = S_a - T_c + 3R_{Aa} + 3R_{AF}$	$C_{11} = S_a - S_c + 3R_{Aa}$
$A_{12} = S_b + 3R_{AF}$	$B_{12} = S_b - T_a + 3R_{AF}$	$C_{12} = -S_a + S_b - 3R_{Ab}$
$A_{13} = S_c + 3R_{AF}$	$B_{13} = S_c - T_b + 3R_{AF}$	$C_{13} = S_c - S_b$
$A_{14} = T_a$	$B_{14} = -U_c + T_a - 3R_{BF}$	$C_{14} = T_a - T_c$
$A_{15} = T_b$	$B_{15} = -U_a + T_b - 3R_{Bb} - 3R_{BF}$	$C_{15} = T_b - T_a$
$A_{16} = T_c$	$B_{16} = -U_b + T_c - 3R_{BF}$	$C_{16} = T_c - T_b$
$A_{22} = S_a + 3R_{Ab} + 3R_{AF}$	$B_{21} = S_a - T_b + 3R_{Aa} + 3R_{AF}$	$C_{21} = S_a - S_b + 3R_{Aa}$
$A_{33} = S_a + 3R_{Ac} + 3R_{AF}$	$B_{22} = S_b - T_c + 3R_{AF}$	$C_{23} = -S_a + S_c - 3R_{Ac}$
$A_{44} = U_a + 3R_{Ba} + 3R_{BF}$	$B_{23} = S_c - T_a + 3R_{AF}$	$C_{32} = S_a - S_c + 3R_{Ab}$
$A_{45} = U_b + 3R_{BF}$	$B_{24} = -U_b + T_a - 3R_{BF}$	$C_{33} = -S_a + S_b - 3R_{Ac}$
$A_{46} = U_c + 3R_{BF}$	$B_{25} = -U_c + T_b - 3R_{BF}$	$C_{44} = U_a - U_c + 3R_{Ba}$
$A_{55} = U_a + 3R_{Bb} + 3R_{BF}$	$B_{26} = -U_a + T_c - 3R_{Bc} - 3R_{BF}$	$C_{45} = -U_a + U_b - 3R_{Bb}$
$A_{66} = U_a + 3R_{Bc} + 3R_{BF}$	$B_{32} = S_a - T_b + 3R_{Ab} + 3R_{AF}$	$C_{46} = U_c - U_b$
	$B_{34} = -U_a + T_c - 3R_{Ba} - 3R_{BF}$	$C_{54} = U_a - U_b + 3R_{Ba}$
	$B_{42} = S_a - T_c + 3R_{Ab} + 3R_{AF}$	$C_{56} = -U_a + U_c - 3R_{Bc}$
	$B_{46} = -U_a + T_b - 3R_{Bc} - 3R_{BF}$	$C_{65} = U_a - U_c + 3R_{Bb}$
	$B_{53} = S_a - T_c + 3R_{Ac} + 3R_{AF}$	$C_{66} = -U_a + U_b - 3R_{Bc}$
	$B_{54} = -U_a + T_b - 3R_{Ba} - 3R_{BF}$	
	$B_{63} = S_a - T_b + 3R_{Ac} + 3R_{AF}$	
	$B_{65} = -U_a + T_c - 3R_{Bb} - 3R_{BF}$	

Equalities

Equation (1)

$A_{12} = A_{23} = A_{31} \quad A_{45} = A_{56} = A_{64}$
$A_{13} = A_{21} = A_{32} \quad A_{46} = A_{54} = A_{65}$
$A_{14} = A_{25} = A_{36} = A_{41} = A_{52} = A_{63}$
$A_{15} = A_{26} = A_{34} = A_{42} = A_{53} = A_{61}$
$A_{16} = A_{24} = A_{35} = A_{43} = A_{51} = A_{62}$

Equation (2)

$B_{12} = B_{43} = B_{51} \quad B_{22} = B_{33} = B_{61}$
$B_{13} = B_{41} = B_{52} \quad B_{23} = B_{31} = B_{62}$
$B_{14} = B_{45} = B_{66} \quad B_{24} = B_{35} = B_{66}$
$B_{16} = B_{44} = B_{55} \quad B_{25} = B_{36} = B_{64}$

Equation (3)

$C_{13} = -C_{22} = C_{31}$
$C_{14} = -C_{26} = C_{35} = C_{41} = -C_{53} = C_{62}$
$C_{15} = -C_{24} = C_{36} = C_{42} = -C_{51} = C_{63}$
$C_{15} = -C_{25} = C_{34} = C_{43} = -C_{52} = C_{61}$
$C_{46} = -C_{55} = C_{64}$

where:

$S_a = (Z_{A1} + Z_{C1}) + (Z_{A2} + Z_{C2}) + (Z_{A0} + Z_{C0}) \quad T_a = Z_{C1} + Z_{C2} + Z_{C0} \quad U_a = (Z_{B1} + Z_{C1}) + (Z_{B2} + Z_{C2}) + (Z_{B0} + Z_{C0})$

$S_b = a(Z_{A1} + Z_{C1}) + a^2(Z_{A2} + Z_{C2}) + (Z_{A0} + Z_{C0}) \quad T_b = aZ_{C1} + a^2Z_{C2} + Z_{C0} \quad U_b = a(Z_{B1} + Z_{C1}) + a^2(Z_{B2} + Z_{C2}) + (Z_{B0} + Z_{C0})$

$S_c = a^2(Z_{A1} + Z_{C1}) + a(Z_{A2} + Z_{C2}) + (Z_{A0} + Z_{C0}) \quad T_c = a^2Z_{C1} + aZ_{C2} + Z_{C0} \quad U_c = a^2(Z_{B1} + Z_{C1}) + a(Z_{B2} + Z_{C2}) + (Z_{B0} + Z_{C0})$

TABLE II

CONSTANTS FOR EQUATIONS (1), (2) AND (3) WHEN POSITIVE AND NEGATIVE SEQUENCE IMPEDANCES ARE EQUAL

Equation (1)	Equation (2)	Equation (3)
$A_{11} = S_a + 3R_{Aa} + 3R_{AF}$	$B_{11} = S_a - T_b + 3R_{Aa} + 3R_{AF}$	$C_{11} = S_a - S_b + 3R_{Aa}$
$A_{12} = S_b + 3R_{AF}$	$B_{12} = S_b - T_a + 3R_{AF}$	$C_{12} = -S_a + S_b - 3R_{Ab}$
$A_{14} = T_a$	$B_{13} = S_b - T_b + 3R_{AF}$	$C_{13} = 0$
$A_{15} = T_b$	$B_{14} = -U_b + T_a - 3R_{AF}$	$C_{14} = T_a - T_b$
$A_{22} = S_a + 3R_{Ab} + 3R_{AF}$	$B_{15} = -U_a + T_b - 3R_{Bb} - 3R_{BF}$	$C_{23} = -S_a + S_b - 3R_{Ac}$
$A_{33} = S_a + 3R_{Ac} + 3R_{AF}$	$B_{16} = -U_b + T_b - 3R_{AF}$	$C_{44} = U_a - U_b + 3R_{Ba}$
$A_{44} = U_a + 3R_{Ba} + 3R_{BF}$	$B_{26} = -U_a + T_b - 3R_{Bc} - 3R_{BF}$	$C_{45} = -U_a + U_b - 3R_{Bb}$
$A_{45} = U_b + 3R_{BF}$	$B_{32} = S_a - T_b + 3R_{Ab} + 3R_{AF}$	$C_{56} = -U_a + U_b - 3R_{Bc}$
$A_{56} = U_a + 3R_{Bb} + 3R_{BF}$	$B_{34} = -U_a + T_b - 3R_{Ba} - 3R_{BF}$	
$A_{66} = U_a + 3R_{Bc} + 3R_{BF}$	$B_{63} = S_a - T_b + 3R_{Ac} + 3R_{AF}$	

Equalities

Equation (1)

$A_{12} = A_{13} = A_{21} = A_{23} = A_{31} = A_{32}$
$A_{45} = A_{46} = A_{54} = A_{56} = A_{64} = A_{65}$
$A_{14} = A_{25} = A_{36} = A_{41} = A_{52} = A_{63}$
$A_{16} = A_{16} = A_{24} = A_{26} = A_{34} = A_{35}$
$= A_{42} = A_{43} = A_{51} = A_{53} = A_{61}$
$= A_{62}$

Equation (2)

$B_{11} = B_{21}$ $B_{15} = B_{65}$
$B_{12} = B_{23} = B_{31} = B_{43} = B_{51} = B_{62}$
$B_{13} = B_{22} = B_{33} = B_{41} = B_{52} = B_{61}$
$B_{14} = B_{24} = B_{35} = B_{45} = B_{56} = B_{66}$
$B_{16} = B_{25} = B_{36} = B_{44} = B_{55} = B_{64}$
$B_{32} = B_{42}$ $B_{34} = B_{54}$
$B_{26} = B_{46}$ $B_{53} = B_{63}$

Equation (3)

$C_{13} = C_{16} = C_{22} = C_{25} = C_{31} = C_{34}$
$= C_{43} = C_{46} = C_{52} = C_{55} = C_{61}$
$= C_{64} = 0$
$C_{14} = -C_{15} = C_{24} = -C_{26} = C_{35}$
$= -C_{36} = C_{41} = -C_{42} = C_{51}$
$= -C_{53} = C_{62} = -C_{63}$
$C_{11} = C_{21}$ $C_{12} = -C_{32}$ $C_{23} = C_{33}$
$C_{44} = C_{54}$ $C_{45} = -C_{65}$ $C_{56} = C_{66}$

where:

$$
\begin{aligned}
S_a &= 2(Z_{A1} + Z_{C1}) + (Z_{A0} + Z_{C0}) \\
S_b &= -(Z_{A1} + Z_{C1}) + (Z_{A0} + Z_{C0}) \\
T_a &= 2Z_{C1} + Z_{C0} \\
T_b &= -Z_{C1} + Z_{C0} \\
U_a &= 2(Z_{B1} + Z_{C1}) + (Z_{A0} + Z_{C0}) \\
U_b &= -(Z_{B1} + Z_{C1}) + (Z_{A0} + Z_{C0})
\end{aligned}
$$

3. Special Cases

The application of the three sets of equations (1), (2) and (3), will be illustrated with a few examples. For simple cases, such as a single or double line-to-ground fault at one location, the equations reduce to formulas frequently found in the literature on this subject.

From set (1) equations for faults to ground at one or two locations can be obtained directly when the zero sequence impedance is finite. Set (2), obtained from (1), is the most convenient set for solutions of faults to ground in isolated systems in which capacitance has been neglected. The phase-to-phase fault currents are best obtained from set (3).

3.1 Single Line-to-Ground Fault at A

Consider a fault to ground on phase "*b*" at *A*. The solution can be obtained from (1) by letting:

$$R_{Aa} = R_{Ac} = R_{Ba} = R_{Bb} = R_{Bc} = \infty \tag{7}$$

This results in:

$$I_{Aa} = I_{Ac} = I_{Ba} = I_{Bb} = I_{Bc} = 0 \tag{8}$$

Striking out all columns in (1) containing the currents in (8) and the corresponding rows indexed by *Aa*, *Ac*, *Ba*, *Bb* and *Bc* only one equation is left:

$$A_{22}I_{Ab} = 3a^2E \tag{9}$$

The numerical value of A_{22} can be calculated directly from Table I, or on substituting the symbolic value of A_{22} in equation (9) the result will be:

$$I_{Ab} = \frac{3a^2E}{Z_1 + Z_2 + Z_0 + 3R_F} \tag{10}$$

where

$$\begin{aligned} Z_1 &= Z_{A1} + Z_{C1} \\ Z_2 &= Z_{A2} + Z_{C2} \\ Z_0 &= Z_{A0} + Z_{C0} \\ R_F &= R_{Ab} + R_{AF} \end{aligned} \tag{11}$$

Equation (10) is the well-known formula for a single line-to-ground fault at one location in a three-phase system.

3.2 Double Line-to-Ground Fault at A

Consider a double line-to-ground fault on phases "*a*" and "*b*" at *A*. Then:

$$R_{Ac} = R_{Ba} = R_{Bb} = R_{Bc} = \infty \tag{12}$$

and

$$I_{Ac} = I_{Ba} = I_{Bb} = I_{Bc} = 0 \tag{13}$$

Striking out the columns of (1) containing the currents in (13) and the corresponding rows (Ac, Ba, Bb and Bc) the following two equations remain:

$$\begin{aligned} A_{11}I_{Aa} + A_{12}I_{Ab} &= 3E \\ A_{21}I_{Aa} + A_{22}I_{Ab} &= 3a^2E \end{aligned} \tag{14}$$

from which on substituting the numerical values for the A's from Table I the two currents I_{Aa} and I_{Ab} can be found. The total fault current to ground at A is:

$$I_{AF} = I_{Aa} + I_{Ab} \tag{15}$$

In the special case where R_{Aa}, R_{Ab} and R_{AF} are zero the expression for I_{AF} can be reduced to the following expression after a direct substitution for the A's in (14) is made:

$$I_{AF} = \frac{-3aZ_2E}{Z_0Z_1 + Z_0Z_2 + Z_1Z_2} \tag{16}$$

where:

$$\begin{aligned} Z_1 &= Z_{A1} + Z_{C1} \\ Z_2 &= Z_{A2} + Z_{C2} \\ Z_0 &= Z_{A0} + Z_{C0} \end{aligned} \tag{17}$$

3.3 Simultaneous Double Line-to-Ground Fault at A and Double Line-to-Ground Fault at B

Consider a fault-to-ground on phases "a" and "b" at A and phases "a" and "c" at B. Then:

$$R_{Ac} = R_{Bb} = \infty \tag{18}$$

Hence:

$$I_{Ac} = I_{Bb} = 0 \tag{19}$$

Striking out the two columns containing I_{Ac} and I_{Bb} and the two corresponding rows (Ac and Bb), the four following equations remain:

$$\begin{aligned} A_{11}I_{Aa} + A_{12}I_{Ab} + A_{14}I_{Ba} + A_{16}I_{Bc} &= 3E \\ A_{21}I_{Aa} + A_{22}I_{Ab} + A_{24}I_{Ba} + A_{26}I_{Bc} &= 3a^2E \\ A_{41}I_{Aa} + A_{42}I_{Ab} + A_{44}I_{Ba} + A_{46}I_{Bc} &= 3E \\ A_{61}I_{Aa} + A_{62}I_{Ab} + A_{64}I_{Ba} + A_{66}I_{Bc} &= 3aE \end{aligned} \tag{20}$$

A symbolic solution in terms of the sequence impedances for these

currents becomes quite involved and it is advisable to substitute numerical values of the constants before solving for the four currents. The total fault currents at A and B, respectively, are (from (4) and (5) in connection with (19)):

$$I_{AF} = I_{Aa} + I_{Ab} \tag{21}$$

$$I_{BF} = I_{Ba} + I_{Bc} \tag{22}$$

In a similar manner faults to ground for any other combination of faulted phases can be found.

3.4 Single Line-to-Ground Faults at A and B in an Isolated System

Consider a fault-to-ground on phases "a" at A and "b" at B in an isolated system in which capacity can be neglected. Then:

$$R_{Ab} = R_{Ac} = R_{Ba} = R_{Bc} = \infty \tag{23}$$

and

$$I_{Ab} = I_{Ac} = I_{Ba} = I_{Bc} = 0 \tag{24}$$

Striking out the columns of (2) containing the currents in (24) and the corresponding rows $Aa - Bc$, $Ab - Ba$, $Ab - Bc$, $Ac - Ba$ and $Ac - Bb$ (all rows containing Ab, Ac, Ba and Bc), leaves only one equation in (2), which together with (2a) gives:

$$\begin{aligned} B_{11}I_{Aa} + B_{15}I_{Bb} &= 3(1 - a^2)E \\ I_{Aa} + I_{Bb} &= 0 \end{aligned} \tag{25}$$

Solving for these currents the result is:

$$I_{Aa} = - I_{Bb} = \frac{3(1 - a^2)E}{B_{11} - B_{15}} \tag{26}$$

Inserting the values of the B's from Table I this reduces to:

$$I_{Aa} = - I_{Bb} = \frac{3(1 - a^2)E}{Z_{1i} + Z_{2i} + Z_{0i} + 3(R_A + R_B)} \tag{27}$$

$$\begin{aligned} Z_{1i} &= Z_{A1} + Z_{B1} + 3Z_{C1} \\ Z_{2i} &= Z_{A2} + Z_{B2} + 3Z_{C2} \\ Z_{0i} &= Z_{A0} + Z_{B0} \\ R_A &= R_{Aa} + R_{AF} \\ R_B &= R_{Bb} + R_{BF} \end{aligned} \tag{28}$$

The subscript i (isolated) is used to distinguish these impedances for the isolated system from those used in (11), (17) and (38).

3.5 Phase-to-Phase Fault at A

Consider a fault between phases "a" and "b" at A. Then let:

$$I_{Ac} = I_{Ba} = I_{Bb} = I_{Bc} = 0 \tag{29}$$

Striking out the columns in (3) containing the currents in (29) and the corresponding rows (all rows containing Ac, Ba, Bb and Bc) only one equation is left:

$$C_{11}I_{Aa} + C_{12}I_{Ab} = 3(1 - a^2)E \tag{30}$$

It is further known from (3a) that:

$$I_{Ab} = - I_{Aa} \tag{31}$$

Substituting (31) and the constants C_{11} and C_{12} from Table I in (30) the result is:

$$I_{Aa} = - I_{Ab} = \frac{(1 - a^2)E}{Z_1 + Z_2 + R_{Aa} + R_{Ab}} \tag{32}$$

$$\begin{aligned} Z_1 &= Z_{A1} + Z_{C1} \\ Z_2 &= Z_{A2} + Z_{C2} \end{aligned} \tag{33}$$

which is a well-known expression for a phase-to-phase fault.

3.6 Three-Phase Fault at A

For this case:

$$I_{Ba} = I_{Bb} = I_{Bc} = 0 \tag{34}$$

Striking out the columns in (3) containing the currents in (34) and the corresponding rows, three equations remain, any two of which together with the equation from (3a) relating to the currents at A give:

$$\begin{aligned} C_{11}I_{Aa} + C_{12}I_{Ab} + C_{13}I_{Ac} &= 3(1 - a^2)E \\ C_{21}I_{Aa} + C_{22}I_{Ab} + C_{23}I_{Ac} &= 3(1 - a)E \\ I_{Aa} + I_{Ab} + I_{Ac} &= 0 \end{aligned} \tag{35}$$

from which the currents can be found.

In the special case where the fault resistances are all zero, the three currents are equal in magnitude and related as follows:

$$I_{Aa} = aI_{Ab} = a^2I_{Ac} \tag{36}$$

The rank of the system determinant in (35) is therefore 1. Using any of the first two equations in (35) in connection with (36) and the constants in Table I, the result is:

$$I_{Aa} = aI_{Ab} = a^2I_{Ac} = \frac{E}{Z_1} \tag{37}$$

$$Z_1 = Z_{A1} + Z_{C1} \tag{38}$$

3.7 Phase-to-Phase Fault at A and Phase-to-Ground Fault at B

Consider a fault between phases "a" and "b" at A and a fault to ground on phase "c" at B. Then:

$$R_{Ac} = R_{Ba} = R_{Bb} = \infty \tag{39}$$

and

$$I_{Ac} = I_{Ba} = I_{Bb} = 0 \tag{40}$$

As explained in a preceding section the three first equations in (3) together with the first in (3a) and the three last equations in (1) may be used for this case.

Striking out the columns I_{Ac}, I_{Ba} and I_{Bb} and the corresponding rows $Aa - Ac$, and $Ab - Ac$ in the three first equations in (3) leaves only the first equation. Similarly by striking out the columns I_{Ba}, I_{Bb}, and the corresponding rows Ba and Bb in the three last equations in (1) leaves only the last equation. Hence:

$$\begin{aligned} C_{11}I_{Aa} + C_{12}I_{Ab} + C_{16}I_{Bc} &= 3(1 - a^2)E \\ A_{61}I_{Aa} + A_{62}I_{Ab} + A_{66}I_{Bc} &= 3aE \end{aligned} \tag{41}$$

and finally from (3a):

$$I_{Aa} + I_{Ab} = 0 \tag{42}$$

from which the three currents can be found. The A's and C's are given in Table I and Table II.

4. Conclusion

While the probability of all phases being faulted at both locations simultaneously is very remote, the three sets of equations (1), (2) and (3) have been given in such a form that they conveniently will provide a solution for any combination of phases faulted from a single line-to-ground fault at one location to the most involved fault condition.

In Section (3) of this paper, in which special cases have been treated, only simple types of fault conditions have been shown in order to illustrate the method to be used and to prove that the general equations reduce to well-known formulas.

The constants given in Table I consist of the nine quantities S_a, S_b, S_c, T_a, T_b, T_c, U_a, U_b and U_c arranged as shown for each set of equations. Table II gives somewhat simpler values for the constants in cases where the positive and negative sequence impedances are assumed equal.

The voltages to ground at the two fault locations are given by (46) and (47) in the Appendix.

It is hoped that this development will provide a more unified presentation of fault current calculations in power networks.

ACKNOWLEDGMENT

The author wishes to express his appreciation to P. A. Jeanne and J. Riordan for valuable suggestions and checking of the mathematical relations in preparation of the paper.

APPENDIX

The standard notation for phase and sequence quantities is usually indicated by a subscript. Thus I_a, I_b, etc., means the current at the point of fault of phase "a" and "b," respectively. I_1, a^2I_1 and aI_1 are the positive sequence currents in phase "a," "b" and "c," respectively. In this treatment, however, complication arises from the fact that two points of faults are involved and it will be necessary to distinguish between the quantities at these two locations. This is most conveniently done by a double subscript, the first referring to the point of fault and the second to the phase or sequence in question. Thus I_{Aa}, I_{Ba}, etc., are the currents in the fault at A and B of phase "a" and I_{A1}, I_{B1} the positive sequence current at the two points of fault, respectively. Making use of this notation the fundamental equations for the sequence currents at fault A are:

$$\begin{aligned} I_{A0} + I_{A1} + I_{A2} &= I_{Aa} \\ I_{A0} + a^2I_{A1} + aI_{A2} &= I_{Ab} \\ I_{A0} + aI_{A1} + a^2I_{A2} &= I_{Ac} \end{aligned} \tag{43}$$

And at fault B:

$$\begin{aligned} I_{B0} + I_{B1} + I_{B2} &= I_{Ba} \\ I_{B0} + a^2I_{B1} + aI_{B2} &= I_{Bb} \\ I_{B0} + aI_{B1} + a^2I_{B2} &= I_{Bc} \end{aligned} \tag{44}$$

where the coefficient "a" is the sequence operator, having the value:

$$\begin{aligned} a &= -\frac{1}{2} + j\frac{\sqrt{3}}{2} \\ a^2 &= -\frac{1}{2} - j\frac{\sqrt{3}}{2} \end{aligned} \tag{45}$$

The voltages to ground at the two fault locations are given by:

$$\begin{aligned} V_{Aa} &= V_{A0} + V_{A1} + V_{A2} \\ &= (R_{Aa} + R_{AF})I_{Aa} + R_{AF}I_{Ab} + R_{AF}I_{Ac} \\ V_{Ab} &= V_{A0} + a^2V_{A1} + aV_{A2} \\ &= R_{AF}I_{Aa} + (R_{Ab} + R_{AF})I_{Ab} + R_{AF}I_{Ac} \\ V_{Ac} &= V_{A0} + aV_{A1} + a^2V_{A2} \\ &= R_{AF}I_{Aa} + R_{AF}I_{Ab} + (R_{Ac} + R_{AF})I_{Ac} \end{aligned} \tag{46}$$

$$\begin{aligned} V_{Ba} &= V_{B0} + V_{B1} + V_{B2} \\ &= (R_{Ba} + R_{BF})I_{Ba} + R_{BF}I_{Bb} + R_{BF}I_{Bc} \\ V_{Bb} &= V_{B0} + a^2V_{B1} + aV_{Ba} \\ &= R_{BF}I_{Ba} + (R_{Bb} + R_{BF})I_{Bb} + R_{BF}I_{Bc} \\ V_{Bc} &= V_{B0} + aV_{B1} + a^2V_{B2} \\ &= R_{BF}I_{Ba} + R_{BF}I_{Bb} + (R_{Bc} + R_{BF})I_{Bc} \end{aligned} \tag{47}$$

Consider the positive sequence diagram in Fig. 2. Evidently:

$$\begin{aligned} V_{A1} &= E - (Z_{A1} + Z_{C1})I_{A1} - Z_{C1}I_{B1} \\ V_{B1} &= E - Z_{C1}I_{A1} - (Z_{B1} + Z_{C1})I_{B1} \end{aligned} \tag{48}$$

where V_{A1} and V_{B1} are the positive sequence voltages to ground at the two fault locations. Similar expressions can be obtained for V_{A2}, V_{B2}, V_{A0} and V_{B0}, except for the fact the E is zero in these cases and the impedances and currents are the negative and zero sequence quantities. They are:

$$\begin{aligned} V_{A2} &= -(Z_{A2} + Z_{C2})I_{A2} - Z_{C2}I_{B2} \\ V_{B2} &= -Z_{C2}I_{A2} - (Z_{B2} + Z_{C2})I_{B2} \end{aligned} \tag{49}$$

$$\begin{aligned} V_{A0} &= -(Z_{A0} + Z_{C0})I_{A0} - Z_{C0}I_{B0} \\ V_{B0} &= -Z_{C0}I_{A0} - (Z_{B0} + Z_{C0})I_{B0} \end{aligned} \tag{50}$$

Solving (43) and (44) for the sequence currents the result is:

$$\begin{aligned} I_{A0} &= \tfrac{1}{3}(I_{Aa} + I_{Ab} + I_{Ac}) \\ I_{A1} &= \tfrac{1}{3}(I_{Aa} + aI_{Ab} + a^2I_{Ac}) \\ I_{A2} &= \tfrac{1}{3}(I_{Aa} + a^2I_{Ab} + aI_{Ac}) \end{aligned} \tag{51}$$

$$\begin{aligned} I_{B0} &= \tfrac{1}{3}(I_{Ba} + I_{Bb} + I_{Bc}) \\ I_{B1} &= \tfrac{1}{3}(I_{Ba} + aI_{Bb} + a^2I_{Bc}) \\ I_{B2} &= \tfrac{1}{3}(I_{Ba} + a^2I_{Bb} + aI_{Bc}) \end{aligned} \tag{52}$$

Substituting the expressions for the sequence currents in (48), (49) and (50) the result is:

$$\begin{aligned} V_{A1} &= E - \tfrac{1}{3}(Z_{A1} + Z_{C1})(I_{Aa} + aI_{Ab} + a^2I_{Ac}) \\ &\qquad - \tfrac{1}{3}Z_{C1}(I_{Ba} + aI_{Bb} + a^2I_{Bc}) \\ V_{A2} &= -\tfrac{1}{3}(Z_{A2} + Z_{C2})(I_{Aa} + a^2I_{Ab} + aI_{Ac}) \\ &\qquad - \tfrac{1}{3}Z_{C2}(I_{Ba} + a^2I_{Bb} + aI_{Bc}) \\ V_{A0} &= -\tfrac{1}{3}(Z_{A0} + Z_{C0})(I_{Aa} + I_{Ab} + I_{Ac}) \\ &\qquad - \tfrac{1}{3}Z_{C0}(I_{Ba} + I_{Bb} + I_{Bc}) \end{aligned} \tag{53}$$

$$
\begin{aligned}
V_{B1} &= E - \tfrac{1}{3}Z_{C1}(I_{Aa} + aI_{Ab} + a^2I_{Ac}) \\
&\quad - \tfrac{1}{3}(Z_{B1} + Z_{C1})(I_{Ba} + aI_{Bb} + a^2I_{Bc}) \\
V_{B2} &= -\tfrac{1}{3}Z_{C2}(I_{Aa} + a^2I_{Ab} + aI_{Ac}) \\
&\quad - \tfrac{1}{3}(Z_{B2} + Z_{C2})(I_{Ba} + a^2I_{Bb} + aI_{Bc}) \\
V_{B0} &= -\tfrac{1}{3}Z_{C0}(I_{Aa} + I_{Ab} + I_{Ac}) \\
&\quad - \tfrac{1}{3}(Z_{B0} + Z_{C0})(I_{Ba} + I_{Bb} + I_{Bc})
\end{aligned}
\tag{54}
$$

Substituting (53) and (54) in (46) and (47) the six original equations in (1) are obtained.

References

1. C. L. Fortescue: "Method of Symmetrical Coordinates Applied to the Solution of Polyphase Networks," *Trans. A. I. E. E.*, 1918, Vol. 37, pages 1027–1114.
2. E. Clarke: "Simultaneous Faults on Three-Phase Systems," *Trans. A. I. E. E.*, 1931, Vol. 50, pages 919–941.
3. Wagner and Evans: "Symmetrical Components" (McGraw-Hill, 1933).
4. Monseth and Robinson: "Relay Systems—Theory and Application" (McGraw-Hill, 1935).
5. Lyon: "Applications of the Method of Symmetrical Components" (McGraw-Hill, 1937).
6. G. Oberdorfer: "Der Erdschluss und seine Bekämpfung" (Springer, 1930).
7. M. J. Fallou: "Propagation des Courants de haute Frequence Polyphasés le long des Lignes Aériennes de Transport d'Energie Affectées de Courts—Circuits ou de Defauts d'Isolement," *Bulletin de la Société Française des Électriciens*, August, 1932, pages 787–864.

A Single Sideband Musa Receiving System for Commercial Operation on Transatlantic Radio Telephone Circuits*

By F. A. POLKINGHORN

In the operation of short-wave radio telephone circuits selective fading is observed which is a result of the combination at the receiving antenna of waves which have arrived from the transmitter over paths of different lengths. The poor quality resulting from this fading may be mitigated by increasing the directivity of the receiving antenna in the vertical plane so as to favor the waves arriving at one angle to the exclusion of others. Friis and Feldman have described an experimental system designed to accomplish this end which they call a "Musa" receiving system. This system was found under certain transmission conditions to give an improvement in the grade of circuit which could be obtained. A commercial installation of this type has now been constructed for use on the single sideband circuits of the American Telephone and Telegraph Company from England. Two receivers have been provided for the operation of four radio telephone circuits.

The antenna system consists of a row of sixteen rhombic antennas two miles long, each antenna connected by a separate transmission line to receivers located near the center of the row of antennas. In each receiver the signals from the antennas are combined in the proper phase to permit simultaneous reception from three adjustable vertical angles. The three signals are then added through delay equalizing circuits or discretely selected on the basis of amplitude to obtain diversity reception. A fourth branch of the receiver has its vertical angle of reception continuously varying and is used to set automatically the angles of reception of the three diversity branches. The delay equalization is also automatically adjusted. A recorder is provided which continuously registers the relative carrier field strength with variation of vertical angle of reception, and the amount of delay equalization.

Introduction

IN the operation of short-wave radio telephone circuits fading is observed which is caused by the combination at the receiving antenna of waves which arrive at different vertical angles and which have traveled from the transmitter over paths of different lengths. This fading may be mitigated by increasing the directivity of the receiving antenna in the vertical plane so as to favor the waves arriving over one path to the exclusion of the others.[1, 2] It is not

* *Proc. I.R.E.*, April 1940.

[1] E. Bruce, "Developments in Short Wave Directive Antennas," *Proc. I.R.E.*, Vol. 19, pp. 1406–1433, August 1931.

[2] E. Bruce and A. C. Beck, "Experiments with Directivity Steering for Fading Reduction," *Proc. I.R.E.*, Vol. 23, pp. 357–371, April 1935.

possible, however, to increase this directivity to any great extent with an ordinary antenna system before it is found that the signal arrives outside the angular range of the antenna an appreciable part of the time. To overcome this difficulty Friis and Feldman experimented with a receiving system consisting of a number of antennas, each having moderate directivity and each connected by a separate transmission line to a receiver where the outputs are phased by a variable phase shifting system in such a manner as to give a system of high, variable directivity. A system of this kind, which they called a "musa" system from the initial letters of "multiple unit steerable antenna," was built and found to give under most transmission conditions an improvement in the grade of circuit which could be obtained.[3] Accordingly it was decided that a commercial system should be built for use on the circuits of the American Telephone and Telegraph Company from England. A corresponding system of modified design has been built by the British General Post Office.[4] The purpose of this paper is to review a few of the principles upon which a musa receiver operates, to describe the equipment which has been built in this country, and to discuss some of its operating characteristics.

The transmissions which are to be received are of the so-called twin single-sideband reduced-carrier type described by Oswald [5] and consist of two sidebands, representing two distinct speech channels, on opposite sides of a carrier which is 16 to 26 db below the maximum sideband amplitude. Under normal conditions one of the sidebands is adjacent to the carrier while the second is spaced by the width of one sideband from the carrier. A single sideband receiver for this type of transmission has been described by Roetken [6] and many of the features discussed by him were developed for use also in the musa receivers. These features include highly stable oscillators, crystal filters and automatic tuning circuits.

Outline Description of Receivers

A block schematic of one channel of the commercial musa system is shown in Fig. 1. The sixteen rhombic antennas are placed in a line two miles long in the direction of the English transmitting station.

[3] H. T. Friis and C. B. Feldman, "A Multiple Unit Steerable Antenna for Short-Wave Reception," *Proc. I.R.E.*, Vol. 25, pp. 841–917, July 1937; *B.S.T.J.*, Vol. XVI, No. 3, pp. 337–419, July 1937.

[4] A. J. Gill, Wireless Section, Chairman's Address, *Jour. I.E.E.*, Vol. 84, No. 506, pp. 248–260, February 1939.

[5] A. A. Oswald, "A Short-Wave Single Sideband Radio Telephone System," *Proc I.R.E.*, Vol. 26, No. 12, pp. 1431–54, December 1938.

[6] A. A. Roetken, "A Single Sideband Receiver for Short-Wave Telephone Service," *Proc. I.R.E.*, Vol. 26, No. 12, pp. 1455–65, December 1938.

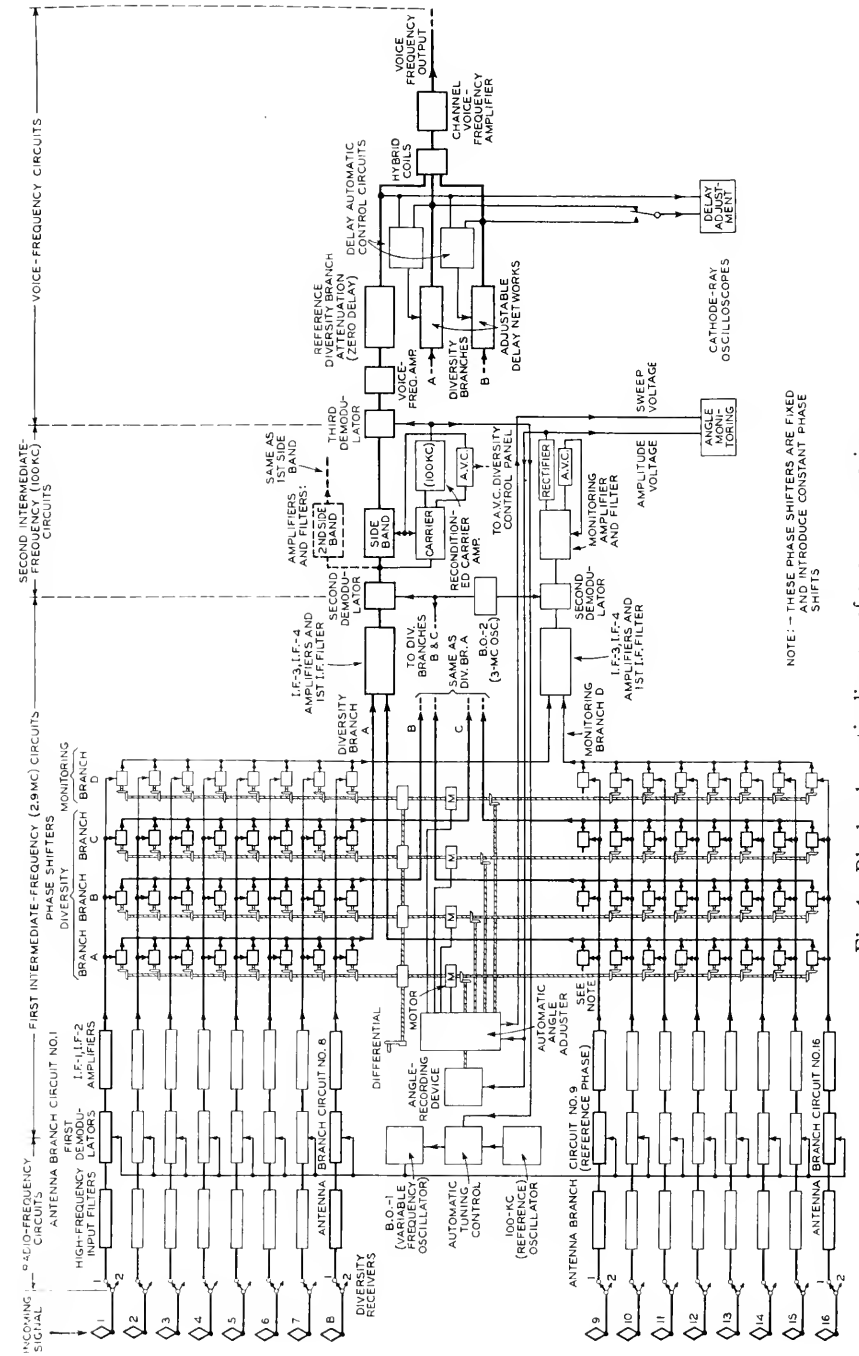


Fig. 1—Block schematic diagram of one musa receiver.

Separate transmission lines lead from each antenna to a building placed a little to one side of the rear of the ninth antenna. Two receivers, only one of which is shown in the figure, are connected in parallel to each transmission line. Each receiver is designed to receive five specific frequencies, ranging from 4,810 kc. to 18,620 kc., assigned to the corresponding transmitter in England.

After passing through selective input circuits the signals from each antenna are demodulated by a common oscillator to a band adjacent to a carrier frequency of 2,900 kc. The signals, after going through two stages of intermediate frequency amplification are then applied to the inputs of four phase shifter systems in parallel. In each of these phase shifter systems the signals from the sixteen antennas are combined so as to give reception from a particular vertical angle. This angle can be varied by a mechanical movement of a phase shifter drive shaft. Three of the phase shifter system outputs are used for a three-branch angular-diversity system in which the signals arriving over three separate paths are separately received and then either combined or individually selected for connection to the line, while the fourth branch is used for monitoring to determine where the phase shifters of the three diversity branches should be set in order to receive the best signals. From each phase shifter group the circuit continues through the first intermediate frequency filter and two further stages of amplification to the second demodulator where the 2,900 kc. carrier frequency is shifted to 100 kc. The carriers and sidebands of each diversity branch are then amplified separately and again combined in the final demodulators to give three distinct voice frequency outputs for each sideband. These three outputs are either combined after inserting variable delay in two of the branches or, optionally, the branch having the greatest signal at any instant is connected to the line. Both of these operations are performed automatically. The output of the monitoring phase shifter group is also heterodyned to 100 kc. and after amplifying the carrier only it is rectified and applied to an automatic system for adjusting the phase shifters of the three diversity branches.

A general view of the receivers is shown in Fig. 2. The principal parts of the two musa receivers occupy three rows of bays each about 25 feet long and $11\frac{1}{2}$ feet high. The row shown on the right contains the input circuits and first demodulators for both receivers. The middle row contains the remaining equipment for one receiver and the left row that for the second receiver. In addition there are five bays of rectifiers and power control equipment located in a fourth row which is not shown.

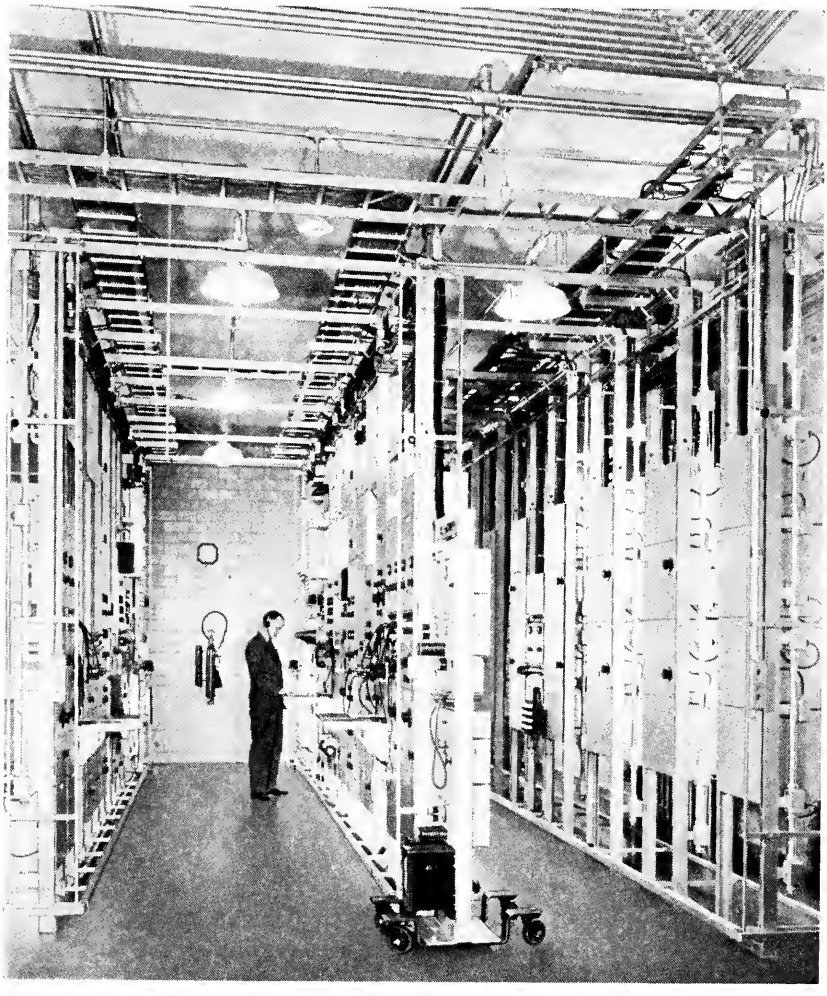


Fig. 2—View of musa receivers.

General

When waves arriving over several paths from the same transmitter are demodulated in a simple receiver the severity of the resultant selective fading is dependent upon the relative amplitudes at the demodulator of the several path contributions, the differences in the times of transmission over the several paths, and the rates at which the path lengths are varying.[7] When the difference in the time of

[7] R. K. Potter, "Transmission Characteristics of a Short-Wave Telephone Circuit," *Proc. I.R.E.*, Vol. 18, No. 4, pp. 581–648, April 1930.

transmission over two paths is t there are alternate maxima and minima in the frequency spectrum caused by these two components which are separated by $1/2t$. Continuous small changes in the lengths of the paths cause these maxima and minima to wander back and forth through the spectrum. By separating the waves arriving at distinctly different angles the musa receiver succeeds, for the most part, in separating those waves which have greatly different transmission times and thus widens the frequency interval between a maximum and an adjacent minimum. As the interval increases the fading appears less selective. The signal appearing to arrive at any one angle, however, is in reality composed of a bundle of waves, the components of which have traveled over slightly different paths and which might be expected to be nearly alike in amplitude and transmission time but not in phase. As a consequence it is to be expected that the general fading on a single-angle musa receiver will be greater than on an ordinary receiver and it is essential that some form of diversity be used to insure a satisfactory output amplitude at all times. Sudden shifts in the received angle of signals will also give general fading which will be greater the greater the angular discrimination of the musa system.

A musa receiver differs from an ordinary receiver in that there are a number of separate antenna branches, the outputs of which must be added in the proper phase over an appreciable band of frequencies.

When delay equalization is used between the various diversity branches, these branches must also have equal phase shifts if the audio-frequency outputs are to add properly. In both antenna and diversity branches the problem of keeping equal phase shifts is complicated by the action of the automatic volume control system which changes the operating condition of the vacuum tubes over a wide range. In designing these receivers a nominal overall value of non-uniformity of $\pm 10°$ was taken as acceptable and an effort made to keep the phase uniformity of individual elements to within one or two degrees wherever possible.

Within the receiving station all radio-frequency wiring is made with a flexible coaxial cable having rubber insulation. The various panels composing the receivers are placed on the racks with a view to operation and maintenance rather than ease of wiring and consequently long leads between panels are frequently necessary. For this purpose the circuit impedance is dropped to 70 ohms and at a number of points brought out to jack panels to facilitate testing.

Coaxial jacks are used which fit into the usual jack strips. Normalling jacks are not available and consequently it is necessary to

have plugs in jacks during operation. In order to avoid cords, which would be in the way, the jacks to be connected regularly are mounted adjacent to each other and connected together by two plugs mounted in a shell similar to that commonly used for terminating resistances. Alternating current for cathode heating is supplied to conduit outlets near each panel. Flexible cords with plugs complete the circuit to the panels. All audio frequency, bias, and signal wiring is made into cables in the usual telephone manner. Wires having a potential of over 150 volts to ground are placed in conduit and safety switches are provided to remove the voltage from a panel when the panel cover is removed.

Antenna System

The degree of vertical resolution and the signal-to-noise improvement of a musa antenna system are functions of the overall length and number of unit rhombic antennas used. The decision to build a sixteen-antenna system was based on experience with the six-antenna system and took into consideration the land necessary, the cost of antennas and transmission lines, and the complexity of the receiving equipment, as well as the resolution which it would be practical to use and still have it possible for the operator or automatic equipment to follow changes in the direction of signal arrival.

When the spacing between unit antennas of a musa system is several wave-lengths there will be more than one vertical angle at which the phase shifters will simultaneously phase the antenna outputs. The spacing between antennas is so chosen that the range traversed by the lowest of these angles will be the range covered by useful signals. Fortunately the angle of useful signals varies with frequency in such a manner as to permit a variation in frequency over the range desired with a fixed spacing. The unit rhombic antenna is designed to have a null at the position of the second phasing maximum and reception is thus confined essentially to the lowest phasing maximum.

Extensive study did not disclose a better unit rhombic antenna than the one used in the experimental system and consequently an antenna 590 feet (180 meters) long, 60 feet high, and having each side angle equal to 140 degrees was used. The spacing is 656 feet (200 meters) between corresponding parts of adjacent antennas.

A representative directive diagram of a unit rhombic and the 16-unit array in a vertical plane in the line of the antennas is shown plotted in rectangular coordinates in Fig. 3. A polar diagram of the major lobe in three different positions, corresponding to three possible angles of diversity reception, is shown on Fig. 4*a*. In this figure the

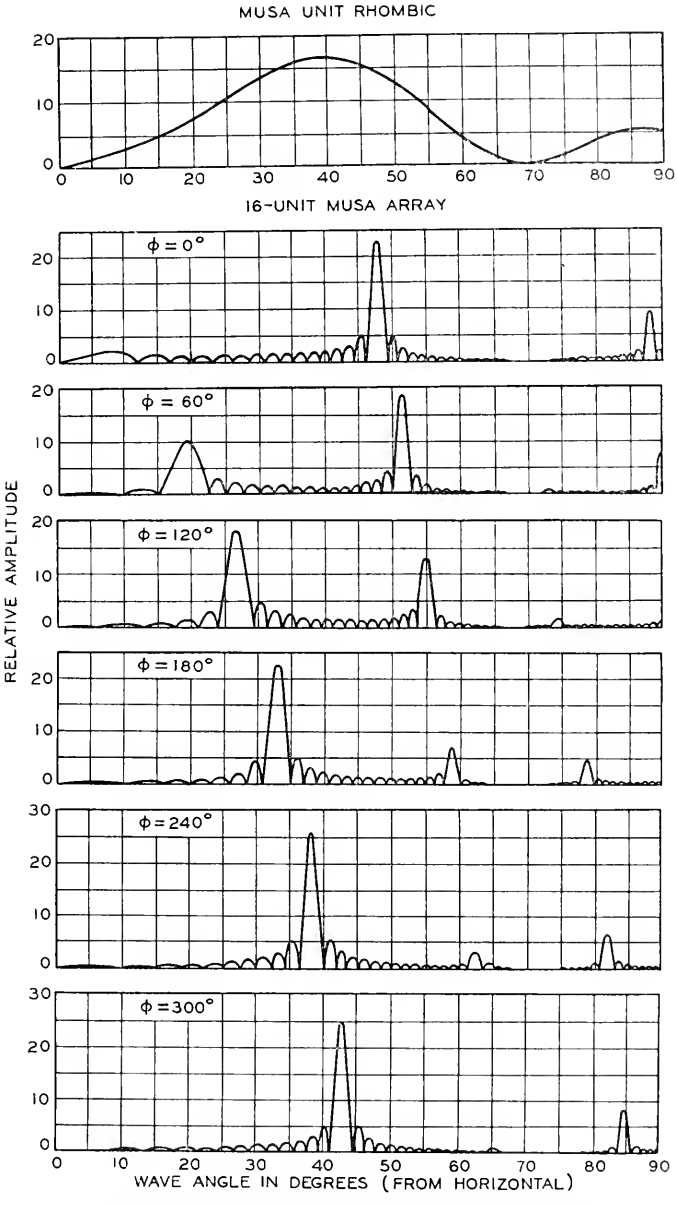


Fig. 3—Directive diagram in the vertical plane of a unit rhombic antenna and the 16 unit musa array. Frequency 4700 Kc. ϕ = phase shifter setting.

dotted outline is the diagram of the unit rhombic antenna characteristic enlarged 16 times. The complete diagram is of course a solid. Figure 4*b* is an attempt to show how the middle lobe of Fig. 4*a* looks when viewed from the ground plane at a horizontal angle of 45° from the line of the antennas. The contour lines on this leaf-shaped figure are lines of equal reception. All of these diagrams are for a frequency of 4,700 kc., near the low end of the range of received signals. At higher frequencies the lobes will be more slender and the angle of maximum reception will be lower. Fortunately the latter corresponds to the trend of the received signals.

When the outputs of several antennas are connected to a receiver and added in the proper phase the total signal can be made equal to

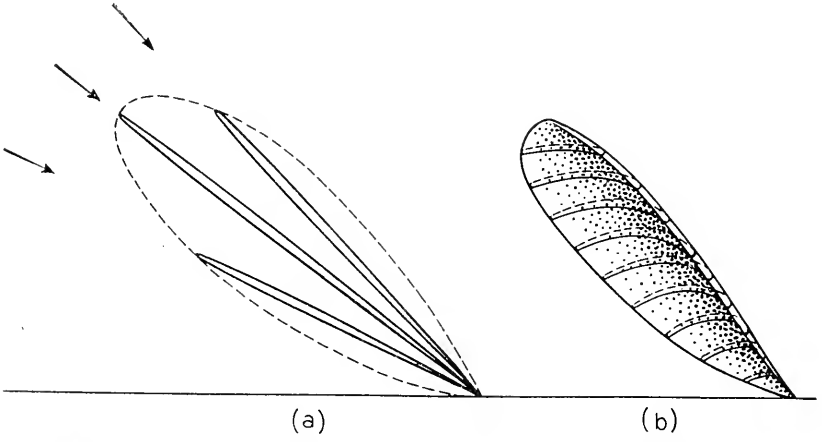


Fig. 4—Polar diagrams of musa antenna system. (*a*) Showing three possible locations of the major lobe. (*b*) Solid polar diagram of middle lobe shown in (*a*).

the sum of the signal voltages. The set noise, however, adds at random phase with the result that there is an improvement in signal-to-noise ratio over a single antenna equal to the square-root of the number of antennas, or $10 \log_{10} n$ in decibels. The theoretical improvement of a 16-antenna system is, therefore, 12 decibels. On the assumption that received noise comes from a random direction with respect to the signal a similar result is obtained. At a particular time, however, the principal noise may be arriving from such a direction as to allow of no discrimination by the antenna system, or at another time to allow much more than 12 db discrimination.

The received signal in an antenna is a result of both the direct wave, arriving from some angle above the horizon, and the same wave front

reflected from the ground at a point ahead of the antenna. If the phase of the induced voltage is to progress uniformly from antenna to antenna it is essential that the ground be homogeneous and flat. For this and other reasons a flat marsh near Manahawkin, N. J. was chosen for the station site. The ground is level to within less than one foot, except where there are inlets, for the entire length of the antenna and for a considerable distance ahead.

As shown in Fig. 3, in addition to the major lobe caused by the phasing factor there are fourteen minor lobes. The amplitude of the first three starting from a major lobe are approximately $2/3\pi$, $2/5\pi$, $2/7\pi$ of the major lobe. However, when the major lobe is at a very low or very high angle it is greatly reduced in amplitude by the unit antenna directional characteristic while the adjacent minor lobes on one side may not be reduced to any such extent. Consequently the ratio of the amplitude of the major to minor lobes may be much less than the values given and signals from two or more angles might be received simultaneously with comparable amplitude on the same diversity branch and so defeat the purpose of the system.

It has been shown by John Stone Stone and others that if the amplitudes of the currents contributed by the various units of the antenna system are tapered in such a manner that the central units contribute more than the end units a reduction in the amplitude of the minor lobes can be obtained. However, this is accompanied by a widening of the major lobe and a reduction in the signal-to-noise improvement obtained. For antenna systems having only a few units there appears to be a net advantage in tapering but for the sixteen-unit system under discussion a large amount of tapering broadens the major lobe so that it extends over the normal first and possibly second minor lobes. Since the remaining lobes are already of a low enough amplitude to be comparable with those which might be produced by inescapable errors in phase and amplitude of the various unit contributions, there appears to be no particular advantage in much tapering in this system. Provision has, however, been made to obtain tapering should it ever be found desirable. Under normal conditions all antenna branch amplifiers are operated at the same gain so as to use only the small tapering caused by the losses in transmission lines.

The antennas are coupled to the transmission lines through metallic core transformers which pass a band from 4,000 kc. to over 20,000 kc. with a loss of less than 1 db. The transformers are equipped with lightning arrestors and arranged so that the total d.c. loop-resistance of the transmission line, antenna, transformer, and antenna terminating resistance can be checked from the station.

The transmission lines from the antennas to the receiver are of the coaxial type made of copper tubing $1\frac{1}{4}''$ in inside diameter surrounding a central conductor of $\frac{3}{8}''$ outside diameter. Ceramic insulators $\frac{1}{8}''$ thick are spaced every 16'' and a locking insulator is placed every 250 feet to prevent creeping of the inside pipe. The velocity of propagation of the line is 0.980 of that in space. The attenuation amounts to about 1 db per 1,000 feet at 20 megacycles.

The lines are buried to protect them from mechanical injury and to prevent phase errors due to differences in expansion. Bacterial growth in the marsh makes the soil extremely corrosive and it was necessary to protect the lines by coatings of tar and asbestos tape. The lines are kept under gas pressure at all times.

In a musa receiving system a saving of nearly one half of the transmission line can be made if the receiving equipment is located at the center of the antenna system rather than at one end. Furthermore, since the average length of transmission line will be cut in half, the diameter of the coaxial transmission line also could be cut in half and still maintain the same signal loss. The economic, and to a lesser extent the locational, advantages of the center position were so great that the equipment was so placed in spite of certain technical disadvantages.

Theory of Phase Shifter System with Center Location of Receivers

At this point some of the theory of the musa phase shifting system will be reviewed and extension made to cover the situation of the receiver being located near the middle of the antenna system.

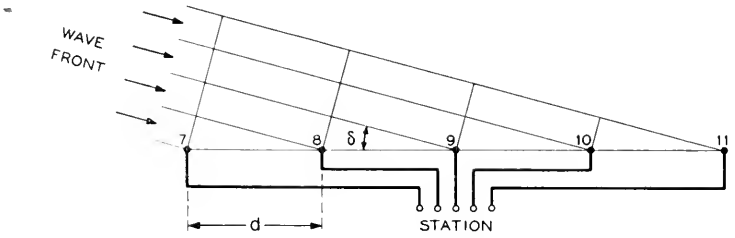


Fig. 5—Diagram of wave front approaching antenna.

Assume a plane wave front progressing toward the earth from the Kennelly-Heaviside layer at an angle δ with respect to the horizon and impinging upon the receiving antennas indicated as points 7, 8, 9, 10, 11, etc. (Fig. 5). Let the receiving station be located at antenna 9. Let the time of arrival at the receiving station of the voltages induced

in the various antennas by this wave front be computed with respect to that of antenna 9. The wave front will arrive at antenna 9 at a time $\frac{1}{c}(d \cos \delta)$ later than at antenna 8. The voltage from antenna 8 will require a time $\frac{d}{v}$ to reach the receiving station over the transmission line, making the net time delay between the two outputs equal to

$$t_{9-8} = d\left(\frac{1}{v} - \frac{1}{c}\cos \delta\right), \tag{1}$$

where d is the horizontal distance between antennas
c is the velocity of transmission in space
v is the velocity of transmission in the transmission line

Similarly the difference in time of arrival of the voltages from antennas 9 and 7 will be:

$$t_{9-7} = 2d\left(\frac{1}{v} - \frac{1}{c}\cos \delta\right). \tag{2}$$

The voltages from all three antennas could be made to add in phase if a delay of

$$2d\left(\frac{1}{v} - \frac{1}{c}\cos \delta\right) \tag{3}$$

were added to the output from antenna 9, and half as much to the output of antenna 8. For a greater number of antennas these delays would have to be correspondingly increased.

Now consider the outputs of antenna 9, and those antennas which lie behind it. It will immediately be seen that the time between the arrival of the voltage caused by the wave front at antenna 9 and antenna 10 will be the sum of the times of transmission of the wave front from 9 to 10 in space and on the transmission line back from 10 to 9, which is similar to equation (1) except that the sign between the space and transmission line components is reversed. Similarly the equation for the delay between the outputs of antennas 9 and 11 is similar to that between 7 and 9 with the same sign reversed, and likewise with the other antennas.

Considering either the group of antennas ahead of the station or the group of antennas behind the station, and considering antenna 9 as a part of whichever group is being considered, it will be seen that the delay compensation which it is necessary to insert in the various antenna outputs to make the signal voltages add up in phase at the re-

ceiving station is always an integral multiple of either:

$$d\left(\frac{1}{v}-\frac{1}{c}\cos\delta\right) \tag{4}$$

or

$$d\left(\frac{1}{v}+\frac{1}{c}\cos\delta\right), \tag{5}$$

so that for either the front or the rear group the delay compensation could be adjusted by means of a single shaft geared to various individual delay compensators through gears having integral ratios.

Practically, the difficulty in building continuously variable delay circuits resulted in the use of continuously variable phase shifters. For any particular delay t there is a corresponding phase shift $\frac{2\pi ct}{\lambda}$ radians. Since this phase shift is a function of frequency, for a given phase shifter setting the frequencies in a wide band transmission will not all be in phase. The substitution of phase shifters for delay compensating circuits therefore results in a restriction of the band of simultaneous reception from a given angle. For the group of antennas ahead of the station the band restriction is small, since the space and transmission line paths give a partial delay compensation. For the antennas behind the station the net delay difference between antennas is large and the band is restricted to a much greater extent.

When using phase shifters it is possible to get minus as well as plus values of phase shift. This makes it possible to reverse the order in which delay compensation was assumed to be added in the above discussion so that no phase compensation is added to antenna 9, one negative unit is added to antenna 8, two to antenna 7, etc. In the equipment herein described the phase shifters were so designed.

Assume that the phase shifters of the front group are all set alike and then changed so as to receive from an angle δ. There will have been introduced a change of phase of

$$\phi=\frac{2\pi cd}{\lambda}\left(\frac{1}{v}-\frac{1}{c}\cos\delta\right) \tag{6}$$

between each of the antenna outputs.

Similarly if the phase shifters of the rear group were initially set alike, and similar to the initial condition of the front group, and then changed to receive from an angle δ the change of phase which would be required would be

$$\phi=-\frac{2\pi cd}{\lambda}\left(\frac{1}{v}+\frac{1}{c}\cos\delta\right). \tag{7}$$

The negative sign ahead of this equation takes account of the fact that an increase in the angle δ requires that more phase will have to be taken from line 8 while less phase will have to be subtracted from 10.

It will be noted that for a given change in δ, however, the amount of change of phase is the same absolute amount for the two groups. This indicates that it should be possible to connect the front and rear groups of phase shifters together and drive them as a single unit. To do this it will be necessary to connect the shafts together after they have been moved from their initial position by the amounts given in the above equations. The difference in phase is

$$\phi_{8-10} = \frac{4\pi d}{\lambda} \frac{c}{v}. \tag{8}$$

It will be noted that for a given installation this value is dependent upon the received frequency and will have to be changed when the received frequency is changed.

In the receivers herein described the phase shifters connected to antenna 9 are not driven. The phase shifters connected to antennas 1 to 8 are driven at ratios of 8 : 1, 7 : 1, 6 : 1, 5 : 1, 4 : 1, 3 : 1, 2 : 1 and 1 : 1 respectively. The phase shifters of antennas 10 to 16 are driven in the opposite direction at ratios of 1 : 1, 2 : 1, 3 : 1, 4 : 1, 5 : 1, 6 : 1, and 7 : 1 respectively. A differential gear mechanism is inserted between the two groups. Changing the position of the ring gear of the differential permits a mechanical shift equivalent to the phase shift given in equation (8) to be inserted between the two groups. This change may be made while the drive shafts are in motion. The ring gear drives of the differentials of the four groups of phase shifters are all connected by a common adjustment shaft so as to insure that the monitoring branch does not give a false indication of receiver performance by being set differently than the diversity branches.

A front view of the phase shifting system is shown on the left of Fig. 6. The vertical shafts drive the individual phase shifting condensers through spiral gears, of which there are eight inside each of the lower cast boxes and seven inside each of the upper cast boxes. The horizontal shafts connect to the cam switches used with the automatic adjustment feature. A rear view showing a few of the individual phase shifting condensers is given in Fig. 7. Each phase shifting condenser consists of four quadrantal stators each consisting of two plates between which revolves an eccentric circular plate. The rotor plates of the condensers in a horizontal row (see Figs. 1 and 7) are

connected in parallel through the demountable brushes, and a wire which is behind the brush support, to the output of one of the antenna branch amplifiers. The corresponding stator plates of the condensers in a vertical row are connected together with coaxial wiring and to the four terminals of an artificial quarter-wave transmission line, which

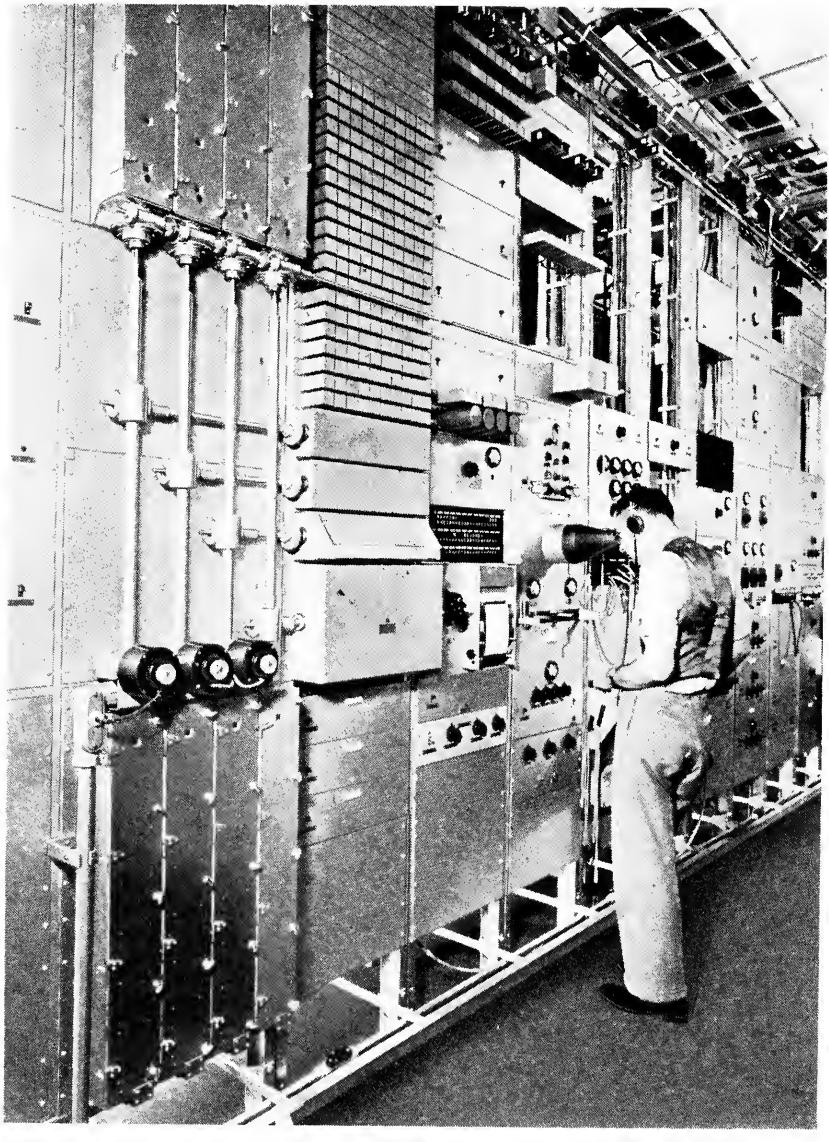


Fig. 6—View of second row of bays showing phase shifter drive system (left).

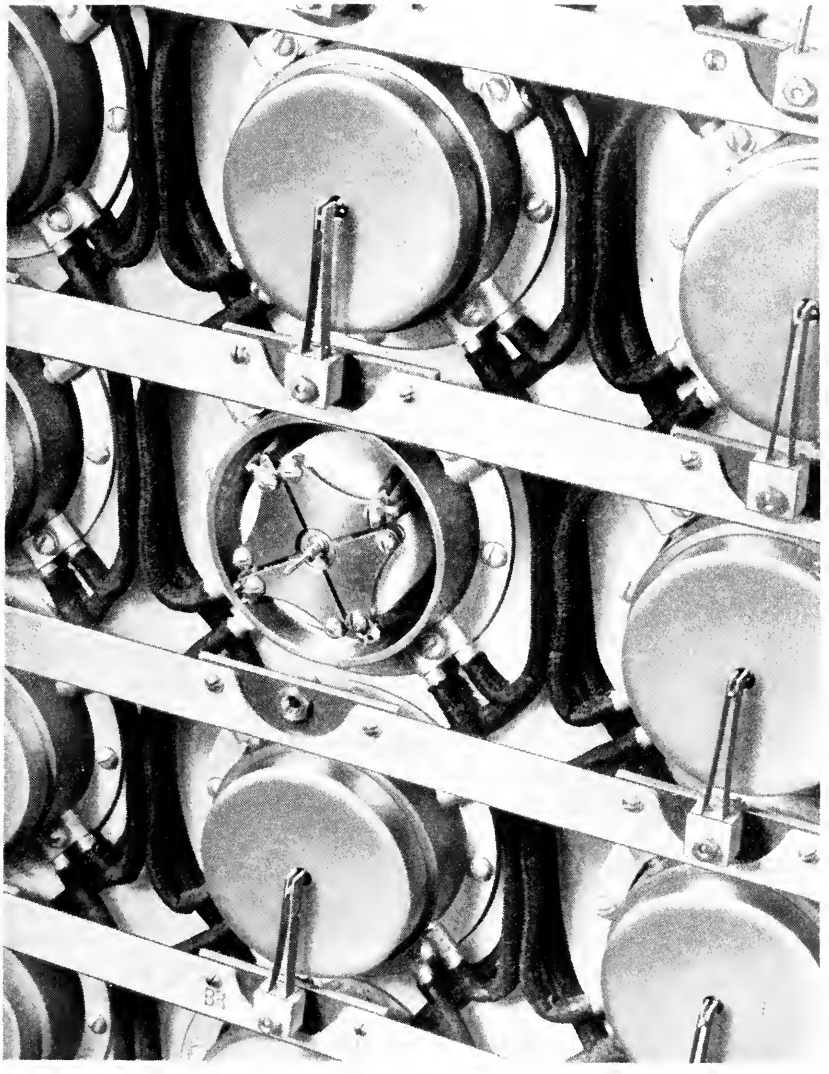


Fig. 7—Portion of a group of phase shifting condensers. Brush and cover removed from center condenser to show construction.

in turn is connected to the output amplifier. The quarter-wave line forms a combining network for the contributions of the various phase shifters. The two groups of eight phase shifters forming one diversity branch are each treated as units so as to facilitate isolating one group in case of trouble or the necessity of reducing the resolution of the antenna system.

There are two opportunities for undesirable interchanges of power in the parallel phase shifter system; there can be an interchange between antenna branches through condensers in the same diversity group, and an interchange between diversity branches through condensers in the same antenna branch. Both of these can be reduced to the required values by a proper proportioning of input and output impedances with respect to the condenser reactance. This results in a large loss, about 40 db, through the phase shifting system.

Input Circuits and First Demodulators

Obtaining uniformity of phase shift in the circuits preceding the first detectors is facilitated by reducing the number of circuit elements and selectivity to a minimum. No high-frequency amplification is used for this reason. The selectivity required to avoid image reception is materially lowered by the choice of a high operating frequency for the first intermediate frequency amplifiers. In addition to the usual requirement of high selectivity and high voltage transformation ratio, the input circuits must have uniformity in gain and phase, and must properly terminate the transmission lines so that multiple reflections will not shift the effective input amplitude or phase.

These factors seemed to preclude use of the usual variable tuned circuits on account of the time required to change from one frequency to another, and consequently fixed tuned circuits were used. Five input circuits are mounted on a panel and the switches for changing the input and output circuits of the two panels of each receiver which are mounted on a single bay are ganged. Each input circuit consists of two anti-resonant circuits capacity-coupled to each other and to the transmission line. It operates into the grids of the two first demodulator tubes which are paralleled. The first beating oscillator input is applied in push-pull between the cathodes of the tubes.

Oscillators and Automatic Tuning System

Three highly stable oscillators are used in the receivers, one for beating the signal frequency down to 2,900 kc., the second for beating it down further to 100 kc., and the third for use as a reference frequency for the automatic tuning system, or it may be used in the final demodulator when the automatic branch selector is used.

The first beating oscillator is of the coil-and-condenser type and covers the range from 7,000 kc. to 17,000 kc. Automatic tuning is used to compensate for long-time variations in its frequency, as well as any variations in the transmitter frequency, but every effort has been made to keep short-time variations to a minimum. The oscil-

lator is contained in a cast box mounted on rubber supports. The inductance coils have an extremely low temperature coefficient and are mounted on cast supports for rigidity. The condenser is also of rigid construction. A variation in plate voltage of one volt gives a frequency variation of about 6 cycles at 18,000 kc.

A buffer amplifier connects between this oscillator and a push-pull power amplifier which delivers ten watts of output. This output is delivered to a transformer located on the center bay of the row of first demodulator bays. Coaxial cables of equal length distribute it to the first demodulators on the adjacent bays. A vacuum tube voltmeter connected across this transformer gives an alarm if the voltage fails.

In the automatic tuning system the incoming carrier at approximately 100 kc. is beaten with the local 100 kc. oscillator. The phase of the beat frequency is then split and the resultant two-phase output applied to a motor which drives a condenser in the first beating oscillator circuit until the beat frequency is reduced to zero. In order to avoid interruption of control due to fading, all three diversity branch carriers are simultaneously connected to the circuit.

The second beating oscillator operates at 3,000 kc. and is a standard broadcast oscillator which has been slightly modified.

The 100 kc. oscillator is required to have the same frequency as the center of the pass band of the carrier filters. Since these are only 40 cycles wide both filters and oscillators are made with low temperature coefficient crystals. The oscillator is of the bridge type described by Meacham [8] but without temperature control.

Automatic Volume Control and Final Demodulators

Only the carrier is rectified for automatic volume control purposes. In the 100 kc. amplifiers where the carrier and sideband are amplified separately, separate automatic volume control circuits are used. The time-constant of the carrier amplifier control is made 0.1 second, which is as fast as is practical with the narrow carrier filter used, and the time-constant of the sideband volume control is made variable but is generally set at a value of 1 second.

With the common automatic volume control used with diversity receivers the rectifiers are so connected as to give outputs which vary according to the square or first power of the branch input. The sum of the rectifier outputs is held substantially constant. If the combined signal output is also to be held constant, the final demodulators of the

[8] L. A. Meacham, "The Bridge-Stabilized Oscillator," *Proc. I.R.E.*, Vol. 26, No. 10, p. 1278, October 1938; *B.S.T.J.*, Vol. XVII, No. 4, p. 574, October 1938.

branch circuits must follow the same law, i.e., if linear rectifiers are used, the final demodulators should be linear, and if square-law rectifiers are used, the final demodulators should be square-law. When linear demodulators are used the output noise is independent of the strength of the incoming carrier and since the gains of all branch amplifiers are the same the noise output of each branch will be the same, assuming that received noise does not vary with the vertical angle of reception. As a consequence the total noise will be equal to the product of a single branch noise and the number of branches regardless of the signal contributions of each branch. If square-law detectors are used, however, the noise output of a branch will go down when the carrier in the final demodulator of that branch goes down and consequently the total noise will be proportional to the total signal. In a three-branch diversity system a theoretical improvement in signal-to-noise ratio varying up to 4.77 db can be had by using square-law demodulators rather than linear. For this and other reasons square-law final demodulators have been used in this equipment.

When delay equalization is used between the various diversity branches it is essential that the received carrier be used in the final demodulation process. Small changes in the lengths of the paths in space traversed by the sidebands being received by the various diversity branches make the phases at random and if all branches were demodulated by a common carrier the audio outputs would not add in phase. By using the carrier arriving over each path for the demodulation of the accompanying sideband the random relation disappears and the audio outputs can be added in phase.

When using an automatic branch selector which discretely chooses one branch at a time for connection to the line it is no longer necessary to consider phases in the diversity branches and a local carrier is used because it reduces output amplitude variations. With only one diversity branch connected to the output, only the corresponding volume control rectifier should be contributing to the automatic volume control voltage if the output volume is to be held as constant as possible. This result is obtained by putting a rectifier in the d.c. output lead of each branch volume control rectifier so that only the volume control rectifier having the highest amplitude will supply current to the load resistance.

Delay Circuits and Switches

On account of the fact that waves arriving at different vertical angles have taken different times in transit it is necessary to insert

delay compensation in two of the three diversity branches when they are all connected to the output at once so that the audio outputs will add in phase. The branch receiving the waves at the highest angle, which are always the waves which have traversed the greatest distance, does not contain a delay circuit. The other two branches contain variable delay circuits having a maximum of 2,768 microseconds delay in steps of 31 microseconds. The band covered by the delay circuits is 6,000 cycles, the same as the width of the filter in the last intermediate frequency amplifier. The delay steps were made small enough so that it would be technically possible to phase properly over the entire band within less than a quarter cycle provided that the phase distortions of the transmission path, and the other parts of the receiver made it practical to do so.

The delay circuits each consist of 8 units having a delay of 31 microseconds and 9 units having a delay of 280 microseconds. Hand operated switches are provided to vary the delay in the usual decade switch manner. A motor driven switch is also provided which is arranged with two shafts connected by an intermittent movement so that when the eight small units of delay have been added a continued movement of the shaft removes these units from the circuit and simultaneously connects one of the large units, which is equivalent to nine small units. Further movement in the same direction successively adds in the smaller units again.

Automatic Delay Adjusting Circuits

A block schematic of the automatic delay adjusting apparatus as well as the automatic angle adjusting and recording equipment is shown in Fig. 8.

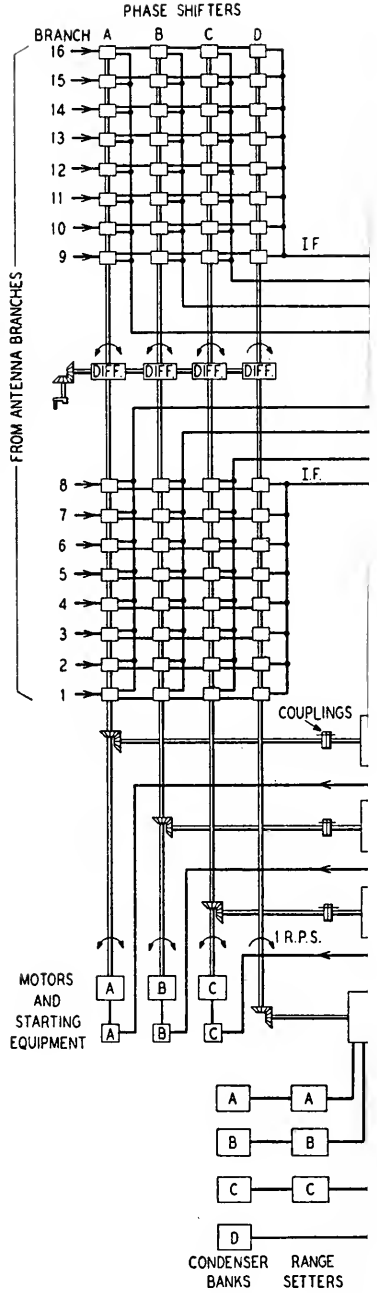
The proper delay compensation to phase the output of two diversity branches can be determined by connecting the outputs of the two branches to the two pairs of plates of a cathode ray oscilloscope and varying the delay in the lower angle path until a straight line pattern is obtained on the oscilloscope screen. This adjustment is facilitated by the restriction of the band in each branch at the oscilloscope to about an octave. If only a single frequency were used for this adjustment several positions of the delay adjustment might be found to give a straight line on the oscilloscope, the number of positions depending upon the frequency. In order that there may be only one position when the maximum delay is 2,768 microseconds the phase shift caused by this delay must be less than 180°, and consequently the frequency of operation must be less than 180 cycles. Where a band of frequencies a few hundred cycles wide is used, however, this difficulty is not

encountered and there is only one adjustment of the delay which gives a straight line on the oscilloscope.

With the cathode ray oscilloscope there is no indication as to whether the delay in the circuit is too small or too great, but only that it is either correct or incorrect. A direct indication of whether the delay should be increased or decreased can be obtained by connecting one diversity branch to the push-pull input of a balanced modulator and another diversity branch to the parallel input of the same modulator, after having shifted the phase between the two branches by 90°. When the two receiver branch outputs are in the same phase the two modulator input voltages will add in quadrature and the currents in the plate circuits of the two modulators will be the same, but when the phases are not the same the current in one plate circuit will be greater and the other less than in the in-phase case, the sense of the unbalance depending upon whether too little or too much delay is in the circuit. A center-zero meter in a bridge connection in the plate circuits therefore can be made to indicate the sense of the necessary correction. By substituting a voltmeter relay for the indicating meter a motor drive of the variable delay circuit can be operated in such a manner as to adjust the delay to the correct value.

This automatic equipment must operate satisfactorily with circuits having types of privacy in which the energy bearing components of speech are shifted from their normal position in the frequency spectrum. The equipment is made, therefore, to operate on a band of frequencies from 250 to 750 cycles. Volume limiters are provided which keep the input to the automatic equipment from this band substantially constant. If the delay is never incorrect by more than 1,000 microseconds, which has been found to be true under all conditions of normal operation, the automatic equipment will bring it to the correct value. For greater errors the equipment tries to set the delay at values about 2,000 microseconds higher or lower than the correct value.

Since the delay adjustment operates on speech, the relay operation will be intermittent at a syllabic rate and the motor drive relay system must incorporate a suitable hangover circuit to keep the motor operation as constant as possible when the delay is far from the proper value, and still not cause an over-run when the proper adjustment is reached. Freedom from over-run caused by motor inertia is obtained by using a motor having a multiple-pole permanent magnet armature with a speed of 75 r.p.m., which under no load will brake itself in about twenty degrees of armature travel.


PHASE SHIFTERS
BRANCH A B C D
16
15
14
13
12
11
10
9
IF
DIFF.
DIFF.
DIFF.
DIFF.
FROM ANTENNA BRANCHES
8
7
6
5
4
3
2
1
I.F.
COUPLINGS
I.R.P.S.
MOTORS AND STARTING EQUIPMENT
A
B
C
A
B
C
A
A
B
B
C
C
D
CONDENSER BANKS
RANGE SETTERS

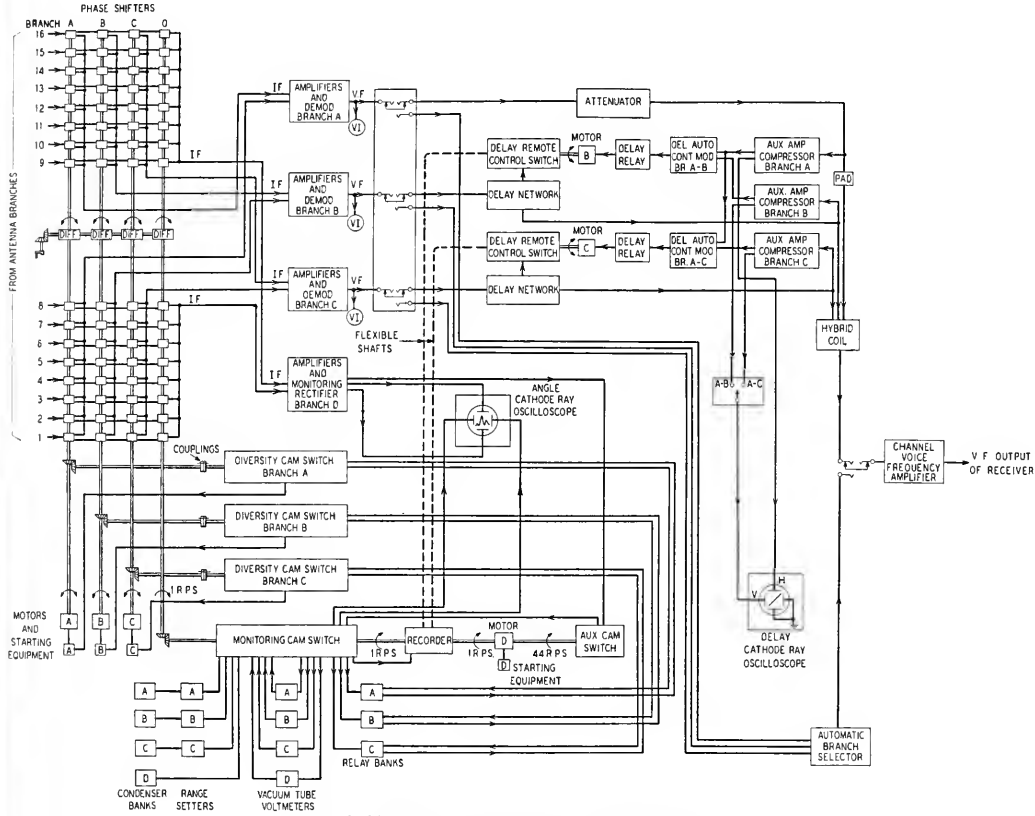


Fig. 8—Block schematic diagram of automatic adjusting circuits.

Automatic Branch Selector

As an alternative to the combination of the outputs of the diversity branches after delay equalization it is possible to use a system which discretely chooses one branch for connection to the line. Equipment to do this has been provided.

The common automatic volume control which is used with both systems operates on the carrier and if the carrier and sideband fade alike the output volume is held substantially constant. To the extent that the received noise on the various branches is the same, the branch having the highest output volume will be the most desirable to use. In the equipment used part of the audio output of each of the three diversity branches is rectified and applied differentially to three polar relays in such a manner that the relay corresponding to the branch having the highest amplitude is operated. This relay reduces the bias of an amplifier which connects between that branch and the output line from a high to a normal value, and thus permits voice frequency signals to flow through that branch to the output. The noticeable effects of switching are eliminated by several expedients. Push-pull amplifiers with feedback are used so as to balance out the low-frequency thump. The variation of biasing takes place with a time-constant of 0.01 seconds in order to aid in this matter as well as to render unnoticeable the instantaneous differences in the two channels. A biasing winding on each relay insures that once a contact is broken the relay moves to the opposite contact in a fixed time, which permits the selection of time-constants for the suppression and build-up which are as nearly complementary as possible, and so keeps the volume constant.

A difference in volume of 1 to 2 db is required to cause a switch. During the switching period, which lasts about 20 milliseconds, the output varies about $\pm$ 1 db. When no speech is being transmitted the relays remain inoperative and consequently the line may not be connected to the branch which at the next instant may deliver the highest volume. It might be expected that clipping of the succeeding initial syllable would be intolerable. To reduce selective fading to an unnoticeable amount it is only necessary to suppress the unwanted branches by 12 to 15 db. With the equipment adjusted to give this suppression it is found that there is always sufficient signal transmitted through one or more of the branches to practically eliminate noticeable clipping.

The use of the automatic branch selector has the disadvantage that the effect of having more than one diversity branch contribute

Automatic Branch Selector

As an alternative to the combination of the outputs of the diversity branches after delay equalization it is possible to use a system which discretely chooses one branch for connection to the line. Equipment to do this has been provided.

The common automatic volume control which is used with both systems operates on the carrier and if the carrier and sideband fade alike the output volume is held substantially constant. To the extent that the received noise on the various branches is the same, the branch having the highest output volume will be the most desirable to use. In the equipment used part of the audio output of each of the three diversity branches is rectified and applied differentially to three polar relays in such a manner that the relay corresponding to the branch having the highest amplitude is operated. This relay reduces the bias of an amplifier which connects between that branch and the output line from a high to a normal value, and thus permits voice frequency signals to flow through that branch to the output. The noticeable effects of switching are eliminated by several expedients. Push-pull amplifiers with feedback are used so as to balance out the low-frequency thump. The variation of biasing takes place with a time-constant of 0.01 seconds in order to aid in this matter as well as to render unnoticeable the instantaneous differences in the two channels. A biasing winding on each relay insures that once a contact is broken the relay moves to the opposite contact in a fixed time, which permits the selection of time-constants for the suppression and build-up which are as nearly complementary as possible, and so keeps the volume constant.

A difference in volume of 1 to 2 db is required to cause a switch. During the switching period, which lasts about 20 milliseconds, the output varies about $\pm$ 1 db. When no speech is being transmitted the relays remain inoperative and consequently the line may not be connected to the branch which at the next instant may deliver the highest volume. It might be expected that clipping of the succeeding initial syllable would be intolerable. To reduce selective fading to an unnoticeable amount it is only necessary to suppress the unwanted branches by 12 to 15 db. With the equipment adjusted to give this suppression it is found that there is always sufficient signal transmitted through one or more of the branches to practically eliminate noticeable clipping.

The use of the automatic branch selector has the disadvantage that the effect of having more than one diversity branch contribute

to the output at any one time is lost. This is estimated to be equivalent to not more than an increase of one decibel of transmitted power. To offset this there must be considered the possibility of the delay being out of adjustment for brief periods when changes in angle require sudden changes in delay equalization. The necessity for close phase uniformity between the various carrier and sideband amplifiers over a wide range of automatic volume control voltage is also eliminated when the automatic branch selector is used. Further, it is no longer necessary to use the received carriers for demodulation of the various branch outputs and a locally generated carrier of uniform amplitude can be used with a resultant increased stability of output volume. The practicability of trying to phase branch outputs by delay equalization over a range of more than 3,000 cycles from the carrier has not been demonstrated and consequently the use of automatic branch selection with the channel whose sideband is spaced by one sideband width from the carrier is to be recommended. The use of the branch selector also permits simpler operation of the automatic angle adjusting equipment as will be explained later. For all of these reasons it is to be expected that the automatic branch selector may eventually be used to the exclusion of the delay compensating equipment.

Automatic Angle Adjusting Equipment

In the experimental musa receiver the rectified carrier output of the monitoring branch was connected to one set of deflection plates of a cathode ray oscilloscope and a sweep circuit mounted on the monitoring phase shifter shaft connected to the other set of deflection plates. The oscilloscope screen displayed a graph of the amplitude of signal received for each phase shifter setting. The pattern frequently changed rapidly from moment to moment so that only by constantly observing the screen was it possible to determine at what phase shifter settings the best signals were being received and to set the diversity phase shifters accordingly.

The attention necessary to operate satisfactorily the equipment in this manner was believed to be too great for commercial operation, particularly since it might vary widely from hour to hour. With a mind to the fact that improper adjustment might give poorer reception than would be obtained with ordinary receivers it was decided to make the settings of the phase shifters of the diversity branches automatic.

In Fig. 8 the motors A, B, C and D drive the phase shifters of the corresponding branches through the vertical shafts. Motor D operates continuously so as to vary the phase shifting system through its complete range once a second, while motors A, B and C operate

only when a change in the angle of reception is required. Connected to the vertical drive shafts by the horizontal shafts are the three diversity cam switches and the monitoring cam switch.

The incoming carrier in the monitoring branch *D* is amplified in such a manner as to keep the average peak amplitude constant. It is then rectified, and applied to the vertical deflection plates of a cathode ray oscilloscope in the same manner as in the experimental equipment, the monitoring cam switch being provided with a set of contacts and resistances which act as a sweep circuit for the horizontal plates of the oscilloscope.

The rectified signal from the monitoring rectifier is also connected through the auxiliary cam switch, the monitoring cam switch, a high resistance, and the range setters, to three separate banks of condensers, each consisting of 44 four-microfarad condensers. The condensers in each bank are connected successively to the rectifier circuit once each second for a short period by the cam switch so that each is charged at a rate depending upon the amplitude of the received signal for a particular phase shifter position. A vacuum tube voltmeter is connected successively across each condenser of a bank. When one condenser becomes charged to one-half volt or more during the preceding second the vacuum tube voltmeter operates a relay. With the Branch *A*, *B*, and *C* voltmeters this results in a relay corresponding to that particular condenser being locked up and all the condensers in that particular bank being discharged. The operation of the second relay causes the motor of the corresponding diversity branch to start and turn in the right direction so that the branch phase shifters are adjusted with the least movement to the position corresponding to the relay, at which point a contact in the diversity cam switch trips the relay and stops the motion.

If no further control were provided all the diversity branch circuits would be set in the same position and consequently no diversity action would be obtained. To prevent this the range setters, *A*, *B*, and *C*, have been provided which are operated manually to limit the angular range of operation of each diversity branch. These switches merely short-circuit the condensers of a particular branch in the range which it is not desired to use. Since the short-circuited condensers do not acquire a charge the automatic adjusting equipment will never move the phase shifters to a position corresponding to a short-circuited condenser in that particular branch.

In setting the range switches when using delay equalization it is necessary to know what the condenser position is which corresponds to the highest angle which it is possible to receive at the particular

frequency being used. The short circuit will then be removed from this condenser and from the condensers representing successively lower angles in Branch *A* until it seems probable from the recorder or cathode ray oscilloscope pattern that a good signal will be received in that branch. The remaining condensers in that branch are left short-circuited. The short-circuits are then removed in a like manner from part of the remaining range for Branch *B* and in the remainder of the range for Branch *C*, the best division line being determined from the cathode ray oscilloscope or recorder pattern. This procedure is necessary inasmuch as branch *A* is used for a reference in adjusting the audio frequency delay compensation and must always have a satisfactory signal if that equipment is to operate.

When using the automatic branch selector other settings of the range switches are possible. One arrangement is to allow one branch to stop on even numbered contacts for a part of the range and another branch to stop on the remainder of the even numbered contacts. The third branch may then stop on any odd numbered contact. This permits two diversity branches to be set in the range of maximum signal. A difference of one contact has been found sufficient to give satisfactory diversity action in most cases and the recorder pattern always shows that the signal is more than one contact wide. The third branch is free to follow a signal in another part of the range, which may grow to be the strongest at any moment.

In order to improve the accuracy of this equipment and reduce the maintenance of the monitoring cam switch an auxiliary high speed cam switch is used which operates 44 times faster than the main switch, closes just after each contact of the monitoring switch, and opens just before each contact opens. The charging time for all condensers in a given bank is thus determined by the same cam and set of contacts.

To prevent over-running on the diversity branches from mechanical inertia a special motor is used. This motor is similar to the one used for automatic delay adjusting equipment.

At times there may be only one angle at which a satisfactory signal may be arriving. It is possible to get diversity action at these times by setting the diversity branches on opposite sides of the average angle of best reception. Provision is made for doing this with the automatic equipment by allowing one bank of condensers and one voltmeter relay to control all three diversity branches and then mechanically off-setting the phase shifters of two branches by means of adjustable couplings in the diversity cam switch drives.

It will be seen that to obtain accurate operation of the automatic angle adjusting equipment it is necessary that the charging voltage

should be large as compared with the final voltage on any condenser and that the final voltage must be the sum of a number of charges. The time necessary for a condenser to reach the final voltage can be varied from 8 to 45 seconds or longer and successive movements of a diversity branch phase shifter drive will not be oftener than this. Once the motor has started, however, it will move the phase shifter shaft through an angle of 180°, the maximum which would ever be necessary, in 6 seconds.

Recorder

In order properly to set the phase shifters manually or to set the range adjusters of the automatic angle adjusting equipment it is necessary to know the phase shifter positions corresponding to the angles at which signals are arriving. The angle monitoring cathode ray oscilloscope shows how the signal amplitude *vs.* phase shifter position varies from second to second. By using a retentive screen on the oscilloscope it is possible to see the traces for the previous few seconds at the same time as the most recent trace. The traces, however, normally vary appreciably in position of maximum amplitude and it is somewhat difficult to form an opinion from looking at the oscilloscope as to just where to set the diversity branches. By integrating the value of received signal over a number of seconds a better conception can be obtained.

In addition to the cathode ray oscilloscope it also seemed desirable to have a record available to the operator of the variation of signal intensity with phase shifter position as it changes from minute to minute so that he would not continuously have to observe the oscilloscope to determine whether the range adjusting switches were set properly. This required one more variable to be considered than the ordinary recorder is designed to register and it was consequently necessary to devise a new type of device.

The scheme of recording operates in a somewhat similar manner to the automatic angle adjusting equipment. A set of 44 condensers is charged by the incoming signal through the monitoring switch. Each condenser corresponds to a particular position of the phase shifters and consequently with a particular vertical angle of arrival at a particular frequency. A vacuum tube voltmeter is successively connected to the condensers until one is found which has acquired a predetermined potential in the order of two volts. A relay in the plate circuit of the voltmeter then operates, causing only that particular condenser to be discharged and making a record on a paper strip.

The recorder consists of a mechanism for driving a paper strip 5″ wide at a constant speed over a drum having a spiral wire on its

periphery. Above the drum and paper are a typewriter ribbon and a thin bar which may be made to come down on the ribbon, paper tape, and spiral wire, by the action of an electromagnet. The action of this striker bar is to cause a dot to be made on the paper strip at the position where the striker bar and spiral wire intersect. The drum carrying the spiral wire revolves in synchronism with the phase shifters and there is consequently a lateral position on the paper corresponding to each one of the 44 condensers previously mentioned. When each condenser is discharged by the action of the vacuum tube voltmeter a dot is made in a particular lateral position on the paper strip and successive dots caused by the discharge of the same condenser fall in the same longitudinal line on the strip. The frequency of dots in a particular longitudinal line is, therefore, proportional to the relative field strength at a vertical angle corresponding to that line. As a result of the action of the automatic volume control on the monitoring branch amplifier, the maximum frequency of dots along a longitudinal line is kept approximately constant regardless of the absolute value of signal received so that the device does not record the variation of signal at a fixed angle from minute to minute.

A sample of a section of a record is shown on Fig. 9, together with a scale showing the angles corresponding to the rows of dots for a particular received frequency. The angle record is contained in the section above the "Phase Shifter Position" scale.

In order to have a check on, and a record of, the operation of the automatic angle adjusting equipment, provision has been made so that three longitudinal lines are drawn on the paper corresponding to the three angular positions of the diversity branches.

A record of the operation of the automatic delay adjusting device is also made by the recorder. This was done by inserting a mechanism which uses the margins of the paper on either side of the main record. The delay recording device consists of two drums mounted concentrically with the main recorder drive shaft which are similar in nature to the tens and units drums of an ordinary counter. Flexible shafts extend from the delay adjusting switches to the recorder where they drive the drums on each end of the shaft. The two drums on one end are connected together with an intermittent movement so that one revolution of the small units drum causes the large units drum to move forward one step.

With the paper tape normally running at only $\frac{1}{4}''$ per minute it is impractical to stamp numbers on the paper since the delay adjustment varies several times in a minute and thus would cause the numbers to record on top of one another. Recourse was accordingly taken to a

mark in a definite lateral position to indicate the magnitude of the delay. The drums have segmental ridges on their periphery which are displaced in various lateral positions. Cam operated hammers descend on the typewriter ribbon above the drums and paper once each second leaving a mark in a lateral position corresponding to the

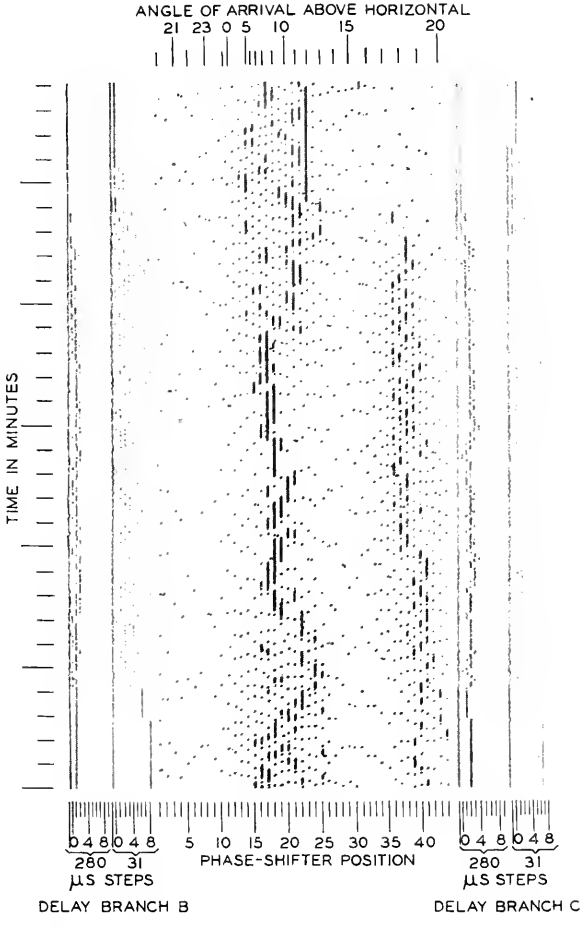


Fig. 9—Sample of musa recorder chart. Station GAW. Freq. 18,200 Kc.

segmental ridge which is beneath and consequently to the delay setting. Two reference lines are used to facilitate the reading of the delay values.

Operation and Performance

The musa system can be expected to give an improvement in signal-to-noise ratio and in selective fading over a receiver using only

a single antenna. The improvement in signal-to-noise ratio caused by the use of 16 antennas should average 12 db, but instantaneous improvements might vary from large negative values, if the equipment were not kept properly adjusted, to values of 25 or 30 db, which might be expected when the noise came from the direction of a null in the musa antenna directive diagram.

In the operation of radio telephone circuits there is a minimum signal-to-noise ratio below which commercial service cannot be given. As the signal-to-noise ratio is increased a value is reached where further increases give little benefit. The range between these two values is about 25 db. Transmitters and receivers generally are designed so that their maximum signal-to-set-noise ratio is somewhat greater than the maximum beneficial circuit value in order that set-noise shall not degrade the circuit. The maximum signal-to-noise ratio obtainable with a musa receiver and a single-antenna receiver should be approximately the same.

The 12 db average improvement which the musa receivers should give should make it possible to obtain, on the average, commercial circuits with signal field strengths 12 db less than those usable with a single-antenna receiver. This in turn will decrease the amount of time in which commercial service cannot be given.[9] On the other hand the musa receiver should produce its maximum signal-to-noise ratio for field strengths 12 db lower than a single-antenna receiver and at field strengths 12 db higher the musa receiver would show no improvement over the single-antenna receiver. The net improvement in signal-to-noise ratio therefore should be expected to average from 12 db at the lowest usable signal-to-noise ratios to 0 db at fields 25 or 30 db higher, with fairly wide variations with time from the average.

The results of a comparison between the musa system and a single sideband receiver operating from one of the same antennas confirm the theoretical expectations to a fair degree. The fraction of the time that given improvements in decibels are obtained follows approximately a normal probability curve.

The reduction in selective fading effected by the musa receivers is difficult to state numerically. Most of the objectionable selective fading is removed. There are times when waves that have traveled over distinctly different paths arrive at so nearly the same angle that they cannot be resolved. Fortunately these times are fairly rare. When waves of closely adjacent angle are present the monitoring system does not give a true indication of reception angles, as can be

[9] R. K. Potter and A. C. Peterson, Jr., "Reliability of Short-Wave Radio Telephone Circuits," *Bell Sys. Tech. Jour.*, Vol. XV, pp. 181–196, April 1936.

shown by theory. It is a fairly common occurrence for a wave group which has apparently traveled over a single general path to have components which vary in transmission times by 100 or 200 microseconds from others of the same group. The fading caused by such small delay differences is not distinctly selective in effect and its chief detriment is in causing volume variations which must be overcome by the use of special devices.

During some severe magnetic storms successive traces on the angle monitoring cathode ray oscilloscope show little relation to each other. It has been reasoned that a reduction in resolution might be beneficial at such times but sufficient experience to prove this has not been obtained. A reduction in resolution can be obtained easily by switching off the amplifiers associated with one group of eight antennas. It has been found that fading on the front group of antennas is generally at random to that on the rear group so that the two groups can be used in space diversity if desired. No particular advantage has been found to this arrangement.

The use of delay compensation between the diversity branches does not seem to have any advantage over the use of the automatic branch selector. The output volume variations are slightly greater with delay compensation because of the use of reconditioned carrier for demodulation. On the other hand the use of the automatic branch selector promises materially to reduce maintenance by eliminating the necessity for keeping the phases of the various carrier and sideband amplifiers alike.

It has been amply demonstrated that the automatic adjusting features provided are essential to the efficient operation of the equipment.

Abstracts of Technical Articles by Bell System Authors

Thermionic Emission, Migration, and Evaporation of Barium on Tungsten.[1] J. A. Becker and G. E. Moore. When barium is deposited on tungsten, the thermionic activity of the tungsten increases, comes to a maximum, and then decreases. It has frequently been found by emission measurements that this optimum corresponds to about a monomolecular layer. However, data obtained in this work show that some regions of the filament require more than five times as much barium as others for optimum emission.

Photographs are presented which show that the rates of both migration and evaporation depend on the crystal surface, the temperature, and amount of barium on the surface. Barium migration on tungsten can be observed at temperatures as low as 970° K., is readily observed at 1025° K., and is rapid at 1070° K. Evaporation is observed on some crystals at temperatures as low as 1025° K., while on others it is slow even at 1260° K. At 1300° K. it is rapid for all crystals. These temperatures probably vary with the oxygen contamination which comes over to the filament with the barium. For barium concentrations near the optimum there exists a range of temperature over which migration is readily observed, but where evaporation is not noticeable.

Measurements of electron emission after all the barium is evaporated show that the filament was contaminated by an electronegative material, probably oxygen.

Barium tends to migrate toward the negative end of the filament, thus indicating ionization of adatoms.

A mechanism for migration is suggested.

The Vocoder—Electrical Re-creation of Speech.[2] Homer Dudley. In the Bell Telephone Laboratories have been developed electrical circuits for the artificial production of speech. One form of the device is itself voice-controlled, thus differing fundamentally from the Voder of the World's Fair which is controlled by keys and pedals. It has been christened the "Vocoder" or "voice coder."

Many startling effects are possible when the code is varied, for the Vocoder then re-creates sounds quite different from those used by the person speaking. Cadences may become monotones, rising inflections may be turned to falling inflections, a vigorous voice may become a

[1] *Philosophical Magazine*, February 1940.
[2] *Jour. S. M. P. E.*, March 1940.

quaver, or a single voice may accompany itself at any desired musical interval—thus converting a solo into a duet, etc. Also non-speech sounds may be coded into intelligible speech and instrumental music into vocal music.

Statistical Measurements on Conversational Speech.[3] H. K. DUNN and S. D. WHITE. Using apparatus designed to collect a large number of data in a short time, the following measurements have been made: peak and r.m.s. pressures in one-eighth-second intervals, and in various bands of frequencies up to 12,000 cycles per second, from the voices of six men and five women; comparison of r.m.s. pressures in one-eighth- and one-fourth-second intervals, from a single male voice; and distribution of the instantaneous pressures in whole speech, from a single voice. Derived from these data are peak factors in one-eighth-second intervals, and frequency distribution of speech energy in long intervals. Both the absolute value and the distribution of energy are found somewhat different from previously published results.

Auditory Patterns.[4] HARVEY FLETCHER. During the last two decades considerable progress has been made in understanding the hearing processes taking place when we sense a sound. The application of the same instrumentalities that have brought such a wonderful development in the radio and sound pictures to this problem is largely responsible for this progress. Such instrumentalities have made it possible to make accurate measurements which are the basis for understanding any physical process.

To understand this problem then we need to know first how to describe and measure the sound reaching the ears; then we need to know how to describe and measure the sensations of hearing produced by such a sound upon a listener. To do this quantitatively we must also know the degree and kind of hearing ability possessed by the listener. It is with these three phases of the problem that this paper deals.

[3] *Jour. Acous. Soc. of America,* January 1940.
[4] *Reviews of Modern Physics,* January 1940.

Contributors to this Issue

KARL K. DARROW, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

W. P. MASON, B.S. in Electrical Engineering, University of Kansas, 1921; M.A., Columbia University, 1924; Ph.D., 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged in investigations on carrier systems and in work on wave transmission networks both electrical and mechanical. He is now head of the department investigating piezoelectric crystals.

A. L. MATTE, B.S. in Electrical Engineering, Massachusetts Institute of Technology, 1909; Graduate Studies, M.I.T., 1912–13. New England Investment and Securities Company, 1910–12; Detroit United Railways, 1913–18. American Telephone and Telegraph Company, Department of Development and Research, 1918–34; Bell Telephone Laboratories, 1934–. Mr. Matte has been engaged principally in transmission studies relating to carrier telegraphy.

G. S. PHIPPS, B.S. in Electrochemical Engineering, Pennsylvania State College, 1930; M.S. in Metallurgy, Columbia University, 1939. Bell Telephone Laboratories, 1930–. Mr. Phipps has been engaged principally in the metallurgical investigation of solders and lead base alloys.

F. A. POLKINGHORN, B.S., University of California, 1922; U. S. Naval Radio Laboratory at Mare Island Navy Yard, California, 1922–24; A-P Radio Laboratories, San Francisco, 1924–25. Pacific Telephone and Telegraph Company, San Francisco, 1925–27; Bell Telephone Laboratories, 1927–. Mr. Polkinghorn's work has been primarily in connection with the design of radio receiving and test equipment for use at high and ultra-high frequencies.

EARLE E. SCHUMACHER, B.S., University of Michigan; Research Assistant in Chemistry, 1916–18. Engineering Department, Western Electric Company, 1918–25; Bell Telephone Laboratories, 1925–. As

Associate Research Metallurgist, Mr. Schumacher is in charge of a group whose work relates largely to research studies on metals and alloys.

A. M. SKELLETT, A.B., 1924, M.S., 1927, Washington University; Ph.D., Princeton University, 1933; Instructor, 1927–28, Assistant Professor of Physics, 1928–29, University of Florida. Bell Telephone Laboratories, 1929–. Dr. Skellett, formerly engaged in investigations pertaining to the transatlantic radio telephone, is concerned with applications of electronic and ionic phenomena.

R. A. SYKES, Massachusetts Institute of Technology, B.S. 1929; M.S. 1930. Columbia University, 1931–33. Bell Telephone Laboratories, Research Department, 1930–. Mr. Sykes has been engaged in the application of piezoelectric crystals to selective networks, and more recently in the use of coaxial lines as filter elements.

EMIL FRIDSTEIN VAAGE, E.E., Technical University of Darmstadt, Germany, 1921–26; M.S., Brooklyn Polytechnic Institute, 1932. Elektrisk Bureau, Oslo, Norway, 1926–27. American Telephone and Telegraph Company, 1927–34; Bell Telephone Laboratories, 1934–. Mr. Vaage has been engaged in inductive coordination studies, 1927–39. Since 1939 his work has been concerned with open-wire television transmission.

L. G. WADE, B.S.E.E., University of Idaho, 1918; M.S., Cornell University, 1924. Engineering Department, Western Electric Company, 1924. Mr. Wade has been engaged in developing equipment and processes for the manufacture of lead covered telephone cable, particularly on drying and dry storage of cable cores preceding the lead sheathing operation.

VOLUME XIX JULY, 1940 NUMBER 3

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XIX *July, 1940* *No. 3*

Crosstalk in Coaxial Cables—Analysis Based on Short-Circuited and Open Tertiaries

By K. E. GOULD

The problem considered herein is that of estimating, from measurements on short lengths of coaxial cable, the crosstalk to be expected in long lengths of the same cable. The method developed, which is particularly applicable to cases in which the effect of tertiary circuits on the crosstalk is large, is based on measurements of crosstalk in a short length, with the tertiaries first short-circuited and then open. The application of this method to the cable described in the companion paper by Messrs. Booth and Odarenko gave crosstalk values in good agreement with their experimental results.

Introduction

FOR a number of years the problem of crosstalk summation in long open-wire lines or cables has been studied by measuring crosstalk, in phase and magnitude, in short lengths. The crosstalk within a short length, between two circuits terminated in their characteristic impedances, would be measured with all important tertiary circuits also approximately terminated. Then the crosstalk between two circuits in adjoining short lengths would be measured with the tertiary circuits terminated. From these coefficient measurements the crosstalk in a long length could be estimated by a process of integration.

The application of this general method to crosstalk in the usual types of coaxial cable would require great accuracy in the coefficient measurements, because in longer lengths the desired crosstalk value depends on the difference between two nearly equal quantities involving the coefficients. In the following analysis the computation of the crosstalk for long lengths of coaxial cable is based on crosstalk measurements, in phase and magnitude, between two coaxial circuits in a single short length with the tertiary circuits first open and then short-circuited, no crosstalk measurements with terminated tertiary circuits being involved.

This method of analysis, when applied to the twin coaxial cable described in the companion paper by Messrs. Booth and Odarenko, gave results in good agreement with the measured crosstalk.

In this analysis it was assumed that all the tertiary circuits could be combined and considered as a single circuit. Although no evidence has been found that with the types of structure studied so far, better accuracy would result from the further refinement of considering two or more dissimilar tertiary circuits with coupling between them, there is one case of practical importance which cannot be handled with the single-tertiary analysis. This case is that of the interaction crosstalk (that is, the crosstalk by way of a tertiary circuit) between two adjoining lengths of coaxial cable, when, at the junction, part of the tertiary conductors are short-circuited to the outer coaxial conductors while the remaining tertiary conductors continue through with no discontinuity. This problem might be of importance where, at a repeater, the outer coaxial conductors and the sheath are bonded together, but paper-insulated pairs in the same sheath provide an uninterrupted tertiary circuit. The near-end crosstalk under such conditions might also differ significantly from the values indicated by the single-tertiary analysis.

The two-tertiary analysis is too long to be given here in detail, and hence has been outlined only to such an extent as to indicate the derivation of the formulas for interaction crosstalk when one of the tertiaries is short circuited and the other terminated in its characteristic impedance. The formula for near-end crosstalk under this condition is given without derivation.

I—Identical Coaxial Lines Symmetrically Placed with Respect to a Single Tertiary

The first case we shall consider herein is that of any number of identical coaxial lines with the outer coaxial conductors in continuous electrical contact and symmetrically placed with respect to a single tertiary circuit, such as that which might be provided by a sheath surrounding the coaxial lines and insulated from the outer coaxial conductors, or by a surrounding layer of paper-insulated pairs. Throughout we shall assume that the reaction of the induced currents upon the disturbing line is negligible.

Following a nomenclature analogous to that of the Schelkunoff-Odarenko paper,[1] we will designate by Z_{12} the mutual impedance per unit length between any two coaxial lines in the presence of the other coaxial lines but in the absence of any other conductors. The mutual

[1] *Bell System Technical Journal*, April, 1937.

impedance per unit length between any coaxial line (in the presence of the other coaxial lines) and the tertiary circuit consisting of all of the coaxial outer conductors with return by way of the sheath or other tertiary conductors, we will designate by Z_{13}.

If we consider the crosstalk between two coaxial lines of length, l, such that the coaxial lines and the tertiary are electrically short, each coaxial line being terminated in its characteristic impedance Z and the tertiary open at each end, the crosstalk (near-end and far-end being identical for such a length) is given by

$$\frac{Z_{12}l}{2Z}.$$

If, now, we consider a case similar except that the tertiary is short-circuited at each end, the crosstalk is the above term plus the effect of the tertiary current I_3, which, for unit current in the disturbing coaxial line, is given by

$$I_3 = \frac{Z_{13}}{Z_{33}},$$

where Z_{33} is the series impedance of the tertiary circuit per unit length. This tertiary current will produce a current $\left(-\frac{Z^2_{13}l}{2ZZ_{33}} \right)$ in the disturbed coaxial line and the total crosstalk will be

$$\frac{Z_{12}l}{2Z} - \frac{Z^2_{13}l}{2ZZ_{33}}.$$

If we designate $\frac{Z_{12}}{2Z}$ by X and $\frac{Z^2_{13}}{Z_{12}Z_{33}}$ by ξ, then, for an electrically short length, X will represent the crosstalk per unit length between two coaxial lines with the tertiary open, and $X(1 - \xi)$ the crosstalk per unit length with the tertiary short-circuited. In the formulas developed below these quantities will be found to be of fundamental importance.

Tertiary Terminated in its Characteristic Impedance

Far-End Crosstalk

From the Schelkunoff-Odarenko paper, the sum of the direct far-end crosstalk (eq. 19) and the indirect far-end crosstalk (eq. 40) for any length under these conditions gives the total far-end crosstalk F_l, as

$$F_l = \frac{Z_{12}l}{2Z} - \frac{Z^2_{13}}{4ZZ_3}\left[\frac{2\gamma_3 l}{\gamma_3^2 - \gamma^2} - \frac{1 - e^{-(\gamma_3 - \gamma)l}}{(\gamma_3 - \gamma)^2} - \frac{1 - e^{-(\gamma_3 + \gamma)l}}{(\gamma_3 + \gamma)^2} \right], \quad (1)$$

where

Z_3 = characteristic impedance of tertiary circuit,

γ, γ_3 = propagation constants of coaxial lines and tertiary circuit respectively.

This may be rearranged and written (since $Z_3\gamma_3 = Z_{33}$)

$$F_t = \frac{Z_{12}l}{2Z} - \frac{Z^2_{13}l}{2ZZ_{33}} - \frac{Z^2_{13}}{2ZZ_{33}} \times \left[\frac{l\gamma^2}{\gamma_3^2 - \gamma^2} - \frac{\gamma_3}{2}\left(\frac{1 - e^{-(\gamma_3-\gamma)l}}{(\gamma_3 - \gamma)^2} + \frac{1 - e^{-(\gamma_3+\gamma)l}}{(\gamma_3 + \gamma)^2} \right) \right] \quad (2a)$$

$$= X\left[l(1 - \xi) - l\xi\frac{\gamma^2}{\gamma_3^2 - \gamma^2} + \frac{\xi\gamma_3}{2} \times \left(\frac{1 - e^{-(\gamma_3-\gamma)l}}{(\gamma_3 - \gamma)^2} + \frac{1 - e^{-(\gamma_3+\gamma)l}}{(\gamma_3 + \gamma)^2} \right) \right]. \quad (2b)$$

This formula has been found to be applicable, with good accuracy, to the types of coaxial cable which have been studied so far. The quantities X and $X(1 - \xi)$ are determined from crosstalk measurements on a short length, and the propagation constants are of course readily determined.

Near-End Crosstalk

A similar approach to the problem of the near-end crosstalk N_t with the tertiary terminated in its characteristic impedance, using equations (10) and (32) of the Schelkunoff-Odarenko paper, gives

$$N_t = X\left[(1 - e^{-2\gamma l}) \left(\frac{1 - \xi}{2\gamma} - \frac{\xi\gamma}{2(\gamma_3^2 - \gamma^2)} \right) + \frac{\xi\gamma_3}{2(\gamma_3^2 - \gamma^2)} (1 + e^{-2\gamma l} - 2e^{-(\gamma_3+\gamma)l}) \right]. \quad (3)$$

Here, as in the case of equation (2b) above, the crosstalk may be computed readily from crosstalk and impedance measurements on a short sample.

Interaction Crosstalk

Far-End Far-End and Far-End Near-End

We will consider the interaction crosstalk between two adjoining sections of lengths l and l', respectively, the tertiary being connected through at the junction, with no discontinuity. The tertiary current $i_3(l)$ at the far end of a section of length l, for unit sending-end current, with the tertiary terminated in its characteristic impedance, is readily formulated as

$$i_3(l) = \frac{Z_{13}}{2Z_3}\frac{e^{-\gamma l} - e^{-\gamma_3 l}}{\gamma_3 - \gamma}. \quad (4)$$

In the adjoining section, with the tertiary terminated in its characteristic impedance, the tertiary current $i_3(y)$ will be given by $i_3(l)e^{-\gamma_3 y}$, where y is the distance measured from the junction of the two sections.

This tertiary current $i_3(y)$ will produce a far-end current in the disturbed coaxial of

$$\frac{Z^2_{13}}{4ZZ_3}\frac{e^{-\gamma l} - e^{-\gamma_3 l}}{\gamma_3 - \gamma}\int_0^{l'} e^{-\gamma_3 y}e^{-\gamma(l'-y)}dy.$$

The equal-level far end-far end interaction crosstalk FF, being this far-end current divided by $e^{-\gamma(l+l')}$, may be obtained as

$$FF = \frac{X\xi\gamma_3}{2(\gamma_3 - \gamma)^2}(1 - e^{-(\gamma_3-\gamma)l})(1 - e^{-(\gamma_3-\gamma)l'}). \tag{5}$$

The near-end current in the disturbed coaxial due to the current $i_3(y)$ is given by

$$\frac{Z^2_{13}}{4ZZ_3}\frac{e^{-\gamma l} - e^{-\gamma_3 l}}{\gamma_3 - \gamma}\int_0^{l'} e^{-\gamma_3 y}e^{-\gamma y}dy.$$

From this the equal-level far end-near end interaction crosstalk FN, being this near-end current divided by $e^{-\gamma l}$, may be obtained as

$$FN = \frac{X\xi\gamma_3}{2(\gamma_3^2 - \gamma^2)}(1 - e^{-(\gamma_3-\gamma)l})(1 - e^{-(\gamma_3+\gamma)l'}). \tag{6}$$

Near-End Near-End

The near-end tertiary current in the section of length l is similarly formulated as

$$i_3(0) = \frac{Z_{13}}{2Z_3}\frac{1 - e^{-(\gamma_3+\gamma)l}}{\gamma_3 + \gamma}. \tag{7}$$

The near-end near-end interaction crosstalk NN is readily obtained, in a fashion similar to that outlined above for the far-end near-end interaction crosstalk, as

$$NN = \frac{X\xi\gamma_3}{2(\gamma_3 + \gamma)^2}(1 - e^{-(\gamma_3+\gamma)l})(1 - e^{-(\gamma_3+\gamma)l'}). \tag{8}$$

Tertiary Short-Circuited

The general case of the crosstalk between coaxial lines of length l with the tertiary short-circuited at each end may be attacked as follows. At any point at a distance x from the sending end, the voltage gradient along the outer surface of the outer coaxial conductors, for unit sending-end current, will be $Z_{13}e^{-\gamma x}$. Each differential element, $Z_{13}e^{-\gamma x}dx$, of this voltage drop will produce a current in the tertiary circuit de-

termined by the impedances, Z' and Z'', of the tertiary as seen in the two directions from this point, these impedances being

$$Z' = Z_3 \tanh \gamma_3 x \tag{9}$$

and

$$Z'' = Z_3 \tanh \gamma_3 (l - x) \tag{10}$$

respectively toward and away from the sending end.

At any other point at a distance y from the sending end, the tertiary current due to the voltage $Z_{13}e^{-\gamma x}dx$ will be given, for $y > x$, by

$$i_3(y) = \frac{Z_{13}e^{-\gamma x}dx}{Z' + Z''} \frac{\cosh \gamma_3(l - y)}{\cosh \gamma_3(l - x)}. \tag{11}$$

From this the transfer admittance $A(x, y)$ between these two points is obtained as

$$A(x, y) = \frac{1}{Z_3} \frac{\cosh \gamma_3(l - y)}{\tanh \gamma_3 x \cosh \gamma_3(l - x) + \sinh \gamma_3(l - x)}. \tag{12}$$

Similarly, for $y < x$, this transfer admittance is obtained as

$$A(x, y) = \frac{1}{Z_3} \frac{\cosh \gamma_3 y}{\sinh \gamma_3 x + \cosh \gamma_3 x \tanh \gamma_3(l - x)}. \tag{13}$$

The tertiary current, $i_3(x)$, is given by

$$i_3(x) = \int_0^l Z_{13}e^{-\gamma y}A(x, y)dy \tag{14}$$

from which we obtain

$$i_3(x) = \frac{Z_{13}}{2Z_3 \sinh \gamma_3 l} \times \left[\begin{array}{l} \dfrac{1}{\gamma_3 + \gamma} \left[\begin{array}{c} \cosh \gamma_3(l - x) + e^{-\gamma x} \sinh \gamma_3 l \\ - e^{-\gamma l} \cosh \gamma_3 x \end{array} \right] \\ - \dfrac{1}{\gamma_3 - \gamma} \left[\begin{array}{c} \cosh \gamma_3(l - x) - e^{-\gamma x} \sinh \gamma_3 l \\ - e^{-\gamma l} \cosh \gamma_3 x \end{array} \right] \end{array} \right]. \tag{15}$$

Far-End Crosstalk

The indirect far-end crosstalk F_s' due to this tertiary current (eq. 15) is given by

$$F_s' = e^{\gamma l} \int_0^l \frac{Z_{13}i_3(x)}{2Z} e^{-\gamma(l-x)}dx \tag{16a}$$

$$= \frac{Z^2_{13}}{2ZZ_{33}} \left[\frac{l\gamma_3^2}{\gamma_3^2 - \gamma^2} - \frac{2\gamma_3\gamma^2}{(\gamma_3^2 - \gamma^2)^2} \frac{\cosh \gamma_3 l - \cosh \gamma l}{\sinh \gamma_3 l} \right]. \tag{16b}$$

If this is combined with the direct far-end crosstalk $\frac{Z_{12}l}{2Z}$, and the terms rearranged as in the case of equation (2b), the total far-end crosstalk F_s is obtained as

$$F_s = X\left[l(1-\xi) - l\xi\frac{\gamma^2}{\gamma_3^2-\gamma^2} + 2\xi\frac{\gamma_3\gamma^2}{(\gamma_3^2-\gamma^2)^2}\frac{\cosh\gamma_3 l - \cosh\gamma l}{\sinh\gamma_3 l}\right]. \quad (17)$$

It will be noted that equations (2b) and (17) differ only in the terms which are not proportional to the length and which thus are of decreasing importance as the length becomes great.

Near-End Crosstalk

The indirect near-end crosstalk N_s' due to the tertiary current $i_3(x)$ is given by

$$N_s' = \int_0^l \frac{Z_{13}i_3(x)}{2Z}e^{-\gamma x}dx \quad (18)$$

By substituting $i_3(x)$ from equation (15) herein, and combining the result with the direct near-end crosstalk,

$$\frac{Z_{12}}{2Z}\frac{1-e^{-2\gamma l}}{2\gamma}$$

we obtain the total near-end crosstalk N_s, which may be written in the form

$$N_s = X\left[\frac{1-e^{-2\gamma l}}{2\gamma}\left((1-\xi)+\frac{\xi\gamma^2(\gamma_3^2+\gamma^2)}{(\gamma_3^2-\gamma^2)^2}\right) - \left(\frac{\xi\gamma_3\gamma^2}{(\gamma_3^2-\gamma^2)^2}\right)\left(\frac{(1+e^{-2\gamma l})\cosh\gamma_3 l - 2e^{-\gamma l}}{\sinh\gamma_3 l}\right)\right]. \quad (19)$$

II—Identical Coaxial Lines Symmetrically Placed with Respect to Each of Two Dissimilar Tertiaries

We will now consider the case of any number of identical coaxial lines with the outer conductors in continuous electrical contact and symmetrically placed with respect to each of two dissimilar tertiaries with coupling between them.

In an unpublished memorandum by J. Riordan, the general forms are developed for the currents and voltages in two parallel circuits having uniformly distributed self and mutual impedances and admittances, when these circuits are subjected to impressed axial fields.

These currents (I_1 and I_2) and voltages (V_1 and V_2) (the subscripts applying, of course, to the respective tertiaries) are given by the coefficient array,

	$(a_1 + P_1)e^{-\gamma_1 x}$	$(b_1 + Q_1)e^{\gamma_1 x}$	$(a_2 + P_2)e^{-\gamma_2 x}$	$(b_2 + Q_2)e^{\gamma_2 x}$
I_1	1	-1	η_2	$-\eta_2$
I_2	η_1	$-\eta_1$	1	-1
V_1	K_1	K_1	$-\eta_1 K_2$	$-\eta_1 K_2$
V_2	$-\eta_2 K_1$	$-\eta_2 K_1$	K_2	K_2

where a_1, b_1, a_2 and b_2 are constants to be determined from the boundary conditions, and

$$\begin{matrix} P_1 \\ Q_1 \end{matrix} = \frac{1}{2K_1(1 - \eta_1\eta_2)} \int e^{\pm\gamma_1 x}(f_1 + \eta_1 f_2)dx, \tag{20}$$

$$\begin{matrix} P_2 \\ Q_2 \end{matrix} = \frac{1}{2K_2(1 - \eta_1\eta_2)} \int e^{\pm\gamma_2 x}(\eta_2 f_1 + f_2)dx, \tag{21}$$

f_1 and f_2 being the impressed fields along circuits 1 and 2 respectively.

If we consider the two tertiary circuits as consisting of (1) the outer coaxial conductors in parallel with return by tertiary path 1 and (2) tertiary path 1 − tertiary path 2, only tertiary circuit 1 will be subjected to an impressed field. Thus we will have $f_2 = 0$ and $f_1 = Z_{13}e^{-\gamma x}$ (for unit sending-end current in the disturbing coaxial line), where Z_{13} is the mutual impedance, per unit length, between a coaxial (in the presence of the other paralleling coaxials) and tertiary 1, and γ is the propagation constant, per unit length, for the coaxial circuit. The other quantities in this array are circuit parameters given as follows in terms of the series impedances Z_{11}, Z_{22} and Z_{12} per unit length and admittances a_{11}, a_{22} and a_{12} per unit length (subscripts 11 and 22 for self impedance or self admittance of circuits 1 and 2 respectively, and 12 for mutuals):

$$\begin{matrix} \gamma_1^2 \\ \gamma_2^2 \end{matrix} = \tfrac{1}{2}[a_{11}Z_{11} + a_{22}Z_{22} + 2a_{12}Z_{12} \pm ((a_{11}Z_{11} - a_{22}Z_{22})^2 + 4(a_{11}Z_{12} + a_{12}Z_{22})(a_{12}Z_{11} + a_{22}Z_{12}))^{1/2}], \tag{22}$$

$$\eta_1 = \frac{\gamma_1^2 - a_{11}Z_{11} - a_{12}Z_{12}}{a_{11}Z_{12} + a_{12}Z_{22}}, \tag{23}$$

$$\eta_2 = \frac{\gamma_2^2 - a_{12}Z_{12} - a_{22}Z_{22}}{a_{12}Z_{11} + a_{22}Z_{12}}, \tag{24}$$

$$K_1 = \frac{Z_{11} + \eta_1 Z_{12}}{\gamma_1} = \frac{\gamma_1}{a_{11} - \eta_2 a_{12}}, \tag{25}$$

$$K_2 = \frac{Z_{22} + \eta_2 Z_{12}}{\gamma_2} = \frac{\gamma_2}{a_{22} - \eta_1 a_{12}}. \tag{26}$$

From equations (20) and (21) above, we have

$$P_1 = \frac{Z_{13}}{2K_1(1 - \eta_1\eta_2)(\gamma_1 - \gamma)} e^{(\gamma_1 - \gamma)x}, \tag{27}$$

$$Q_1 = \frac{-Z_{13}}{2K_1(1 - \eta_1\eta_2)(\gamma_1 + \gamma)} e^{-(\gamma_1 + \gamma)x}, \tag{28}$$

$$P_2 = \frac{Z_{13}\eta_2}{2K_2(1 - \eta_1\eta_2)(\gamma_2 - \gamma)} e^{(\gamma_2 - \gamma)x}, \tag{29}$$

$$Q_2 = \frac{-Z_{13}\eta_2}{2K_2(1 - \eta_1\eta_2)(\gamma_2 + \gamma)} e^{-(\gamma_2 + \gamma)x}. \tag{30}$$

If we designate

$$\frac{Z_{13}}{K_1(1 - \eta_1\eta_2)(\gamma_1^2 - \gamma^2)} \text{ by } \psi_1$$

and

$$\frac{Z_{13}\eta_2}{K_2(1 - \eta_1\eta_2)(\gamma_2^2 - \gamma^2)} \text{ by } \psi_2,$$

we have

$$I_1 = a_1e^{-\gamma_1 x} - b_1e^{\gamma_1 x} + \eta_2 a_2 e^{-\gamma_2 x} - \eta_2 b_2 e^{\gamma_2 x} + (\psi_1\gamma_1 + \psi_2\eta_2\gamma_2)e^{-\gamma x}, \tag{31}$$

$$I_2 = \eta_1 a_1 e^{-\gamma_1 x} - \eta_1 b_1 e^{\gamma_1 x} + a_2 e^{-\gamma_2 x} - b_2 e^{\gamma_2 x} + (\psi_1\eta_1\gamma_1 + \psi_2\gamma_2)e^{-\gamma x}, \tag{32}$$

$$V_1 = K_1 a_1 e^{-\gamma_1 x} + K_1 b_1 e^{\gamma_1 x} - \eta_1 K_2 a_2 e^{-\gamma_2 x} - \eta_1 K_2 b_2 e^{\gamma_2 x} + (\psi_1 K_1\gamma - \psi_2 K_2\eta_1\gamma)e^{-\gamma x}, \tag{33}$$

$$V_2 = -\eta_2 K_1 a_1 e^{-\gamma_1 x} - \eta_2 K_1 b_1 e^{\gamma_1 x} + K_2 a_2 e^{-\gamma_2 x} + K_2 b_2 e^{\gamma_2 x} - (\psi_1 K_1\eta_2\gamma - \psi_2 K_2\gamma)e^{-\gamma x}. \tag{34}$$

Before proceeding with the application of these results to specific crosstalk problems, we will establish certain relations which, as in the single-tertiary analysis, will be fundamental in relating crosstalk measurements on short lengths of cable to the crosstalk to be expected in a longer length.

Let us consider the crosstalk as measured on a short length under the following two conditions: (1) both tertiaries open and (2) tertiary 1 short-circuited at each end and tertiary 2 open. We will designate the crosstalk under condition (1) by Xl and under condition (2) by $Xl(1 - \xi)$. Under condition (2) the tertiary current (I_1) for unit current in the energized coaxial is given by $\frac{Z_{13}}{Z_{11}}$ and the indirect crosstalk current in the disturbed coaxial is thus $\frac{-Z^2_{13}l}{2ZZ_{11}}$, so that we have

$$X\xi = \frac{Z^2_{13}}{2ZZ_{11}}. \tag{35}$$

Interaction Crosstalk with One Tertiary Short-Circuited Far-End

For the sake of simplicity, and with no considerable loss of applicability, we will postulate the restriction that $e^{\gamma_1 l}$ and $e^{\gamma_2 l}$ are large compared with $e^{\gamma l}$, where l is the length of the section in which we are formulating the tertiary currents.

Referring to equations (31) to (34), under the above restrictions the terms involving $e^{-\gamma_1 x}$ and $e^{-\gamma_2 x}$ are negligible in the region near $x = l$ and thus in this region

$$I_1 = - b_1 e^{\gamma_1 x} - \eta_2 b_2 e^{\gamma_2 x} + [(p_1 - q_1) + \eta_2(p_2 - q_2)]e^{-\gamma x}, \tag{36}$$

$$I_2 = - \eta_1 b_1 e^{\gamma_1 x} - b_2 e^{\gamma_2 x} + [\eta_1(p_1 - q_1) + (p_2 - q_2)]e^{-\gamma x}, \tag{37}$$

$$V_1 = K_1 b_1 e^{\gamma_1 x} - \eta_1 K_2 b_2 e^{\gamma_2 x} + [K_1(p_1 + q_1) - \eta_1 K_2(p_2 + q_2)]e^{-\gamma x}, \tag{38}$$

$$V_2 = - \eta_2 K_1 b_1 e^{\gamma_1 x} + K_2 b_2 e^{\gamma_2 x} + [- \eta_2 K_1(p_1 + q_1) + K_2(p_2 + q_2)]e^{-\gamma x}, \tag{39}$$

where [2]

$$p_1 = \frac{Z_{13}}{2K_1(1 - \eta_1\eta_2)(\gamma_1 - \gamma)}, \tag{40}$$

$$q_1 = \frac{- Z_{13}}{2K_1(1 - \eta_1\eta_2)(\gamma_1 + \gamma)}, \tag{41}$$

$$p_2 = \frac{Z_{13}\eta_2}{2K_2(1 - \eta_1\eta_2)(\gamma_2 - \gamma)}, \tag{42}$$

$$q_2 = \frac{- Z_{13}\eta_2}{2K_2(1 - \eta_1\eta_2)(\gamma_2 + \gamma)}. \tag{43}$$

If, now, x is measured from the far end, and the following substitutions are made as a matter of convenience: [3]

$$a_1 = b_1 e^{(\gamma_1 + \gamma)l}, \tag{44}$$

$$a_2 = b_2 e^{(\gamma_2 + \gamma)l}, \tag{45}$$

equations (36) to (39), multiplied by $e^{\gamma l}$ so that the currents and voltages are given for unit received current in the energized coaxial, become

$$I_1 = - a_1 e^{-\gamma_1 x} - \eta_2 a_2 e^{-\gamma_2 x} + [(p_1 - q_1) + \eta_2(p_2 - q_2)]e^{\gamma x}, \tag{46}$$

$$I_2 = - \eta_1 a_1 e^{-\gamma_1 x} - a_2 e^{-\gamma_2 x} + [\eta_1(p_1 - q_1) + (p_2 - q_2)]e^{\gamma x}, \tag{47}$$

[2] The terms involving $e^{-\gamma x}$ here are identical with the corresponding terms in equations (31) to (34) except for the change in nomenclature, which in each case has been chosen so that the a's and b's will be given by simple functions of the parameters employed (p's and q's here: ψ's in the previous equations).

[3] a_1 and a_2 here have no relation to a_1 and a_2 in equations (31) to (34).

$$V_1 = K_1 a_1 e^{-\gamma_1 x} - \eta_1 K_2 a_2 e^{-\gamma_2 x} + [K_1(p_1 + q_1) - \eta_1 K_2(p_2 + q_2)]e^{\gamma x}, \quad (48)$$

$$V_2 = -\eta_2 K_1 a_1 e^{-\gamma_1 x} + K_2 a_2 e^{-\gamma_2 x} + [-\eta_2 K_1(p_1 + q_1) + K_2(p_2 + q_2)]e^{\gamma x}. \quad (49)$$

In the section, of length l', adjacent to the far end of the energized section, the impressed fields are zero and thus (under the condition that $e^{\gamma_1 l'}$ and $e^{\gamma_2 l'}$ are large compared with unity), using primes to indicate currents and voltages in this region, with the distance x' taken positive from $x = l$,

$$I_1' = a_1' e^{-\gamma_1 x'} + \eta_2 a_2' e^{-\gamma_2 x'}, \quad (50)$$

$$I_2' = \eta_1 a_1' e^{-\gamma_1 x'} + a_2' e^{-\gamma_2 x'}, \quad (51)$$

$$V_1' = K_1 a_1' e^{-\gamma_1 x'} - \eta_1 K_2 a_2' e^{-\gamma_2 x'}, \quad (52)$$

$$V_2' = -\eta_2 K_1 a_1' e^{-\gamma_1 x'} + K_2 a_2' e^{-\gamma_2 x'}. \quad (53)$$

With tertiary 1 short-circuited, the boundary conditions to be satisfied are that at $x = x' = 0$, $V_1 = V_1' = 0$ and $I_2 = I_2'$. From these boundary conditions, we obtain

$$a_1 = -(p_1 + q_1) + \frac{\eta_1 K_2(\eta_1 p_1 + p_2)}{K_1 + \eta_1^2 K_2}, \quad (54)$$

$$a_2 = -(p_2 + q_2) + \frac{K_1(\eta_1 p_1 + p_2)}{K_1 + \eta_1^2 K_2}, \quad (55)$$

$$a_1' = \frac{\eta_1 K_2(\eta_1 p_1 + p_2)}{K_1 + \eta_1^2 K_2}, \quad (56)$$

$$a_2' = \frac{K_1(\eta_1 p_1 + p_2)}{K_1 + \eta_1^2 K_2}. \quad (57)$$

The equal-level far-end far-end interaction crosstalk FF_s is given by [4]

$$FF_s = \frac{Z_{13}}{2Z} e^{\gamma l'} \int_0^{l'} I_1' e^{-\gamma(l'-x')} dx'. \quad (58)$$

With I_1' as given by equations (50), (56) and (57), under the restrictions we have placed on $\gamma_1 l'$ and $\gamma_2 l'$, we have

$$FF_s = \frac{Z_{13}^2}{4Z(1 - \eta_1\eta_2)(K_1 + \eta_1^2 K_2)} \times \left[\frac{\eta_1}{K_1(\gamma_1 - \gamma)} + \frac{\eta_2}{K_2(\gamma_2 - \gamma)} \right] \left[\frac{\eta_1 K_2}{\gamma_1 - \gamma} + \frac{\eta_2 K_1}{\gamma_2 - \gamma} \right] \quad (59)$$

[4] As pointed out in the Schelkunoff-Odarenko paper in the section on mutual impedance, since a coaxial circuit is involved, the current distribution external to this circuit does not affect the mutual impedance, and hence the current I_2' contributes nothing to the crosstalk.

or, with the use of equation (35),

$$FF_s = \frac{X\xi Z_{11}}{2(1-\eta_1\eta_2)(K_1+\eta_1^2K_2)} \times \left[\frac{\eta_1}{K_1(\gamma_1-\gamma)}+\frac{\eta_2}{K_2(\gamma_2-\gamma)}\right]\left[\frac{\eta_1K_2}{\gamma_1-\gamma}+\frac{\eta_2K_1}{\gamma_2-\gamma}\right]. \quad (60)$$

The equal-level far-end near-end interaction crosstalk FN_s is given by

$$FN_s = \frac{Z_{13}}{2Z}\int_0^{l'} I_1' e^{-\gamma x'}dx', \quad (61)$$

and under the restrictions we have placed on $\gamma_1 l'$ and $\gamma_2 l'$, we have

$$FN_s = \frac{Z^2{}_{13}}{4Z(1-\eta_1\eta_2)(K_1+\eta_1^2K_2)} \times \left[\frac{\eta_1}{K_1(\gamma_1-\gamma)}+\frac{\eta_2}{K_2(\gamma_2-\gamma)}\right]\left[\frac{\eta_1K_2}{\gamma_1+\gamma}+\frac{\eta_2K_1}{\gamma_2+\gamma}\right] \quad (62a)$$

$$= \frac{X\xi Z_{11}}{2(1-\eta_1\eta_2)(K_1+\eta_1^2K_2)} \times \left[\frac{\eta_1}{K_1(\gamma_1-\gamma)}+\frac{\eta_2}{K_2(\gamma_2-\gamma)}\right]\left[\frac{\eta_1K_2}{\gamma_1+\gamma}+\frac{\eta_2K_1}{\gamma_2+\gamma}\right]. \quad (62b)$$

Near-End

Under the above restriction that $e^{\gamma_1 l}$, $e^{\gamma_1 l'}$, $e^{\gamma_2 l}$ and $e^{\gamma_2 l'}$ are large compared with $e^{\gamma l}$, the currents and voltages in the disturbing section near $x = 0$ are given by equations (31) to (34) with the b-terms omitted, and the currents and voltages in the disturbed section adjacent to the sending end by equations (50) to (53).

The boundary conditions to be satisfied are that at $x = x' = 0$, $V_1 = V_1' = 0$ and $I_2 = -I_2'$. From these boundary conditions we obtain

$$a_1 = -(p_1+q_1) + \frac{\eta_1K_2(q_2+\eta_1q_1)}{K_1+\eta_1^2K_2}, \quad (63)$$

$$a_2 = -(p_2+q_2) + \frac{K_1(q_2+\eta_1q_1)}{K_1+\eta_1^2K_2}, \quad (64)$$

$$a_1' = \frac{\eta_1K_2(q_2+\eta_1q_1)}{K_1+\eta_1^2K_2}, \quad (65)$$

$$a_2' = \frac{K_1(q_2+\eta_1q_1)}{K_1+\eta_1^2K_2}. \quad (66)$$

The near-end near-end interaction crosstalk NN_s is given by

$$NN_s = \frac{Z_{13}}{2Z}\int_0^{l'} I_1' e^{-\gamma x} dx \tag{67a}$$

$$= -\frac{Z^2_{13}}{4Z(1-\eta_1\eta_2)(K_1+\eta_1^2K_2)} \times \left[\frac{\eta_1}{K_1(\gamma_1+\gamma)} + \frac{\eta_2}{K_2(\gamma_2+\gamma)}\right]\left[\frac{\eta_1K_2}{\gamma_1+\gamma} + \frac{\eta_2K_1}{\gamma_2+\gamma}\right] \tag{67b}$$

$$= -\frac{X\xi Z_{11}}{2(1-\eta_1\eta_2)(K_1+\eta_1^2K_2)} \times \left[\frac{\eta_1}{K_1(\gamma_1+\gamma)} + \frac{\eta_2}{K_2(\gamma_2+\gamma)}\right]\left[\frac{\eta_1K_2}{\gamma_1+\gamma} + \frac{\eta_2K_1}{\gamma_2+\gamma}\right]. \tag{67c}$$

Near-End Crosstalk with One Tertiary Short-Circuited

Although the derivation of the formula for near-end crosstalk N_s with one tertiary short-circuited is too long to be included here, it seems advisable to give this formula without derivation. Under the above mentioned restriction that $e^{\gamma_1 l}$ and $e^{\gamma_2 l}$ are large compared with $e^{\gamma l}$,

$$N_s = N_t + NN - NN_s, \tag{68}$$

where

N_t = near-end crosstalk, tertiaries terminated,

NN = near-end near-end interaction crosstalk between two adjoining lengths, tertiaries with no discontinuity,

NN_s = near-end near-end interaction crosstalk between two adjoining lengths with tertiary 1 short-circuited at the junction.

The first two terms (N_t and NN) may, with the types of cable studied so far, be determined with satisfactory accuracy from the single-tertiary analysis. In such a case, the formulas given herein are sufficient for computing the near-end crosstalk with one tertiary short-circuited.

III—Comparison of Computed Crosstalk with Measured Values

With 72-ft. and 145-ft. samples of the twin coaxial cable described in the companion paper by Messrs. Booth and Odarenko, crosstalk and impedance measurements were made in the laboratory, at frequencies from 50 kc to 300 kc, the sheath and quads in parallel being considered as providing a single tertiary, that is, as being connected together at short intervals.

The far-end crosstalk for a length of 5 miles was computed from these laboratory measurements and in Fig. 1 the results are compared with measurements on this length, the crosstalk in either case being practically the same whether the tertiary was terminated or short-circuited.

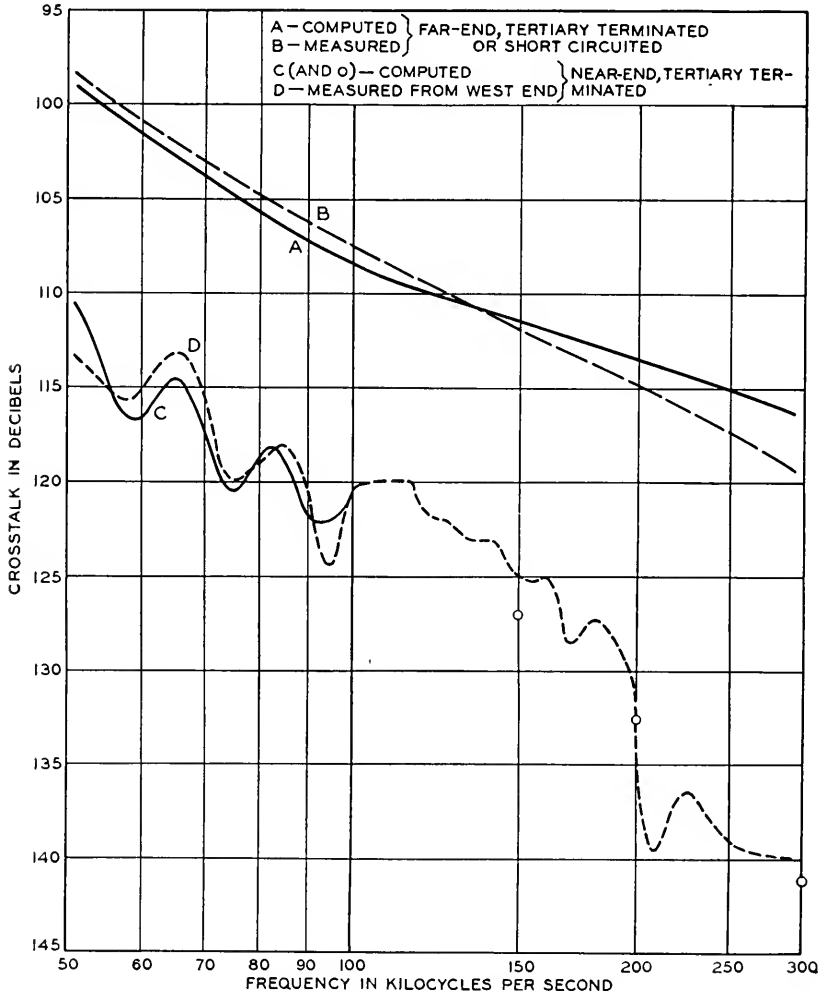


Fig. 1—Far-end crosstalk and near-end crosstalk for 5-mile length.

In Figs. 1 and 2, the computed near-end crosstalk for a length of 5 miles is compared with representative measurements on the above mentioned twin coaxial cable. Figure 1 shows this comparison with the tertiary terminated and Fig. 2 with the tertiary short-circuited.

The assumption of uniformity of the coaxial lines, as regards transfer impedances, and of the tertiary circuits as regards transmission characteristics, is a more serious restriction in the computation of near-end crosstalk than in the case of far-end crosstalk. Even for long lengths of cable, the near-end crosstalk is determined almost entirely by the crosstalk behavior of a relatively short length of the cable near the sending end, whereas the average crosstalk characteristics determine

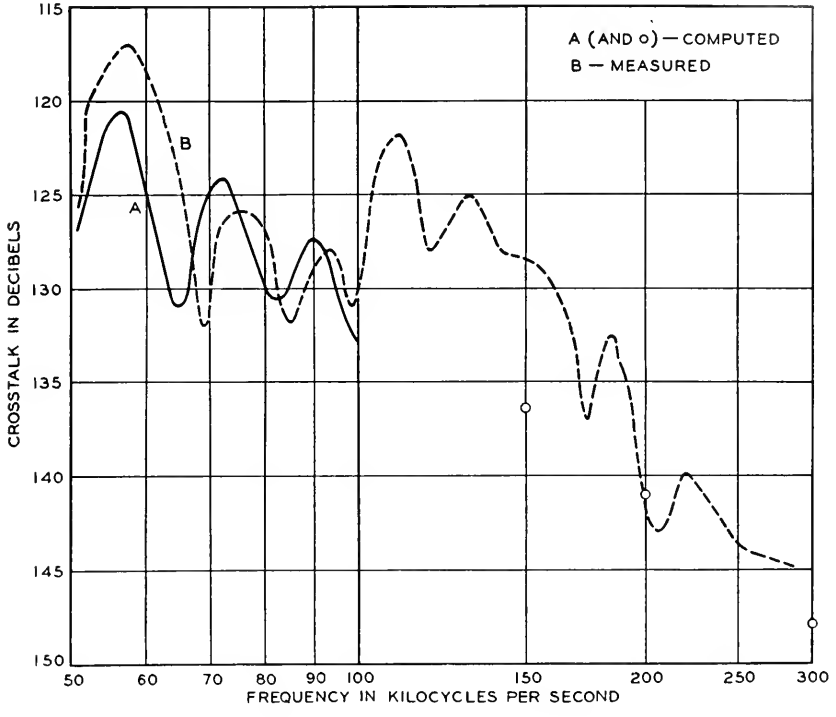


Fig. 2—Near-end crosstalk for 5-mile length, tertiary short-circuited.

the far-end crosstalk for a long length of cable. Thus, from measurements on representative short lengths, the far-end crosstalk for a long length may generally be computed more accurately than can the near-end crosstalk.

Similarly, the various types of interaction crosstalk depend largely upon the crosstalk behavior of relatively short lengths of the cable near the junction. In Fig. 3, the far-end far-end interaction crosstalk has been chosen as an illustration of the correlation which has been obtained between computed interaction crosstalk for the above men-

tioned twin coaxial cable and the measured interaction crosstalk The curves in Fig. 3 are for equal lengths either of 3000 ft. or 12,000 ft. In the case of the measured values, the junction point of the two sections was not the same for these two lengths. Although the agreement between calculated and measured values is only fair, the spread in the experimental results for these two cases, which for uniform cable would be slight, is about the same as the spread between calculated and measured values.

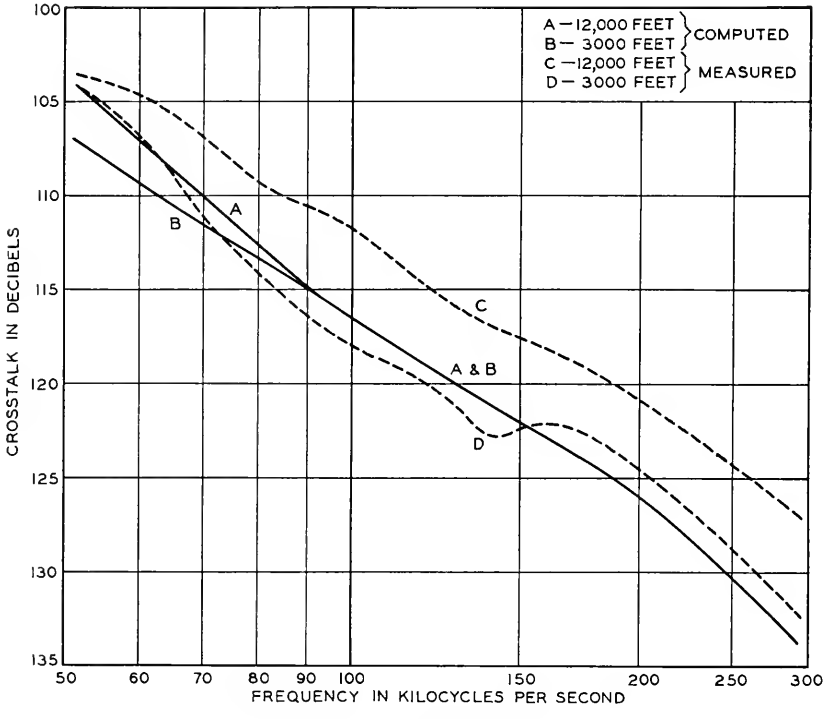


Fig. 3—Far-end far-end interaction crosstalk between two equal lengths.

The two-tertiary formulas have so far been applied only to one type of cable with four coaxial lines and a layer of paper-insulated pairs. The longest length of this type of cable on which crosstalk measurements have been made is 1900 ft. The various types of interaction crosstalk with one tertiary short-circuited, as computed from the formulas given above, agree roughly with the measured interaction crosstalk under this same condition. However, the restriction that the tertiary circuits involved are electrically long, as postulated in deriving the interaction crosstalk formulas for this case, is not satisfied,

and comparisons of the calculated and measured values are not very significant.

It may be remarked that the application of the single-tertiary analysis to all cases in which the two tertiaries were treated alike (either terminated or short-circuited) gave very satisfactory agreement between computed and measured crosstalk for this 1900 ft. length of 4-coaxial cable.

Crosstalk Between Coaxial Conductors in Cable

By R. P. BOOTH and T. M. ODARENKO

The available literature on crosstalk between coaxial conductors in contact makes it clear that the presence of any other conducting material in continuous or frequent contact with the coaxial outer conductors simply reduces the coupling per unit length without altering the law of crosstalk summation with length.

When the conducting material is insulated from the coaxials, as in the case of quads and sheath in coaxial cables, the situation is more complicated. Instead of simply reducing the coupling per unit length the quads and sheath, with the outer conductors for a return, provide a tertiary circuit in which interaction crosstalk can take place between elementary line sections. The summation with length for this type of crosstalk is quite different from that between two coaxials in contact and therefore the combined summation is obviously more involved.

Tests on sections of a five-mile length of coaxial cable were made at Princeton, New Jersey, in the latter part of 1937 and early in 1938 in order to obtain experimental verification of the manner in which the quads and sheath affect crosstalk summation with length. It is shown that the crosstalk component due to the presence of the sheath and quads opposes the component which is present between two coaxials in free space so that the resultant crosstalk is considerably lower than would be computed ignoring the tertiary effects.

Introduction

In spite of the geometrical and electrical symmetry of the coaxial circuit and the excellent shielding properties of the outer conductor, a part of the electromagnetic energy escapes from the circuit through the outer conductor and sets up an electromagnetic field in the space around it. Any circuit, be it even another coaxial placed in this field will absorb a part of the energy stored in the field and deliver it to the terminals of the circuit in the form of an unwanted or interfering current—the crosstalk current. The magnitude of this crosstalk current depends on a variety of factors, such as the physical characteristics of the conductors and of the intervening space, the frequency and the length of the circuit.

Expressions for two important cases of crosstalk between two coaxial circuits in *free space*, namely, the so-called "direct" crosstalk with the outer conductors in continuous contact and the "indirect" crosstalk with the outer conductors insulated from each other, were determined

and discussed in a previously published paper.[1] It was shown there that the direct far-end crosstalk is directly proportional to l and the direct near-end crosstalk is proportional to

$$\frac{1 - \epsilon^{-2\gamma l}}{2\gamma},$$

where l is the length and γ is the propagation constant of either coaxial unit. The indirect crosstalk was shown to be a more complicated function of the length.

The present paper extends this earlier work to include the case where the coaxials are enclosed in a common sheath or, in the general case, paralleled by any conducting material symmetrically disposed.[2] When this conducting material is introduced in the neighborhood of two coaxials in contact the conditions for crosstalk production are naturally changed from those existing in free space. If the material is uniformly distributed along the coaxials and is in continuous or frequent contact with the outer conductors the summation of crosstalk with length is the same as before but the magnitude is reduced. This reduction is due to the fact that part of the current formerly flowing on the disturbed outer conductor now flows on the new conducting material instead, thus reducing the direct crosstalk coupling per unit length.

In most cables, the coaxial outer conductors are in contact but the other conducting material (sheath and quads) is insulated from the outer conductors. The quads must obviously be insulated for normal use and the sheath is kept insulated except at the ends of a repeater section in order to permit the use of insulating joints for electrolysis prevention where required. This material thus provides an extra transmission circuit, or tertiary circuit, in which tertiary currents can be propagated up and down the line. In such a case the resulting crosstalk in any length consists of both the direct crosstalk between the contacting coaxials and the indirect crosstalk via the outer conductor-sheath and quad tertiary circuit. The general formulas given in the Schelkunoff-Odarenko paper apply for these components. Since the two components follow different laws regarding summation with length the resultant summation is quite complicated except for very short or very long lengths.

The study of the tertiary effects on crosstalk summation is the main contribution of this paper to crosstalk theory. Emphasis will be placed on the development of a simple physical picture which will help one to

[1] Schelkunoff-Odarenko paper in *Bell Sys. Tech. Jour.*, April, 1937.

[2] In the interim between our tests and this publication a paper by H. Kaden concerning this general subject was published in the Europaischer Fernsprechdienst, no. 50, October, 1938, pp. 366–373.

visualize clearly the influence of the tertiary circuits in the summation process. To produce such a picture a certain amount of review of the general crosstalk problem will be necessary. This is undertaken in Part I of this paper.

Part II is devoted mainly to the presentation of test data taken in November and December, 1937, January and February, 1938 on sections of a five-mile length of a twin coaxial cable near Princeton. These data confirm and graphically illustrate certain relationships developed in Part I. In addition they provide information on the tendency of tertiary circuits to complicate the effectiveness of transpositions and show how interaction crosstalk takes place around repeaters via the tertiary circuits.

PART I—THEORY

In any series of crosstalk tests on short lengths of paired or quadded cable where the problem of combining a number of such lengths is concerned it has generally been the practice to terminate both the test circuits and important tertiary circuits in characteristic impedance. Under such a condition the normal influence of all circuits in the production of crosstalk within each short section is provided for and the summation process, including interaction between successive sections, can be studied under actual line conditions. This is a general method applicable to any type of coupling and was adopted for the Princeton investigation. The effect of discontinuities such as short-circuited tertiaries at the extreme ends of a repeater section can be readily handled mathematically as correction terms due to "end effect."

To simplify the presentation of the factors involved, the discussion in this section will be confined mainly to the case of far-end crosstalk. In a twin coaxial cable where the transmission in the two units is in opposite directions there actually exists no far-end crosstalk problem since only talker echo, a near-end crosstalk phenomenon, is involved.[3] In multi-unit cable, however, there will be far-end crosstalk between different systems. Since this type of crosstalk tends to increase directly with the number of repeater sections it is important to understand its nature thoroughly. Moreover, in a study of fundamentals it is possible to avoid certain complications not essential to an understanding of the problem by investigating far-end rather than near-end crosstalk.

To present a clear picture of the physical meaning of some of the forthcoming mathematical expressions their derivations will be ap-

[3] This statement may not hold if the repeater impedances fail to match the line impedance since in that case the far-end crosstalk can be reflected and appear as near-end crosstalk.

proached in as elementary a fashion as possible. In order to do this we shall start with the simple arrangement of two coaxial conductors in free space, a case already covered in previous papers. To the crosstalk equations covering this case will then be added terms to allow for the effects of quads and sheath. In all that follows in Part I the quads and sheath will be considered as one unit referred to as the "sheath." This is a good approximation as will be shown in Part II.

The conception of two independent crosstalk components—a direct or transverse component between coaxials in contact and an indirect or interaction component via the sheath tertiary circuit—is not necessary for the solution of the problem. It is preserved here, however, as offering a familiar and much simpler approach to a clear understanding of the processes involved in crosstalk summation with length.

Far-End Crosstalk

Consider first an elementary section, dl, of a long single coaxial in free space as indicated in Sketch (a) of Fig. 1. If the current at this point in the center conductor is I_1 the current in the outer conductor is practically $-I_1$ since there is no other return path (except through the air dielectric which offers a high impedance especially at the lower broad-band frequencies considered here). Using Schelkunoff's nomenclature we may state that an open-circuit voltage equal to $e_1 = I_1 Z_{\alpha\beta} dl$ is developed on the outer surface of the outer coaxial conductor. The term $Z_{\alpha\beta}$ represents the surface transfer impedance (mutual impedance) per unit length between the inner and outer surfaces of the outer coaxial conductor.

Now suppose that we place another long coaxial parallel to the first one and, for generality, insulated from it as shown by Sketch (b) of Fig. 1. The open-circuit voltage e_1 on length dl of the first coaxial outer conductor will now cause current to flow in the intermediate circuit composed of the two outer conductors. The parameters of this circuit are γ_3 and Z_3 as shown on the sketch. In returning on the second coaxial outer conductor this current causes crosstalk into the second coaxial circuit.

It is convenient at this point to replace the original impressed voltage e_1 by the set of emf's shown in Sketch (c) of Fig. 1. The insertion of equal and opposite voltages $e_1/2$ on the outer surface of the disturbed coaxial outer conductor does not change conditions but enables us to consider certain effects separately. The first effect to be considered is that due to the pair of equal and opposite voltages $e_1/2$ in the loop composed of the two coaxial outer conductors. These voltages combine to form a "balanced" voltage e_1 which tends to drive current

around the balanced circuit composed of the two outer conductors. For the present we shall not consider the voltages $e_1/2$ which are in the same direction in the outer conductors.

The current in the "balanced" intermediate circuit of characteristic impedance Z_3 and propagation constant γ_3 due to the balanced voltage e_1 in the elementary length dl is $i_3 = e_1/2Z_3$. This current flowing along the outer coaxial conductor of the disturbed circuit produces a voltage $e_2 = i_3 Z_{\alpha\beta} dl$ on the inner surface of this outer conductor and this voltage in turn causes a current i_{2a} in the disturbed coaxial circuit

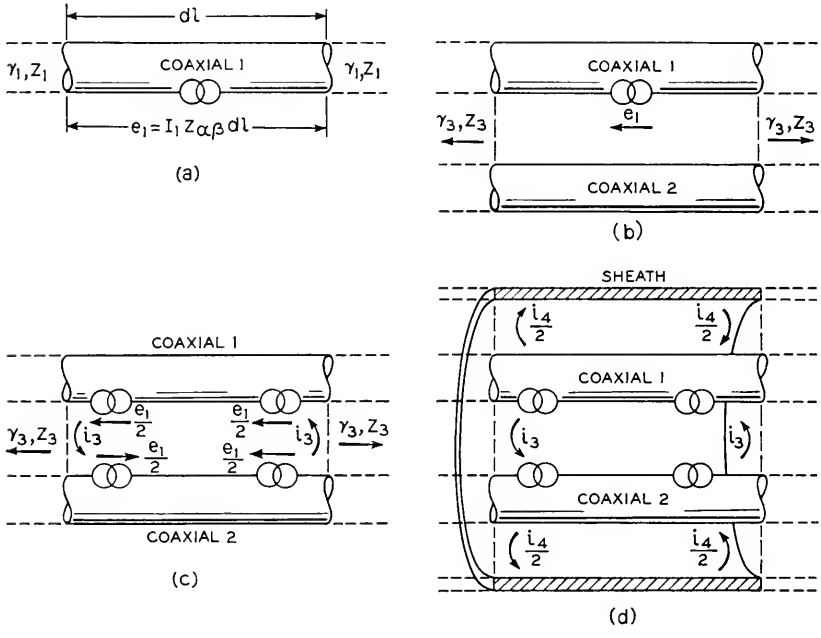


Fig. 1.—Coaxial crosstalk schematics.

equal to $e_2/2Z$, where Z is the coaxial characteristic impedance.[4] In a long line other elementary lengths of the disturbed coaxial are also affected by i_3 because of its propagation along the intermediate circuit. (This crosstalk by way of a tertiary circuit from one length into another is known as indirect or "interaction crosstalk" and because of its presence the summation of crosstalk with length is not a simple function of length even for systematic coupling such as occurs with coaxials.) This is a crosstalk case for which the general solution is already

[4] The subscript "a" in i_{2a} relates this current to the so-called "mode a" current used by Carson and Hoyt in their paper entitled "Propagation of Periodic Currents Over a System of Parallel Wires," *Bell Sys. Tech. Jour.*, July, 1927.

available. When the effects are integrated it is found that the far-end crosstalk is quite a complicated function of length and of the tertiary and coaxial propagation and impedance characteristics.[5] However, if the coaxial units are in actual contact as in the case of the coaxial cable to be considered here, the formula for the far-end crosstalk F_3 expressed as a current ratio is quite simple, namely,

$$F_3 = \frac{Z_{\alpha\beta}^2}{2ZZ_{33}} \cdot l, \tag{1}$$

where $Z_{33} = Z_3\gamma_3$ is the series impedance per unit length of the circuit composed of one coaxial outer conductor with return on the other. Thus, for this component, the far-end crosstalk is directly proportional to length. This simple relation results from the fact that the intermediate circuit, being continuously shorted, has such high attenuation that no interaction crosstalk between elementary lengths can exist.

We shall now consider the crosstalk contribution due to the longitudinal voltage $e_1/2$ acting along both coaxial outer conductors in parallel. Suppose that a sheath is placed symmetrically around the two coaxials but insulated from them as shown in Sketch (d) of Fig. 1. The longitudinal voltage sends a current around the circuit composed of the two parallel outer conductors with sheath return equal to $i_4 = e_1/(2)(2Z_4)$, where Z_4 is the characteristic impedance of this circuit. Half of this longitudinal current flows on the disturbed coaxial outer conductor *in opposition* to the balanced current i_3 flowing there.

Following previous procedure it can be shown that in the elementary length a crosstalk current $i_{2c} = i_4 Z_{\alpha\beta} dl/4Z$ will flow in the disturbed coaxial circuit.[6] Other elementary lengths are also affected by i_4 thus producing interaction crosstalk. When the effects are integrated over a length l the far-end crosstalk for this component is found to be as follows:

$$F_4 = -\frac{Z_{\alpha\beta}^2}{16ZZ_4}\left[\frac{2l}{\gamma_4}\cdot\frac{\gamma_4^2}{\gamma_4^2-\gamma^2} - \left(\frac{2(\gamma_4^2+\gamma^2)}{(\gamma_4^2-\gamma^2)^2} - \frac{\epsilon^{-(\gamma_4-\gamma)l}}{(\gamma_4-\gamma)^2} - \frac{\epsilon^{-(\gamma_4+\gamma)l}}{(\gamma_4+\gamma)^2}\right)\right], \tag{2}$$

where γ_4 is the propagation constant of the sheath-outer conductor circuit. If the sheath is in actual contact with the coaxial units the

[5] See equation (40) in the Schelkunoff-Odarenko paper in *Bell Sys. Tech. Jour.*, April, 1937.

[6] The subscript "c" in i_{2c} relates this current to the "mode c" current used by Carson and Hoyt in their paper of July, 1927.

formula reduces to the simple relation

$$F_4 = -\frac{Z_{\alpha\beta}^2}{16ZZ_4}\cdot\frac{2l}{\gamma_4} = -\frac{Z_{\alpha\beta}^2}{8ZZ_{44}}\cdot l, \tag{3}$$

where $Z_{44} = Z_4\gamma_4$ = series impedance per unit length of the circuit composed of the outer conductors with sheath return.[7] For a sheath in contact this component is thus directly proportional to length since interaction crosstalk from one elementary length into another has been eliminated.

Now, while we are actually concerned with the insulated sheath as covered by (2) it is of considerable interest to study equations (1) and (3) at this point. The total far-end crosstalk when the outer conductors and sheath are in contact is the sum of the crosstalk components F_3 and F_4 as given in equations (1) and (3), or

$$F_3 + F_4 = F_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\frac{l}{Z_{33}} - \frac{l}{4Z_{44}}\right]. \tag{4}$$

This simple addition follows from the fact that the circuits for the two modes of propagation covered by equations (1) and (3) are mutually non-inductive because of symmetry so that there is no reaction between them. The recognition of this fact does away with the necessity of complicated mathematics which would otherwise have to be used in the general solution.[8]

In formula (4) the second term in the bracket represents the contribution of the tertiary circuit involving the sheath and is seen to be opposite in sign to the first term which represents the crosstalk which would exist in the absence of the sheath. The equation illustrates mathematically the previous statement that conducting material in contact with the coaxials acts to reduce the crosstalk. Since both components are directly proportional to length, the total is also directly proportional to length.

It is apparent in formula (4) that the crosstalk would be zero if the values of Z_{33} and $4Z_{44}$ were equal. In cables where steel tapes are used on the outer surface of the coaxials this condition is approached. For example, if we neglect external inductance and proximity effects, Z_{33} would be equal to twice the surface self-impedance of a single outer

[7] It should be noted here that it is not really necessary to postulate a separate sheath return in order to obtain expression (3) for F_4 due to the longitudinal voltage $e_1/2$, since the return *in continuous contact* with the outer conductors will actually tend to lose its identity. The device of introducing sheath return insulated from the outer conductors and then shorting it to the conductors serves only to simplify the concepts of Z_{44}, Z_4, γ_4.

[8] This principle of symmetry can be extended to the case of four coaxial units whether insulated or in contact.

conductor while Z_{44} would equal one-half of the surface self-impedance of a single outer conductor (neglecting the self-impedance of the lead sheath in comparison with the iron outer conductors). Thus, neglecting differences in external inductance and in proximity effects $1/Z_{33}$ would equal $1/4Z_{44}$ and the crosstalk would vanish. Actually, the observed reduction in crosstalk due to these opposing terms is about 32 db at 50 kilocycles in a 145-foot section of twin coaxial with quads and sheath shorted to the coaxials at the ends only. Physically this means that the current due to the voltage on the outer conductor surface of the disturbing coaxial flows mainly in the sheath and quads rather than on the high impedance surface of the disturbed coaxial.

Now let us consider the case where the sheath is insulated from the coaxial outer conductors. For this case equations (1) and (2) may also be added directly to give

$$F_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\left(\frac{l}{Z_{33}} - \frac{l}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2-\gamma^2}\right) + \frac{\gamma_4}{4Z_{44}}\left(\frac{\gamma_4^2+\gamma^2}{(\gamma_4^2-\gamma^2)^2}\right) - \frac{\gamma_4}{4Z_{44}}\left(\frac{\epsilon^{-(\gamma_4-\gamma)l}}{2(\gamma_4-\gamma)^2} + \frac{\epsilon^{-(\gamma_4+\gamma)l}}{2(\gamma_4+\gamma)^2}\right)\right]. \quad (5)$$

This equation appears quite formidable but it has been split purposely into three terms which will be examined individually. The first term is directly proportional to length, the second term is independent of length and the third term involves length exponentially. For lengths where the tertiary circuit is electrically long the third term vanishes and we have

$$F_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\left(\frac{l}{Z_{33}} - \frac{l}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2-\gamma^2}\right) + \frac{\gamma_4}{4Z_{44}}\left(\frac{\gamma_4^2+\gamma^2}{(\gamma_4^2-\gamma^2)^2}\right)\right]. \quad (6)$$

In electrically short lengths we get

$$F_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\left(\frac{l}{Z_{33}} - \frac{l}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2-\gamma^2}\right) + \frac{\gamma_4}{4Z_{44}}\left(\frac{\gamma_4 l}{\gamma_4^2-\gamma^2} - \frac{l^2}{2}\right)\right] = \frac{Z_{\alpha\beta}^2}{2Z}\left[\frac{l}{Z_{33}} - \frac{\gamma_4}{4Z_{44}}\cdot\frac{l^2}{2}\right] \backsimeq \frac{Z_{\alpha\beta}^2}{2Z}\left[\frac{l}{Z_{33}}\right], \quad (7)$$

in which it is seen that terms two and three of (5) combine to cancel the second half of term one.

From equations (5), (6) and (7) we are now ready to build a physical picture of what takes place as l is increased for cable sections where the sheath is insulated from the coaxial outer conductors but terminated to them at each end in characteristic impedance, Z_4. Starting with equation (7) we see that for very short lengths the term involving

l^2 becomes negligible, that is, the crosstalk is practically all due to the component which exists in the complete absence of a sheath (see equation (1)). In the range of lengths where this is true *the crosstalk increases directly with length.*

Quite a different state of affairs exists for a section electrically long enough for equation (6) to hold. The first bracketed term is still proportional to length but now consists of the difference of two components. The first of these represents the crosstalk between the coaxials *with no sheath present* while the second is a part of the crosstalk component introduced by the presence of the sheath. Except for the factor $\gamma_4^2/\gamma_4^2 - \gamma^2$ this first bracketed term in equation (6) is the same as equation (4) for a sheath in contact where, as we have already noted, the cancellation of the two components is quite effective when steel tapes are used on the outer conductors. Since γ_4^2 is necessarily considerably greater than γ^2 because of these steel outer conductors, it is reasonable to expect that the factor $\gamma_4^2/\gamma_4^2 - \gamma^2$ is nearly unity and that, therefore, the two components in the first bracketed term of equation (6) will also tend to cancel leaving a residual proportional to length but *much* lower in magnitude than either component *alone.*

The second bracketed term of equation (6) is entirely independent of length. This term has also been introduced by the presence of the tertiary circuit and its magnitude depends on the characteristics of this circuit.

Thus, even without knowing the relative magnitudes of the two components of the first bracketed term of equation (6) for a given length, it is apparent that as l is increased this term must eventually be controlling. The crosstalk will then again be proportional to length as it was for very short lengths but at a reduced level proportional to

$$\frac{\dfrac{1}{Z_{33}} - \dfrac{1}{4Z_{44}} \cdot \dfrac{\gamma_4^2}{\gamma_4^2 - \gamma^2}}{\dfrac{1}{Z_{33}}} = 1 - \frac{Z_{33}}{4Z_{44}} \cdot \frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}.$$

It is quite evident, too, that for a range of lengths where the tertiary circuits are electrically long but where the first term of equation (6) has not had a chance to build up sufficiently the crosstalk will be about constant at a level determined mainly by the second term.

The above analysis may well suffice as a background for an interpretation of the measurements to be given in Part II. However, another and perhaps in some ways a more illuminating approach from a physical standpoint is possible.

Suppose, for example, that far-end crosstalk measurements are made on two cable sections each of length l with tertiaries terminated as illustrated in Sketch (a) of Fig. 2. Let the total crosstalk in each section be equal to F_l as defined by equation (5) above. If these two sections are joined together the total crosstalk is $2F_l$ plus some other terms which represent the interaction crosstalk between the two sections as illustrated in Sketch (b) of Fig. 2. We shall call the component F_{nn}

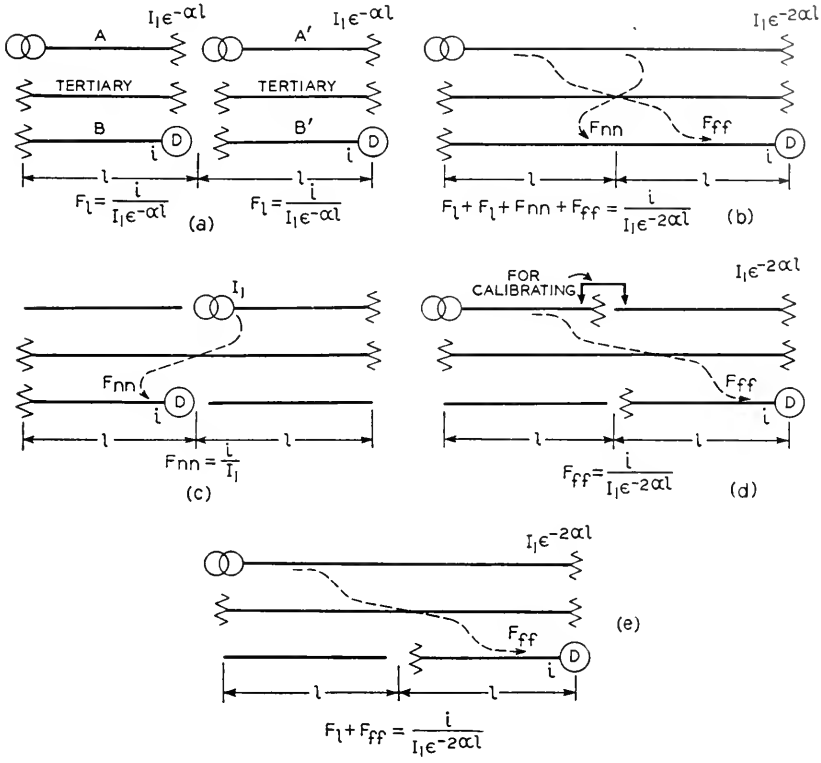


Fig. 2—Schematics illustrating far-end crosstalk summation.

near-end near-end and the component F_{ff} far-end far-end interaction crosstalk. Although inseparable under normal line conditions, these components are definite physical entities and can be isolated as shown schematically on Sketches (c) and (d) of Fig. 2. Thus, both F_{nn} and F_{ff} can be measured readily. In addition it is possible to measure directly $F_l + F_{ff}$ as shown on Sketch (e).

This interaction crosstalk between sections is due to crosstalk currents introduced into the outer conductor-sheath tertiary circuit in one

section and propagated along this circuit into the next section and thence into the disturbed coaxial. Except for interaction crosstalk between sections the total crosstalk in $2l$ would simply be twice that in length l, that is, the crosstalk would be directly proportional to length.

Now, the expressions for far-end crosstalk due to such interactions between two sections each of length l are

$$F_{nn} = -\frac{Z_{\alpha\beta}^2}{4Z} \cdot \frac{\gamma_4}{4Z_{44}} \left[\frac{1 - \epsilon^{-(\gamma_4+\gamma)l}}{\gamma_4 + \gamma} \right]^2, \tag{8}$$

$$F_{ff} = -\frac{Z_{\alpha\beta}^2}{4Z} \cdot \frac{\gamma_4}{4Z_{44}} \left[\frac{1 - \epsilon^{-(\gamma_4-\gamma)l}}{\gamma_4 - \gamma} \right]^2. \tag{9}$$

Since the coefficients [9] outside of the brackets are the same for F_{nn} and F_{ff} the terms may be combined to give the total interaction crosstalk between the two sections, namely,

$$F_{nn} + F_{ff} = -\frac{Z_{\alpha\beta}^2}{2Z} \left[\frac{\gamma_4}{4Z_{44}} \left(\frac{\gamma_4^2 + \gamma^2}{(\gamma_4^2 - \gamma^2)^2} \right) - \frac{\gamma_4}{4Z_{44}} \left(\frac{\epsilon^{-(\gamma_4-\gamma)l}}{(\gamma_4 - \gamma)^2} + \frac{\epsilon^{-(\gamma_4+\gamma)l}}{(\gamma_4 + \gamma)^2} \right) + \frac{\gamma_4}{4Z_{44}} \left(\frac{\epsilon^{-2(\gamma_4-\gamma)l}}{2(\gamma_4 - \gamma)^2} + \frac{\epsilon^{-2(\gamma_4+\gamma)l}}{2(\gamma_4 + \gamma)^2} \right) \right]. \tag{10}$$

As mentioned before, the crosstalk in length $2l$ *exclusive* of interactions *between* the two sections is equal to $2F_l$ or equation (5) multiplied by 2, namely,

$$2F_l = \frac{Z_{\alpha\beta}^2}{2Z} \left[\left(\frac{2l}{Z_{33}} - \frac{2l}{4Z_{44}} \cdot \frac{\gamma_4^2}{\gamma_4^2 - \gamma^2} \right) + \frac{2\gamma_4}{4Z_{44}} \left(\frac{\gamma_4^2 + \gamma^2}{(\gamma_4^2 - \gamma^2)^2} \right) - \frac{2\gamma_4}{4Z_{44}} \left(\frac{\epsilon^{-(\gamma_4-\gamma)l}}{2(\gamma_4 - \gamma)^2} + \frac{\epsilon^{-(\gamma_4+\gamma)l}}{2(\gamma_4 + \gamma)^2} \right) \right]. \tag{11}$$

The total crosstalk in length $2l$ is then the sum of (10) and (11), namely,

$$F_{2l} = 2F_l + F_{nn} + F_{ff} = \frac{Z_{\alpha\beta}^2}{2Z} \left[\left(\frac{2l}{Z_{33}} - \frac{2l}{4Z_{44}} \cdot \frac{\gamma_4^2}{\gamma_4^2 - \gamma^2} \right) + \frac{\gamma_4}{4Z_{44}} \left(\frac{\gamma_4^2 + \gamma^2}{(\gamma_4^2 - \gamma^2)^2} \right) - \frac{\gamma_4}{4Z_{44}} \left(\frac{\epsilon^{-2(\gamma_4-\gamma)l}}{2(\gamma_4 - \gamma)^2} + \frac{\epsilon^{-2(\gamma_4+\gamma)l}}{2(\gamma_4 + \gamma)^2} \right) \right], \tag{12}$$

[9] These near-end near-end and far-end far-end coefficients are equal because the coupling through a coaxial is of a series voltage character. In open wire and non-shielded cables where there is also present coupling due to shunt admittances the coefficients for F_{nn} and F_{ff} are different in magnitude and their effects must be considered separately. See paper by A. G. Chapman in *Bell Sys. Tech. Jour.* for January and April, 1934.

wherein the second term in equation (10) is cancelled completely by the third term of equation (11). This equation (12) is exactly what we would get by substituting $2l$ for l in the general equation (5). The only reason for deriving it in terms of $2F_l$ plus interaction between the sections is to present a better physical picture of the mechanism of far-end crosstalk summation with length, that is, to show how the interaction crosstalk *between* two sections alters what otherwise would be a direct summation with length.

In lengths where the tertiary circuit is electrically long equation (12) for total crosstalk in length $2l$ becomes

$$F_{2l} = 2F_l + F_{nn} + F_{ff} = \frac{Z_{\alpha\beta}^2}{2Z}\left[\left(\frac{2l}{Z_{33}} - \frac{2l}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}\right) + \frac{\gamma_4}{4Z_{44}}\left(\frac{\gamma_4^2 + \gamma^2}{(\gamma_4^2 - \gamma^2)^2}\right)\right], \quad (13)$$

which differs from equation (6) for total crosstalk in length l only by the factor of 2 in the first bracketed term. Thus, as mentioned before, there is a range of lengths wherein the crosstalk will be constant at a level determined by the second term of (6) or (12) until the length becomes sufficient for the first term to become controlling.

In lengths where the tertiary circuit is electrically short equation (11) becomes

$$2F_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\frac{2l}{Z_{33}} - \frac{\gamma_4}{4Z_{44}}\cdot l^2\right], \quad (14)$$

which reduces simply to

$$2F_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\frac{2l}{Z_{33}}\right] = \left[\frac{Z_{\alpha\beta}^2}{ZZ_{33}}\right] l \quad (15)$$

when the length is sufficiently short. The interaction crosstalk *between* two electrically short lengths becomes, from equation (10),

$$F_{nn} + F_{ff} = \frac{-Z_{\alpha\beta}^2}{2Z}\left[\frac{\gamma_4}{4Z_{44}}\cdot l^2\right] = \left[\frac{-Z_{\alpha\beta}^2}{8ZZ_4}\right] l^2, \quad (16)$$

one-half of which is due to component F_{nn} and the other half to component F_{ff}. The sum of (14) and (16) is

$$F_{2l} = 2F_l + F_{nn} + F_{ff} = \frac{Z_{\alpha\beta}^2}{2Z}\left[\frac{2l}{Z_{33}} - \frac{\gamma_4}{4Z_{44}}\cdot 2l^2\right], \quad (17)$$

which is exactly equal to equation (7) if $2l$ is substituted for l therein. From (15) and (16) it is apparent that for very short lengths the total

crosstalk in length $2l$ will be simply twice that in length l since the interaction crosstalk between lengths l is proportional to l^2 and therefore is negligibly small.

The view of the mechanism of far-end crosstalk summation as developed above is illustrated by measurements to be presented in Part II. It may be pointed out here that the measurement of far-end and interaction crosstalk in phase and magnitude on short lengths where equations (15) and (16) hold gives the far-end and interaction crosstalk coefficients from which the crosstalk in any length of line may be computed provided the propagation constants and impedances of the coaxial and the tertiary circuits are known.

A practical difficulty may arise from the fact that the application of this method involves equations (12) or (5) where the first bracketed term consists of the difference of two quantities each of which is very large compared with this difference. Thus, a considerable error may be introduced in the computation of this term because of small errors in the measurement of its components. For some cases it is, therefore, better to use a method based on certain crosstalk measurements in a short length of cable with the tertiary circuits open and shorted.[10] There are cases, however, where the controlling crosstalk in a five-mile section is predominantly due to the second term of equation (5). One such case is for the crosstalk between diagonally opposite coaxials in a four-coaxial cable. In this case tests have shown that the cancellation of components in the first term is so complete that the second term is controlling in five miles. For such a case the more accurate method may be to determine the interaction coefficient from equation (16).

Near-End Crosstalk

It will be sufficient here to give simply the final equations for the two crosstalk components for any length l.

For the component which would exist for two contacting coaxials in free space we have

$$N_3 = \frac{Z_{\alpha\beta}^2}{2Z} \cdot \frac{1}{Z_{33}} \left(\frac{1 - \epsilon^{-2\gamma l}}{2\gamma} \right) \tag{18}$$

and for that component due to the presence of the sheath

$$N_4 = -\frac{Z_{\alpha\beta}^2}{2Z} \cdot \frac{1}{4Z_{44}} \left[\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2} \cdot \frac{1 - \epsilon^{-2\gamma l}}{2\gamma} - \frac{\gamma_4^2}{\gamma_4^2 - \gamma^2} \cdot \frac{1 - 2\epsilon^{-(\gamma_4+\gamma)l} + \epsilon^{-2\gamma l}}{2\gamma_4} \right], \tag{19}$$

[10] The method described in a companion paper by K. E. Gould.

whence, for both components,

$$N_3 + N_4 = N_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\left(\frac{1}{Z_{33}} - \frac{1}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}\right)\frac{1 - \epsilon^{-2\gamma l}}{2\gamma} + \frac{1}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}\left(\frac{1 - 2\epsilon^{-(\gamma_4+\gamma)l} + \epsilon^{-2\gamma l}}{2\gamma_4}\right)\right]. \quad (20)$$

In a section where the tertiary circuit is electrically long equation (20) reduces to

$$N_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\left(\frac{1}{Z_{33}} - \frac{1}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}\right)\frac{1 - \epsilon^{-2\gamma l}}{2\gamma} + \frac{1}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}\left(\frac{1 + \epsilon^{-2\gamma l}}{2\gamma_4}\right)\right] \quad (21)$$

and when l is electrically short it reduces to

$$N_l = \frac{Z_{\alpha\beta}^2}{2Z}\left[\frac{l}{Z_{33}} - \frac{\gamma_4}{4Z_{44}}\cdot\frac{l^2}{2}\right], \quad (22)$$

which is the same as for far-end crosstalk in very short lengths as given in equation (7).

As pointed out earlier the expression for near-end crosstalk even when the tertiary circuit is electrically long is more complicated in form than for far-end crosstalk because of the terms $1 - \epsilon^{-2\gamma l}$ and $1 + \epsilon^{-2\gamma l}$. This may be seen by comparing formulas (6) and (21).

Nevertheless it is possible to see from (21) that the presence of the tertiary circuit acts to reduce near-end crosstalk as it did in the case of far-end crosstalk. The first term of (21) is less than the near-end crosstalk without the sheath (equation (18)) by the factor

$$\frac{\frac{1}{Z_{33}} - \frac{1}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}}{\frac{1}{Z_{33}}} = 1 - \frac{Z_{33}}{4Z_{44}}\cdot\frac{\gamma_4^2}{\gamma_4^2 - \gamma^2}.$$

This is the same factor by which far-end crosstalk is reduced in very long lengths as brought out in the discussion of equation (6). However, the second term in equation (21) prevents this complete reduction from ever taking place in the case of near-end crosstalk.

PART II—EXPERIMENTAL RESULTS

The crosstalk measurements presented here were made on and between sections of twin coaxial cable of various lengths from 73 feet

to about five miles. Primarily the tests were made to indicate the effect of sheath and quads upon the summation of crosstalk with length as a check on theoretical considerations and were the first extensive tests made on a coaxial cable with this end in view. The layup of the cable is shown in Fig. 3.

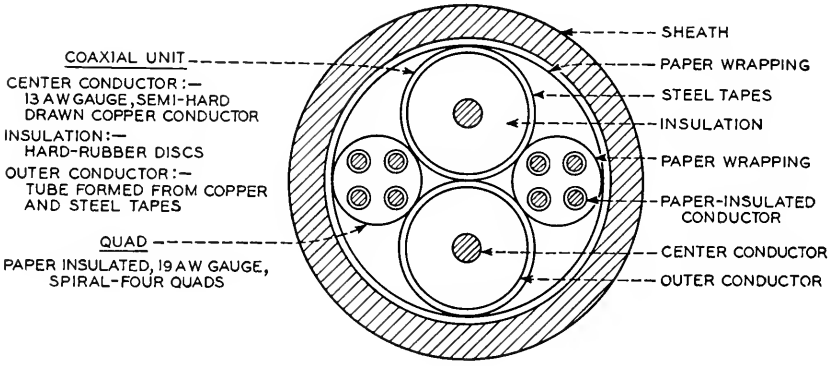


Fig. 3—Cross-section of twin coaxial cable.

As indicated in the latter portion of Part I the general procedure was to measure crosstalk in available sections of equal length, l, with the tertiary circuits terminated in approximately characteristic impedance. Interaction crosstalk between these sections was then measured and finally the two sections were combined to find the total crosstalk in length $2l$. This process was repeated until a total length of about five miles was built up.

Far-End Crosstalk Summation

The results of crosstalk tests on 73 and 146-foot lengths are shown in Fig. 4. The letters on the curves correspond to the crosstalk components discussed in Part I. Only far-end far-end interaction crosstalk was measured but for such short lengths the near-end near-end crosstalk would be nearly the same.

Remembering from the discussion in Part I that the total crosstalk F_{2l} in length $2l$ is equal to $2F_l + F_{nn} + F_{ff}$ it is evident that since in this case the measured components F_{nn} or F_{ff} are quite small the crosstalk in 146 feet should be approximately $2F_l$. That this is the case may be seen from the measured crosstalk in 146 feet which is about 6 db higher than for 73 feet. These lengths are apparently short enough for equations (15) and (16) to hold reasonably well at the lower frequencies.

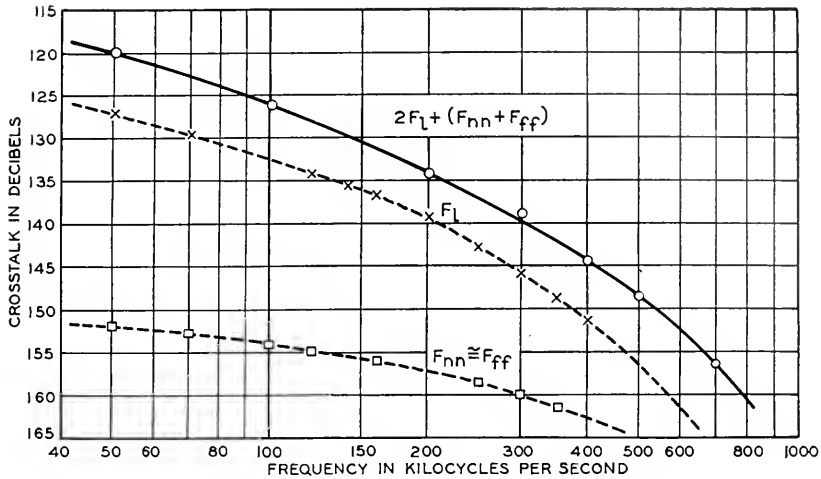


Fig. 4—Crosstalk components in 73-foot and 146-foot lengths.

Now suppose that we consider two lengths which are considerably longer so that equations (6) and (13) more nearly apply. Figure 5 shows the results of tests on and between two 1500-foot cable sections. Here, in contrast with the 73-foot measurements, components F_l and F_{ff} are nearly equal in magnitude while F_{nn} is quite small. Also, F_{ff} and F_l are in general phase opposition since their sum, $F_l + F_{ff}$, is considerably less than either component alone.

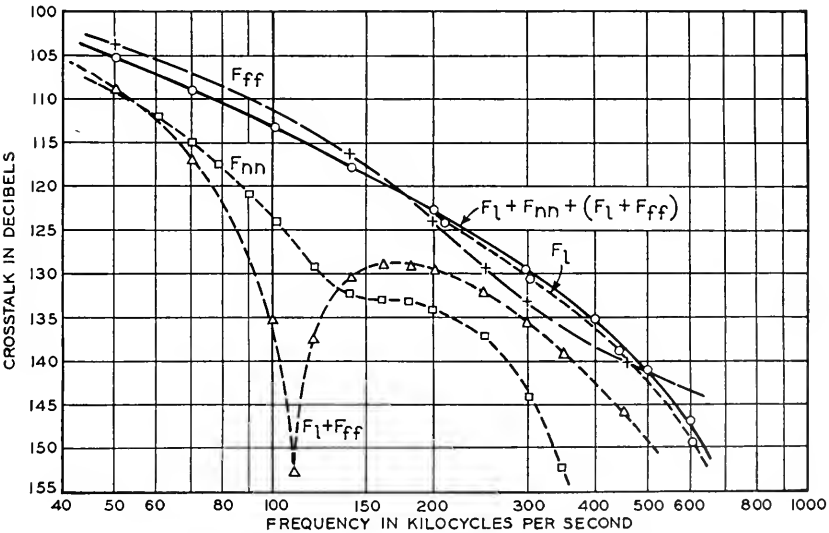


Fig. 5—Crosstalk components in 1500-foot and 3000-foot lengths.

Reference to equations (6) and (9) show that this tendency to cancel is to be expected provided the second term of (6) is the controlling term in F_l. Indeed, in lengths where the tertiary is electrically long, equations (8) plus (9) should exactly cancel the second term of (6). In other words, the total interaction crosstalk *between* two such sections should cancel a portion of the interaction crosstalk *within* a section. Since the portion which is cancelled is the controlling term the net result is that when two sections are combined the total crosstalk in length $2l$ is no more than was measured in length l, as evidenced by the measured curve $F_l + F_{nn} + (F_l + F_{ff})$ of Fig. 5.

This effect persists when two 3000-foot lengths are combined to form a 6000-foot section, as illustrated by the curves of Fig. 6. Here

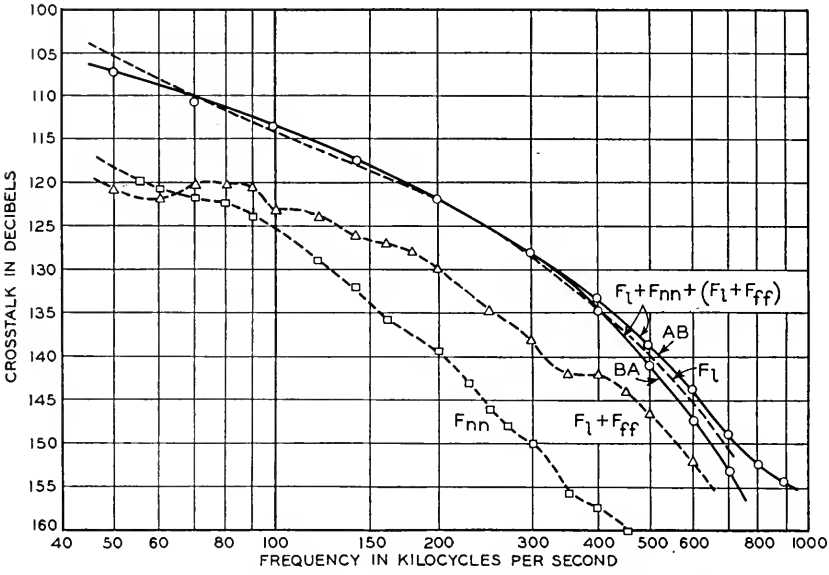


Fig. 6—Crosstalk components in 3000-foot and 6000-foot lengths.

again $(F_l + F_{ff})$ and F_{nn} are considerably smaller in magnitude than F_l so that the total crosstalk in 6000 feet cannot differ materially from the value F_l measured in 3000 feet.

The curves labelled AB and BA were made by using first coaxial A and then coaxial B as the disturbing circuit. The difference between the curves indicates that there is a certain amount of random unbalance within the section. For example, random deviations in the shielding of the two coaxials from a nominal value would result in different values of interaction crosstalk when the disturbed and dis-

turbing circuits are interchanged. The direct crosstalk component would not exhibit this effect.

The results of tests on 6000 and 12,000-foot lengths are given in Fig. 7. Again, the trend is in the same direction as in Figs. 5 and 6 except that in this case $(F_l + F_{ff})$ is nearly equal to F_l and has an appreciable influence when the two components are combined to give far-end crosstalk in 12,000 feet. This indicates that the first term of F_l in equation (6) is becoming more important as l is increased as would be expected.

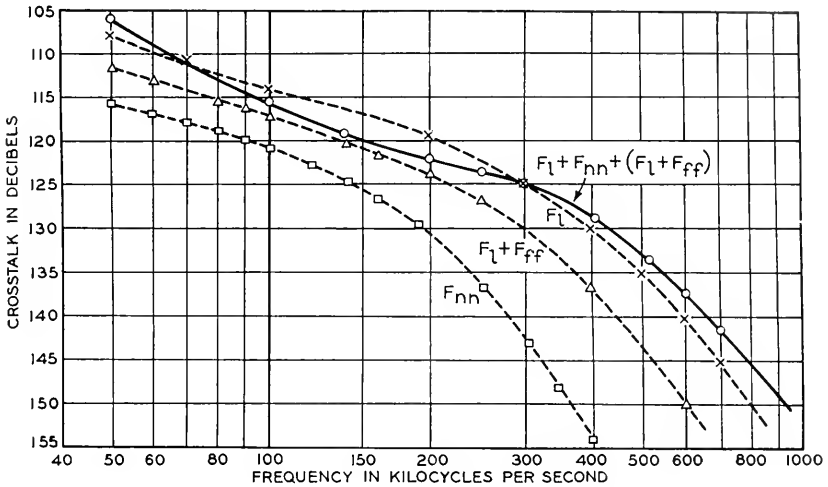


Fig. 7—Crosstalk components in 6000-foot and 12,000-foot lengths.

It may be noted here that curve F_l in Fig. 7 differs considerably from the AB and BA curves of $F_l + F_{nn} + (F_l + F_{ff})$ in Fig. 6 although all represent far-end crosstalk in 6000-foot sections. These differences in magnitude must be due to differences in the construction of the two cable sections. The difference between the curves varies from 3 to 8 db in the frequency range above 200 kilocyles. However, up to about 150 kilocycles the differences are not greater than 1 db. At the higher frequencies such differences naturally will introduce difficulties in any analysis since they superpose sizeable random effects on the major component of crosstalk which is systematic.

The curves in Fig. 8 present far-end crosstalk tests on 12,000 and 24,000-foot lengths. Here F_l and $(F_l + F_{ff})$ are of the same order of magnitude and combine in such a way that the crosstalk in 24,000 feet is from 3 to 6 db higher than that measured in 12,000 feet. Com-

ponent F_{nn} is again negligible. This behavior indicates that the first term of equation (6) is controlling as the length is increased.

It appears from all these tests that the magnitude of the far-end crosstalk in this cable with tertiaries terminated does not vary materially from 1500-foot to 12,000-foot cable lengths, except for random effects. In other words, for this range of lengths the second term of (6) is controlling. For very short lengths the crosstalk varies directly with length due to the absence of interaction crosstalk of sufficient magnitude to exert any influence. Also, in going from 12,000 to 24,000 feet, there is a definite indication that the crosstalk is increasing with

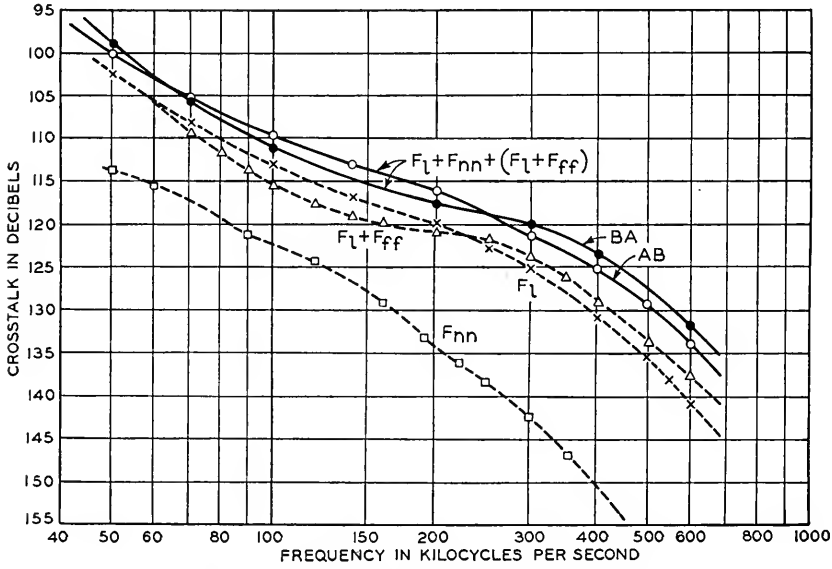


Fig. 8—Crosstalk components in 12,000-foot and 24,000-foot lengths.

length, so that for lengths over 24,000 feet the crosstalk would again tend to be proportional to length. We have shown in Part I that on the basis of theoretical considerations this law of crosstalk summation with length might be expected.

To illustrate this measured behavior the far-end crosstalk versus length for frequencies of 50, 100 and 200 kilocycles has been plotted on Fig. 9. For comparison are also plotted dashed curves based on the 73-foot tests and computed on the assumption that the crosstalk is directly proportional to length. The difference between corresponding curves shows the influence of the tertiary circuits. For a 24,000-foot length this difference amounts to 23, 26 and 27 db at 50, 100 and 200 kilocycles, respectively.

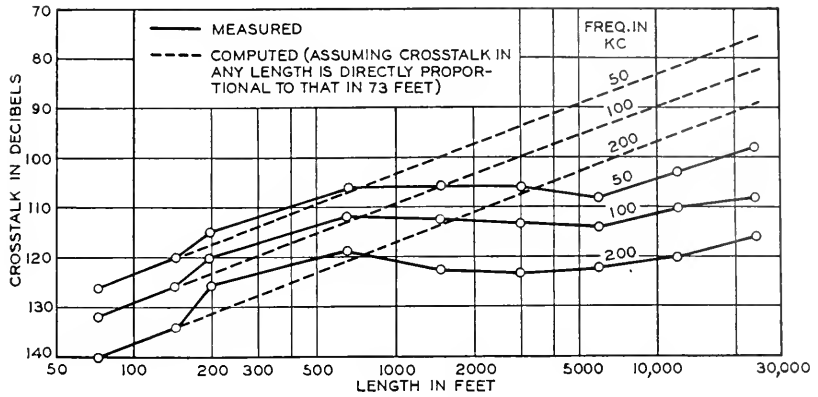


Fig. 9—Far-end crosstalk vs. length with tertiary terminated.

Near-End Crosstalk Summation

The curves on Fig. 10 show the amount of near-end crosstalk reduction due to the presence of the sheath and quads for a length of about five miles. The upper curve was computed from tests on a 73-foot length with tertiaries terminated by raising the values measured

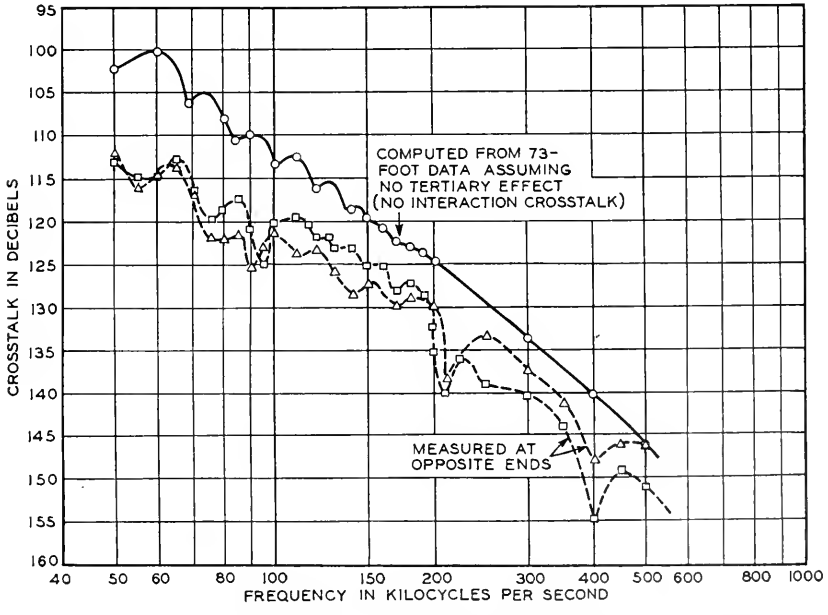


Fig. 10—Near-end crosstalk in a 5-mile length with tertiary terminated.

there by the factor

$$\frac{1 - \epsilon^{-2\gamma L}}{2\gamma l},$$

where $l = 73$ feet and $L = 5$ miles. This is the crosstalk which would exist in five miles *in the absence of a sheath and quads*.

The lower two curves were measured at opposite ends of the cable and the difference of about 10 db between these curves and the upper curve is due to the tertiary circuit effects. As might be expected from the discussion of equation (21), this reduction is considerably less than in the case of far-end crosstalk.

Interaction Crosstalk Between Sections

The methods of measuring the various types of interaction crosstalk between two sections have already been discussed in reference to Fig. 2. Besides showing the influence of interaction crosstalk in the summation of crosstalk *within* a repeater section the results presented below are indicative of the importance of interaction crosstalk which takes place *between* repeater sections, that is, around repeaters, when all or only a part of the tertiary is continuous at repeater points.

Values of near-end near-end interaction crosstalk, F_{nn}, were measured between various section lengths from 73 to 12,000 feet. It was found that the results are roughly independent of the section lengths above 1500 feet, and curve F_{nn} of Fig. 11 for the crosstalk measured between two 12,000 foot sections is typical. This independence of length is because of the high attenuation of the tertiary circuits which annihilates the effects of crosstalk in the more remote portions of the sections as may be seen from equation (8) if γ_4 is made large. The relatively unimportant contribution of this type of interaction crosstalk to the summation of far-end crosstalk *within* a repeater section has been discussed.

Similarly, measured values of far-end far-end and near-end far-end interaction crosstalk between various sections lengths were found to be practically independent of length above 1500 feet. Curves F_{ff} and N_{nf} of Fig. 11 for the crosstalk between 12,000-foot sections are typical. The far-end far-end component of interaction crosstalk has an important influence on the summation of far-end crosstalk within a repeater section as already mentioned in the section on far-end crosstalk summation. The influence of near-end far-end interaction crosstalk N_{nf}, on the summation of near-end crosstalk within a repeater section has not been very thoroughly investigated here but it is respon-

sible for the results described in the discussion of near-end crosstalk in a five-mile length.[11]

The relative importance of various tertiary circuits in the production of interaction crosstalk between two sections was studied for the case of near-end near-end crosstalk between two 12,000-foot lengths. It was found that the outer conductor-quads and outer conductor-sheath circuits were about equally important and that crosstalk via the quad-sheath tertiary circuit was from 20 to 30 db less. These results are about as expected since the outer conductors are the source of the

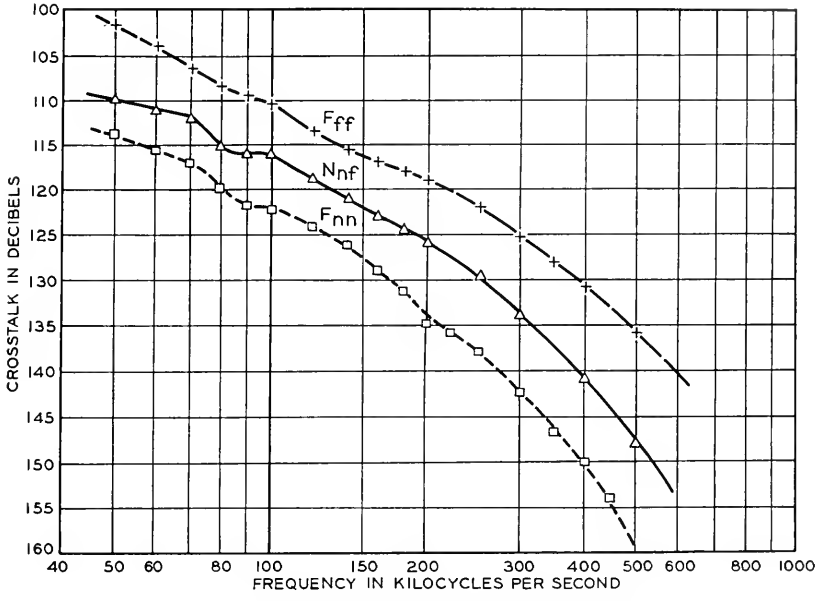


Fig. 11—Interaction crosstalk between two 12,000-foot lengths.

tertiary emf and thus the tertiary circuits involving the outer conductors should be the important ones. It is therefore permissible to consider sheath and quads as a single unit as was done in Part I.

Effectiveness of Transpositions on Far-End Crosstalk Reduction

In a long repeatered system the far-end crosstalk measured in successive individual sections inherently tends to sum up directly since all

[11] It should be noted that while Fig. 11 shows the measured values of the three types of interaction crosstalk between two 12,000-foot sections, the relative importance of the various types acting *between* repeater sections, that is, around repeaters, is not as shown there, since different correction factors have to be applied when estimating the total crosstalk at system terminals.

repeaters have practically the same phase shift and the propagation characteristics of the two coaxials are nearly identical. One way to prevent this direct addition is to transpose one section against another or one group of sections against another group along the line. In the case of unbalanced circuits these "transpositions" take the form of transformers or extra tube stages in one of the systems at repeaters, either of which will produce a 180-degree phase reversal.

If the far-end crosstalk in one transposition section is F_{l1} and that in another is F_{l2} the total in the two sections, *exclusive of interaction crosstalk between sections*, is inherently $F_{l1} + F_{l2}$. With a transposition in one coaxial at the junction the total becomes $F_{l1} - F_{l2}$. Hence, if $F_{l1} = F_{l2}$ it is possible to eliminate this crosstalk component entirely. However, due to irregularities in the cable and the practical impossibility of locating repeater points exactly, F_{l1} will not, in general, equal F_{l2} and even after transposing a small residual may remain.

This residual, however, may be negligible compared with the near-end near-end and far-end far-end interaction crosstalk components F_{nn} and F_{ff} between repeater sections (that is, around repeaters), unless transmission along the tertiary circuits from one repeater section into another is suppressed at repeater points. The interaction crosstalk tests already discussed may be used to compute this effect. However, in order to demonstrate the effectiveness of transpositions, far-end crosstalk tests were made in a 24,000-foot length with and without a transposition in one of the coaxials at the center and with various interaction crosstalk paths suppressed. The results are given in Figs. 12 to 14 and are discussed below.

To suppress entirely the interaction crosstalk between the transposed sections all tertiary circuits were shorted at the transposition point. In these measurements the tertiaries were also shorted at each end of the line in an effort to have both ends of each half of the line terminated as nearly alike as possible. The test results are given in Fig. 12.

For this condition the crosstalk measured in each half of the line is also shown. Curve AB represents the far-end crosstalk in one line section and $A'B'$ that in the other section. Curve $(AB + A'B')$ gives the results when the two sections are combined with no transposition. Curve $(AB - A'B')$ gives the results when a transformer is inserted in one coaxial at the center. (A similar set of curves are given for BA, $B'A'$, etc.)

Note that AB and $A'B'$ coincide very closely in magnitude. When combined with no transposition the crosstalk in two sections is nearly 6 db higher over the entire frequency range than in either individual

section. When combined with a transposition the crosstalk in two sections is from 13 to 27 db below either individual section over the frequency range. *Such a reduction is possible only because AB and A'B' are so nearly equal.*

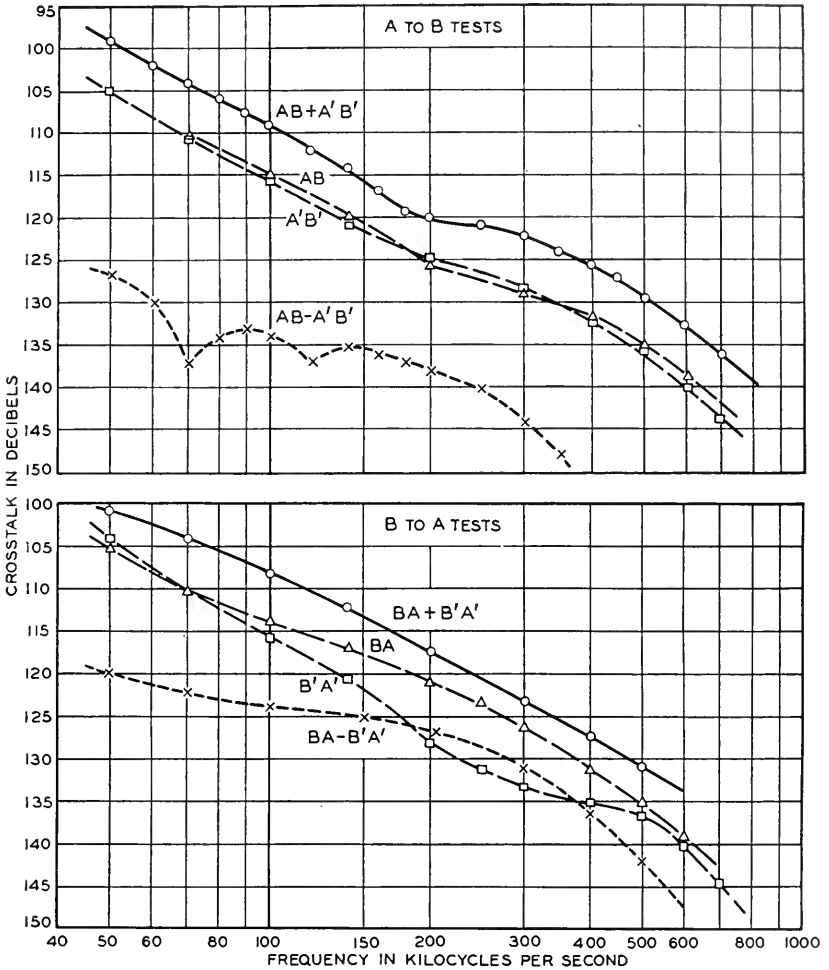


Fig. 12—Effect of a transposition on far-end crosstalk in a 24,000-foot length with all tertiary circuits suppressed at the transposition.

In contrast, BA and $B'A'$ may be seen to differ considerably from each other at the higher frequencies. As a result, the transposition is not nearly so effective in that range. The improvement at the lower frequencies where it is needed most is still about 20 db.

In order to suppress only a portion of the interaction crosstalk between two sections, measurements were made with the coaxial outer conductor-sheath circuit shorted at the transposition point thus permitting continuity of the quad-outer conductor tertiary circuit. This tertiary circuit had been shown previously to be an important

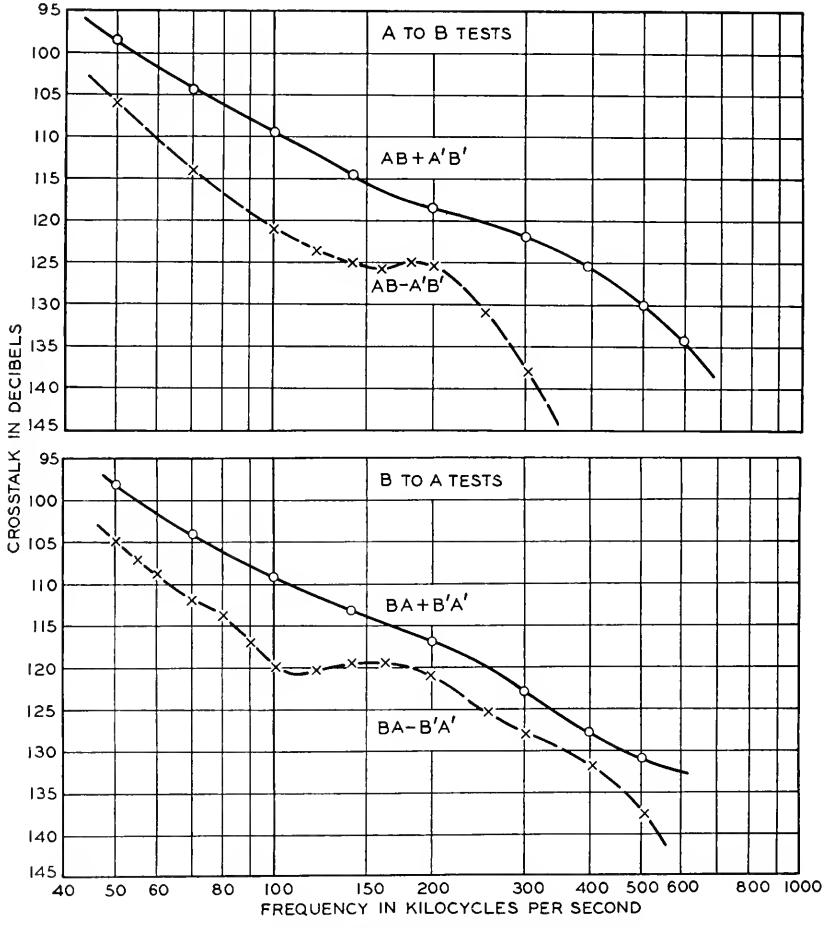


Fig. 13—Same as Fig. 12 except with quad-outer conductor tertiary circuit continuous past the transposition.

one in the production of interaction crosstalk. The measured far-end crosstalk results are given in Fig. 13.

It is at once apparent that the transposition is not so effective in this case. The crosstalk remaining after transposing is about what would be expected due to interaction crosstalk between sections via the quad-

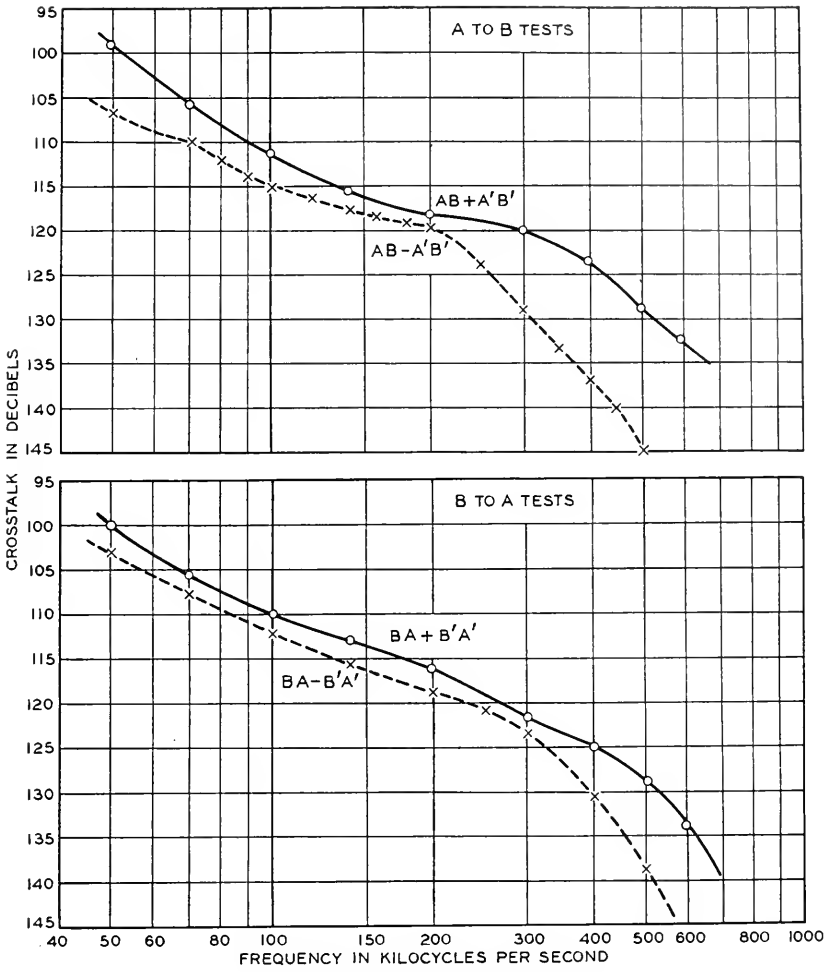


Fig. 14—Same as Fig. 12 except with all tertiaries continuous past the transposition.

outer conductor tertiary circuit.[12] However, a certain portion is also due to differences between AB and $A'B'$ (or BA and $B'A'$).

On Fig. 14 are plotted far-end crosstalk values when two 12,000-foot sections are combined with and without a transposition in one coaxial

[12] It should be noted here that these tests indicate directly the effect of a transposition *at the center* of a 24,000-foot section rather than at a *junction* between two repeater sections in a long repeatered system. If 12,000-foot repeater spacing is assumed with the transposition at the repeater point it is necessary to reduce the measured far-end far-end interaction crosstalk and increase the measured near-end near-end interaction crosstalk by an amount equal to the line loss in 12,000 feet. These corrections put interaction crosstalk between repeater sections on an output-to-output or equal level basis.

at the center and when all tertiary circuits are continuous at the transposition point and terminated at the ends. Curve ($AB + A'B'$) gives the results when the two sections are combined with no transposition. Curve ($AB - A'B'$) shows the result when a transformer is inserted in one coaxial at the junction. (A similar set of curves is given for BA, $B'A'$, etc.)

It is seen that in the 50–200 kc range there is an improvement in overall crosstalk of from 3 to 8 db due to the transposition. However, the overall crosstalk in the combined sections with a transposition is not appreciably less than that in an individual 12,000-foot section as shown by curve F_l on Fig. 8. Reference to Fig. 11 shows that this is due mainly to the far-end far-end interaction crosstalk between the two sections which is unaffected by the transposition.

The results shown in Fig. 12 give some indication of the extent to which far-end crosstalk may be reduced by means of a transposition, *provided interaction crosstalk between sections is entirely suppressed.* As illustrated in Figs. 13 and 14 a transposition at a repeater point is not nearly so effective if the interaction crosstalk is not suppressed.

Acknowledgment

The authors are greatly indebted to Mr. John Stalker and the staff at the Princeton, New Jersey, repeater station of the American Telephone and Telegraph Company and to Mr. William Bresley and Mr. Norman Mathew of the New Jersey Bell Telephone Company, for their cooperation and assistance in the Princeton tests.

Compressed Powdered Molybdenum Permalloy for High Quality Inductance Coils *

By V. E. LEGG and F. J. GIVEN

Molybdenum-Permalloy is now produced in the form of compressed powdered cores for inductance coils. Its high permeability and low losses make possible improved coil quality, or decreased size without sacrificing coil performance. Its low hysteresis loss reduces modulation enough to permit application where large air core coils would otherwise be required.

Introduction

THE introduction of loading coils in the telephone system at about the turn of the century brought special demands on magnetic and electrical properties of core materials, and set in motion investigations which have had wide influence on the theoretical and practical aspects of ferromagnetism. The first step in this development led to cores of iron wire, which sufficed for loading coils on circuits of moderate length.[1] With the development of telephone repeaters and the extension of circuits to transcontinental length some twenty-five years ago, there arose need not only for loading coils, but also for network coils, which would have high stability with time, temperature and accidental magnetization. Magnetic stability was at first secured [2] by employing iron wire cores provided with several air gaps. Later, commercial and technical considerations led to a core structure made from compressed insulated powdered material, first electrolytic iron [3] and later permalloy powder.[4] This type of core is mechanically stable; it introduces in an evenly distributed fashion the requisite air-gaps, while avoiding undesirable leakage fields; and it sub-divides the magnetic material so as to reduce eddy-current losses. Although other means have been suggested,[5, 6] no way has yet been devised which provides these features so well and at so low a cost as the compressed powdered type of core.

Loading coil cores made from electrolytic iron powder generally satisfied the stability requirements for long lines, but on account of their low magnetic permeability they were large and costly. The search for materials with higher permeability and lower hysteresis loss

* Presented at Winter Convention of A.I.E.E., New York, N. Y., January 22–26, 1940.

led to permalloy [7] which, by 1925, had been produced in powdered form and fabricated into cores. This development provided coils for voice frequency applications (loading coils and filter coils) which were cheaper and yet superior electrically to those made from electrolytic iron. These coils became available at a time when the telephone plant was undergoing a very large extension of loaded cables. As a result, large economies in cost and space were realized in the more than six million coils involved in this plant expansion.

Iron powder, and later permalloy powder, ground to a finer size and diluted to lower permeability than used in loading coils, also found application in coils for oscillators, filters and networks of multiplex carrier telephone and telegraph systems employing frequencies up to 30 kc.[8] and in receivers for transoceanic radio telephone communication employing frequencies up to approximately 60 kc.[9] Permalloy powder improved the electrical characteristics—particularly modulation—of coils for use in high frequency circuits, because of its low hysteresis losses.

Continued research for a powdered material having still better intrinsic properties has recently made available new compressed powder cores which permit further important gains in coils for voice-frequency circuits and in coils for high-frequency carrier system applications. The latter take on considerable significance at this time because they play an important part in making practical for commercial use the new broad-band carrier telephone systems intended for use on existing open-wire and cable lines and on new types of cable. Again, therefore, the advent of a new core material is well-timed to be of assistance in further growth of the telephone system.

The development of this core material was based on the discovery [10] that the addition of a small percentage of molybdenum to permalloy increases its permeability and electrical resistivity, and decreases its eddy current and hysteresis losses. Decreased losses are necessary for improvements and economies for both voice and carrier frequency operation. The increased permeability of this alloy is essential for the improvement of voice-frequency coils. It is readily reduced to the proper values for high-frequency coils by diluting the powdered magnetic material with insulating material before compressing into core form. In this development many problems of alloy embrittlement, pulverization, insulation and heat treatment had to be solved both on a laboratory and factory scale. The alloy composition finally selected as giving the best combination of desirable properties, contains approximately 2 per cent molybdenum, 81 per cent nickel, and 17 per cent iron, and is designated as 2–81 molybdenum-permalloy. This new

alloy is manufactured commercially by the Western Electric Company for use in loading coils and filter coils.

Physical and Magnetic Characteristics

The raw materials and necessary embrittling agents [11] are melted together and cast into ingots which are rolled to develop the desired grain structure. The density of this alloy is 8.65 gm/cm³. The brittle material is pulverized to the desired fineness and finally annealed to soften the alloy particles before insulation and pressing into core form.

The distribution by weight of the particle sizes of a sample of 120-mesh powder is given in Fig. 1, showing a root mean square size of 50

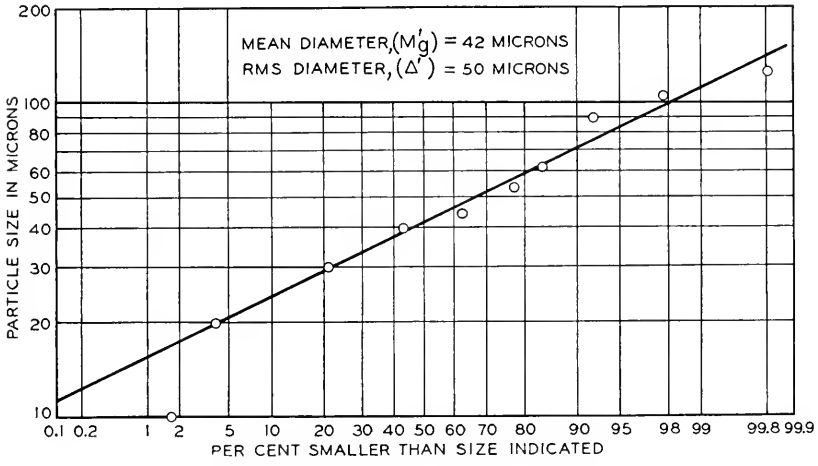


Fig. 1—Distribution of particle size of 120-mesh powder, by weight.

microns.[12] Since the effective resistance of a coil due to eddy-current losses in its core is proportional to the mean square particle diameter,[13] it can be decreased when desired by the use of more finely pulverized material.

The problem of insulating 2–81 molybdenum-permalloy powder is to coat the particles with a minimum thickness of a material which will not break away during the pressing operation, which will not fuse and flux the magnetic particles together during the core heat treatment, which will prevent the flow of eddy currents between metallic particles, and which will be chemically inert throughout the lifetime of the magnetic core. The difficulty of the problem will be appreciated from the fact that the separation between adjacent particles of a core of 125 permeability is approximately equal to the wave-length of visible

light (0.5 micron). This thickness of insulating film may be shown to be approximately $\frac{rt}{300p}$, where r is the percentage of insulating material by volume, p is the packing factor of the magnetic material, and t is the r.m.s. particle diameter (assuming spherical particles).

A new type of ceramic insulation has been introduced with these cores which fulfills the above requirements and which is more inert than the previous type. This new insulating material is free from water soluble residue. It thus eliminates the final washing treatment which was required with the earlier type.

For applications where a low permeability is desired, non-magnetic powder is added to further dilute the magnetic material. The permeability of the finished core depends largely on the quantity, particle size, and thoroughness of admixture of non-magnetic powder. Various attempts to derive theoretical relations between core permeability and dilution have been made,[14,15] but they generally fail in some detail. An empirical representation of this relationship is found to be

$$\mu = \mu_i^p \qquad \text{or} \qquad \log \mu = p \log \mu_i,$$

where μ_i is the intrinsic permeability of the magnetic material, and p is the packing factor, or fraction of the core volume occupied by magnetic material. This equation is found to be valid for a wide range of dilution, effected either by adding insulating material or by reducing the load during core compression. However, the intrinsic permeability must be determined experimentally for each type of particle, size distribution, method of admixture of non-magnetic powder, and annealing process. Figure 2 shows curves of permeability and percentage diluting material vs. metallic packing factor. A permeability of 125 has been selected for most loading coil cores, while permeabilities of 14 and 26 have been chosen for two important types of high-frequency filter coils.

A pressure of 100 tons/sq. in. is employed in forming molybdenum permalloy cores, to attain proper density and mechanical strength. The effect of pressure on core density and strength is shown in Fig. 3 for cores having 2.5 per cent dilution. The tensile strength of diluted cores is decreased somewhat; for example, cores with 25 per cent non-magnetic materials have a tensile strength of about 250 lbs./sq. in.

In annealing the compressed cores to remove stresses incident to pressing, it has been found that the insulating material remains intact at a considerably higher temperature if oxygen is excluded. An improved annealing treatment has therefore been introduced by which

the cores are heated in an atmosphere of hydrogen, with the attainment of high core permeability and low hysteresis loss. The ability of the insulating material to withstand such a heat treatment testifies

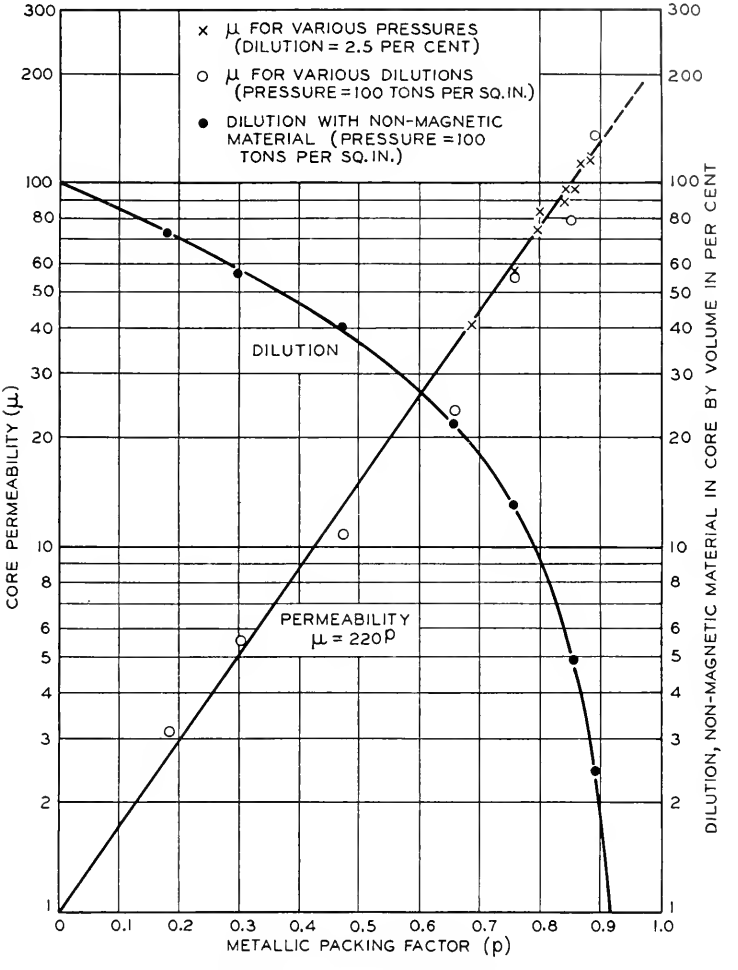


Fig. 2—Relation between metallic packing factor, permeability and percentage dilution.

to its extreme stability and recommends it in preference to organic materials.

An essential core requirement of precision inductance coils is that the permeability remain unaltered during the life of the coil. The greatest difficulty with cores having no air gaps is the large and more or

less permanent shift of permeability due to accidental strong magnetization. Such variations have recently been overcome to a degree in continuous cores made of hard rolled nickel-iron alloy sheet,[6] but they have been found to be rather large immediately after strong magnetization, decreasing slowly to tolerable limits only after two or three days. With compressed powdered molybdenum-permalloy cores, permeability shift due to strong magnetization is remarkably small even within a fraction of a minute after the magnetization is released, and any

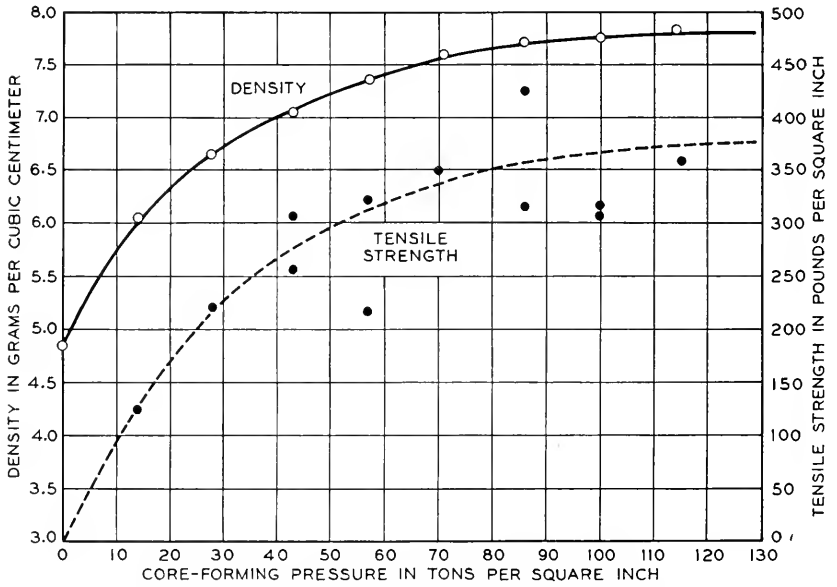


Fig. 3—Effect of core forming pressure on density and tensile strength.

further drift of permeability with time is negligible. In typical cores of the new material, the shift in permeability after strong magnetization is less than 0.2 per cent for cores of permeability 125, and less than 0.05 per cent for cores of permeability 14. Figure 4 shows the residual effect of the application and removal of various magnetizing forces on cores of both these permeabilities.

When a direct current is superposed on an alternating current in the windings of a coil, the inductance is altered because the magnetic field set up by the direct current modifies the core permeability. Figure 5 shows the effect of superposed d-c. fields on the permeability of 2–81 molybdenum-permalloy powder cores of various permeabilities.

A further important core property is the constancy of permeability with respect to flux density B. This is of particular importance in

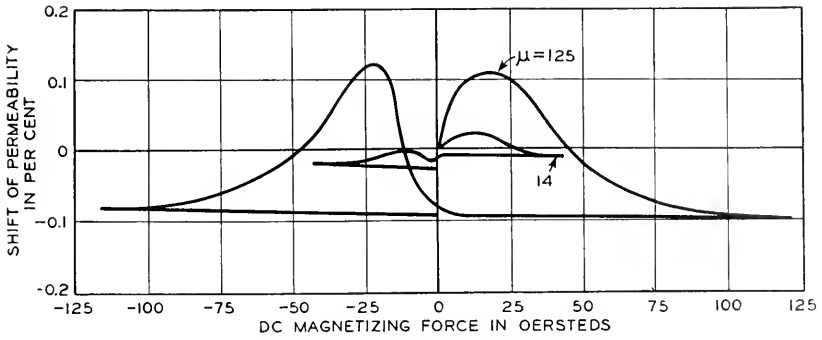


Fig. 4—Residual effect of d-c. magnetization on initial permeability—measured three minutes after release of direct current.

precision filters, to insure that changes in transmission level do not produce serious alterations in the frequency discrimination characteristics. Figure 6 shows the superiority of the new material over the earlier permalloy.

A new requirement for cores has been introduced by quartz crystal filters used in wide-band carrier systems. In order to secure the necessary precision in this type of filter, measures have to be taken to prevent departures from the initial frequency adjustment due to changes in core permeability ordinarily occurring with room temperature changes. Extremely small temperature coefficients of permeability have now been achieved by adding to the new 2–81 molybdenum-permalloy powder a very small percentage of special permalloy powder having a molybdenum content of about 12 per cent. Such an alloy has a non-magnetic or Curie point close to room temperature, and for a

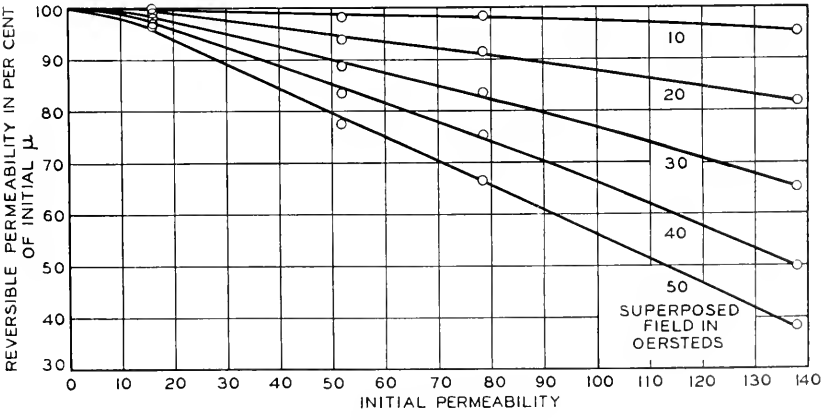


Fig. 5—Effect of superposed magnetization on permeability.

small temperature range just below its Curie point, it has a negative temperature coefficient several hundred times as large as the positive coefficient of 2–81 molybdenum-permalloy. By choosing suitable compositions and percentages of such compensating alloys, the net temperature coefficient of permeability of a core can be adjusted to any reasonable value, positive or negative, over a desired temperature range. Figure 7 shows a permeability vs. temperature curve for a

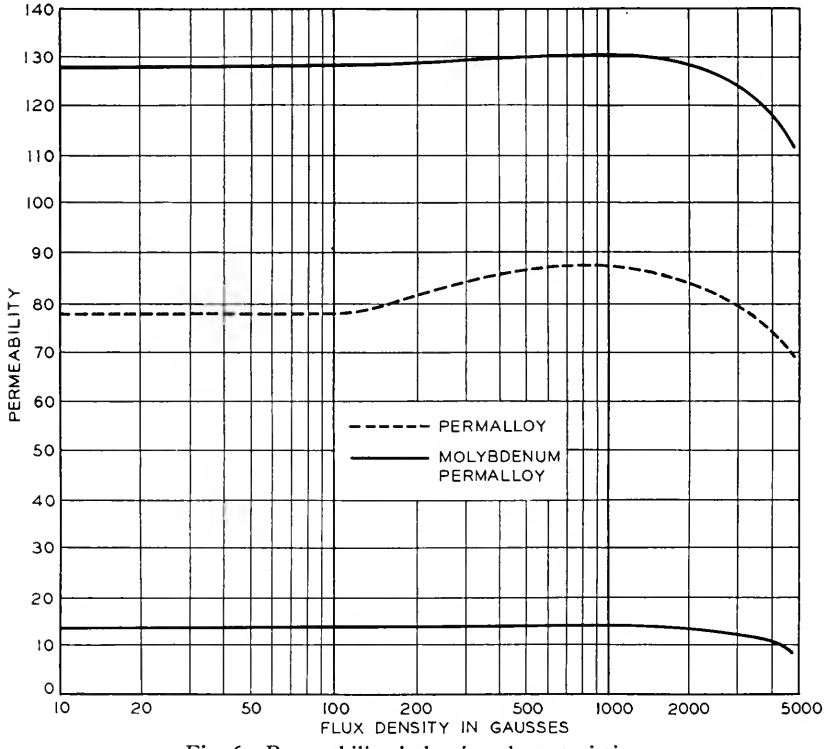


Fig. 6—Permeability-induction characteristics.

core stabilized to give a small negative coefficient, compared to a similar curve for a core not stabilized.

Core Losses

The desirability of core materials increases in general as their loss characteristics decrease. Low total eddy-current and hysteresis losses give low contributions to attenuation. Hysteresis loss is frequently of especial importance because it appears fundamentally as a resistance which varies with coil current, and because it incidentally generates harmonic voltages. A low value of hysteresis loss thus

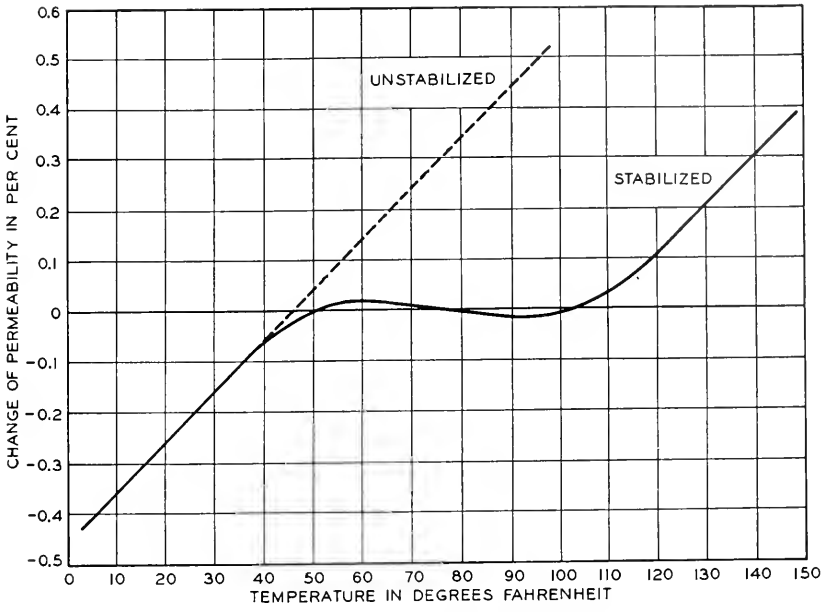


Fig. 7—Effect of temperature on permeability of stabilized and unstabilized cores.

simplifies circuit problems arising from resistances which vary with energy level, and it avoids troublesome modulation conditions.

The total resistance per unit inductance arising from eddy-current and hysteresis losses may be expressed as [16]

$$R_m/L = \mu(aB_m + c)f + \mu ef^2,$$

where the symbols are as given in the Appendix.

Table I gives the loss coefficients for various core materials, and for

TABLE I

Loss Coefficients of Powdered Core Materials

Material	μ	Hysteresis		Residual		Eddy Current	
		$a\times10^6$	$\mu a\times10^3$	$c\times10^6$	$\mu c\times10^3$	$e\times10^9$	$\mu e\times10^6$
Grade B Iron	35	49	1.7	109	3.8	88	3.1
Grade C Iron	26	81	2.1	139	3.6	31	0.8
81 Permalloy	75	5.5	0.41	37	2.8	51	3.8
	26	11.5	0.30	108	2.8	27	0.7
2–81 Molybdenum Permalloy	125	1.6	0.20	30	3.8	19	2.4
	26	6.9	0.18	96	2.5	7.7	0.2
	14	11.4	0.16	143	2.0	7.1	0.1

2–81 molybdenum-permalloy insulated to several permeabilities. The low loss coefficients of the new material as compared with the best previous materials are of importance from two standpoints. First, core permeabilities as much as 50 per cent greater can be now utilized in coils without increasing the total core loss resistance. Second, by utilizing the same permeabilities, core loss resistances about 60 per cent smaller can be obtained.

In many coil design problems, harmonic generation or modulation assume controlling importance. The modulation factor m which denotes the ratio of the generated third harmonic to the applied voltage may be expressed as follows [17]

$$m = E_3/E_1 = 3\mu a B_m/10\pi.$$

The low values of a obtained with the new material yield values of m that are about 6 db and 20 db lower than possible with powdered permalloy and electrolytic iron cores, respectively.

The wide range of core permeability available with this new material permits a ready choice of the proper values of permeability and core size to suit any particular needs. In the usual design problem the following main requirements must be considered, in addition to providing the desired inductance.

1. D-C. Resistance, R_c.
2. Coil quality factor, $Q = \omega L/(R_c + R_m)$.
3. Modulation Factor, m.
4. Coil size (which depends directly on core size).

These requirements can not be satisfied independently however, as fixing any two of them automatically fixes the values of the others. In each case, a particular value of core permeability is required for the proper fulfillment of the conditions. Appendix I lists the formulae essential to the determination of these factors. Although these formulae imply an entire freedom of choice of core permeability and size, it becomes necessary for practical purposes to standardize on a limited number of values of permeability and a limited number of sizes of core. It is possible by the proper choice from these types to approach rather closely to an ideal solution for each problem.

Improved Designs of Loading Coils

A study of alternate ways of utilizing the advantages offered by the new material in coils for voice frequency loaded cables showed that the greatest immediate benefit to the telephone plant would accrue from making the new coils substantially duplicate performance characteristics of previous designs. The new designs chosen are in fact better

in most respects and have been made approximately 50 per cent smaller in volume by using a molybdenum-permalloy core with a nominal permeability of 125. Table II summarizes data comparing

TABLE II

COMPARATIVE SIZE AND WEIGHT DATA OF TYPICAL NEW AND SUPERSEDED COILS

Type of Coil	Type of Compressed Powdered Core	Inductance (Henrys)	Coil Volume (Cu. In.)	Coil Weight (Lbs.)
Small Exchange Area	Permalloy	0.088	2.5	0.4
"	Molybdenum Permalloy	0.088	1.5	0.2
Program Circuit	Permalloy	0.022	11.8	1.6
"	Molybdenum Permalloy	0.022	4.4	0.6
Toll-Side Circuit	Permalloy	0.088	13.5	1.7
"	Molybdenum Permalloy	0.088	5.1	0.7

the electrical characteristics, sizes, and weights of coils commonly used on exchange and toll cables. Figure 8 shows the improved

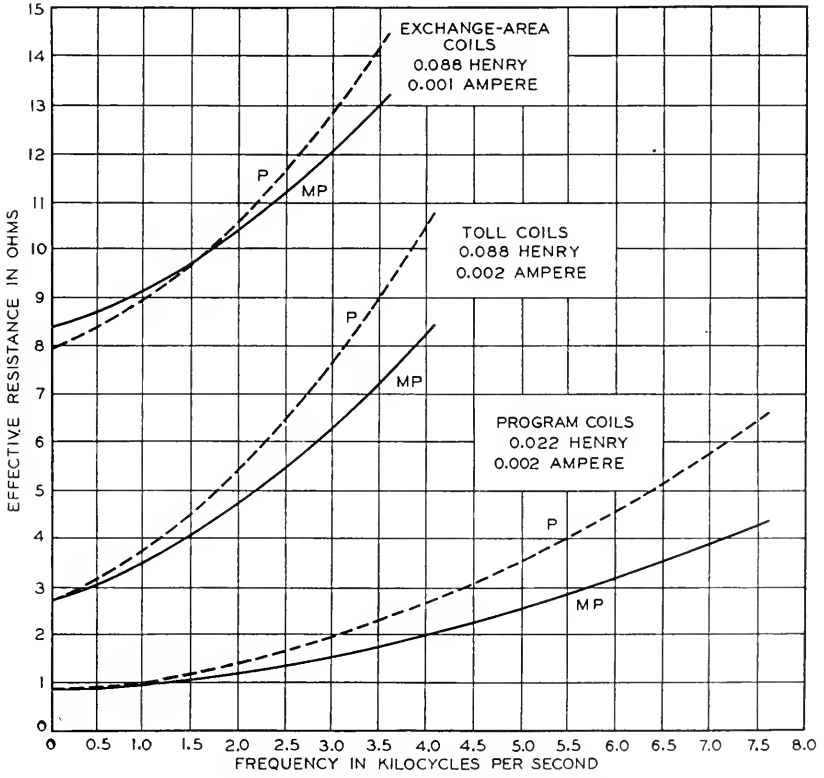


Fig. 8—Effective resistance-frequency characteristics of typical loading coils.

resistance-frequency characteristics for typical coils. Figure 9 shows the improved telegraph flutter [18] characteristics of a commonly used toll type coil.

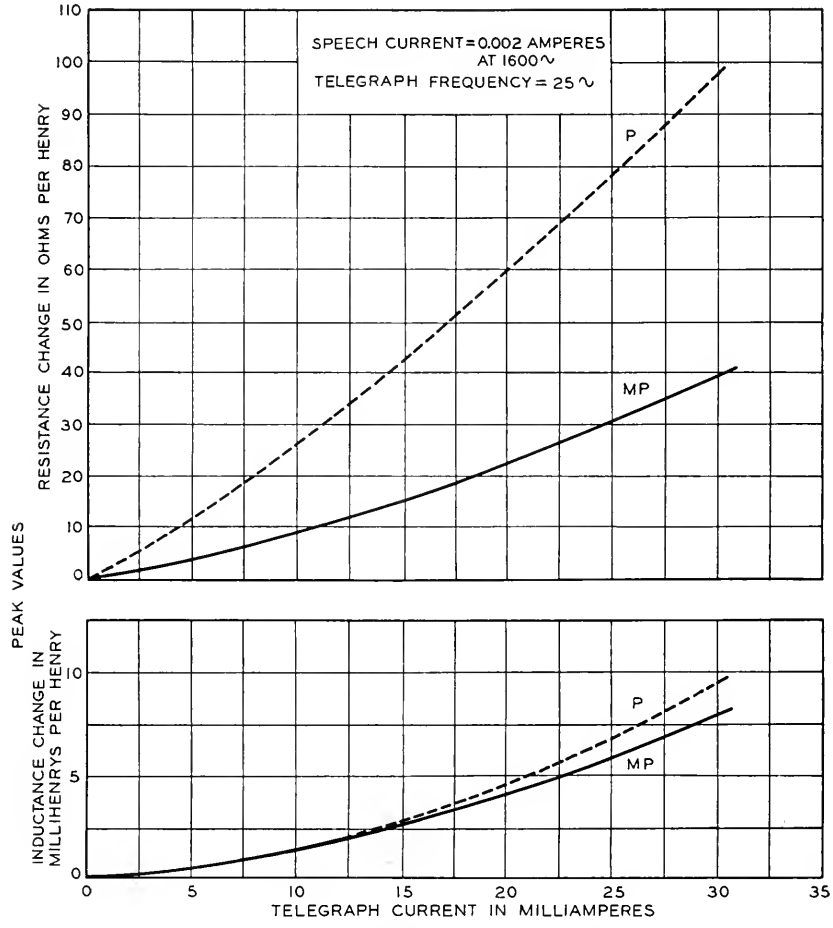


Fig. 9—Flutter characteristics of typical toll loading coil; P—with permalloy core, MP—with molybdenum-permalloy core.

Figure 10 pictures the reduction in size of cores and coils which are commonly used in toll and exchange area circuits. In the preparation for commercial manufacture of the smallest of the new coils, a difficult problem in the development of winding machinery was involved because of the small dimensions of the hole in the finished coil. This problem has been successfully solved by the Western Electric Company.

Fig. 10—Comparative sizes of the new molybdenum-permalloy (A, A') and the superseded permalloy (B, B') cores and coils: left—program circuit loading coils; right—exchange area loading coils.

Aside from manufacturing economies, the reduction in coil size is of importance to the Bell System from the standpoints of plant construction and installation. The size reduction is particularly important in instances where only a few coils are required at a given point, as is the case in program circuit loading. It is now practicable to enclose as many as six coils, even of the larger size employed for loading program circuits, directly in the cable splice at a loading point. This dispenses with the need for conventional cases, and reduces both manufacturing and installation costs. In the field of small complements using the inexpensive lead sheath type of case construction, it is now possible to furnish as many as 100 of the small exchange area coils whereas 15 was the maximum number accommodated with the superseded coil design. Table III gives comparative weights and volumes of typical cases provided for potting exchange area and toll type coils.

TABLE III

COMPARATIVE DATA ON REPRESENTATIVE CASES FOR NEW AND SUPERSEDED LOADING COILS

Type Cable	Size of Complement	Type Coil or Unit	Approx. Volume Cu. Ft.	Approx. Weight Lb.
Exchange Area	200	Permalloy	1.7	480
"	"	Molybdenum-Permalloy	0.8	350
Program-Toll	6	Permalloy	0.15	70
"	"	Molybdenum-Permalloy	0.11	50
Toll	50 units*	Permalloy	5.5	1350
"	"	Molybdenum-Permalloy	3.5	950

* A unit consists of one phantom and two side circuit coils.

Figures 11 and 12 show the comparative sizes of typical steel and lead sleeve type cases for the new and superseded coils. In Fig. 13, the midget proportions of the latest design of case for plotting 100 exchange area coils are contrasted with those of the case utilized up to 1922 containing only 98 exchange area coils. The reduction in size of cases for this type of coil is of particular importance in larger cities where underground vault space is at a premium.

IMPROVED INDUCTANCE COILS FOR FILTERS

The trend of development of toll transmission circuits is now very definitely toward multiple channel carrier systems utilizing a much wider frequency band than has heretofore been employed on wire circuits.[19,20] These systems involve extensive use of selective or equalizing networks at terminals and at repeater stations in order to

obtain proper separation of the frequency bands of the various channels or to insure suitable transmission properties of the individual channels. While these networks involve coils, condensers and crystals, it is frequently the case that their size, cost and performance are determined chiefly by the quality factor Q of the inductance coils. This follows from the fact that Q values of coils are usually considerably

Fig. 11—New and superseded cases containing 200 exchange area coils.

lower than those obtainable readily in condensers and crystals. Accordingly, it is very desirable to have as high a value of Q as possible economically. In addition, such coils must have low hysteresis resistance to limit modulation, and a low temperature coefficient of inductance to secure stability of attenuation or impedance characteristics of the filters and networks.

Due to the improvements in these respects, molybdenum-permalloy core coils can be used quite extensively in new types of carrier tele-

phone systems In such systems for existing lines and cables, as well as for projected new types of cables, those filters are of key importance which separate individual message channels in the frequency range from 3 to 108 kc. By using coils of powdered molybdenum-permalloy insulated to permeabilities of 14 or 26, valuable economies in space and cost of filters are realized.[21,22] Figure 14 shows a typical coil employing a 14 permeability core designed for use in one of these channel filters having its transmitted band in the vicinity of 108 kc,

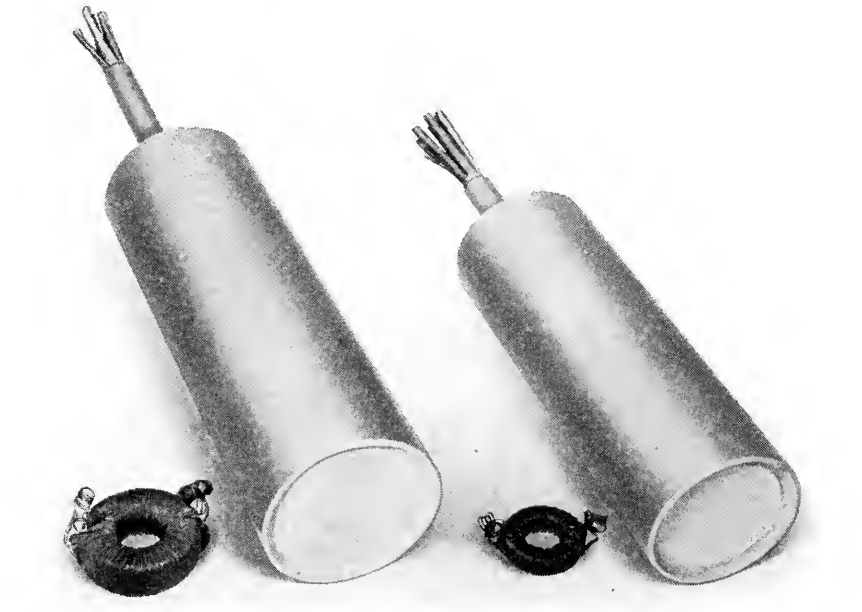


Fig. 12—New and superseded cases containing six program loading coils.

together with a shielded solenoidal air core coil which might be employed for the same purpose. The molybdenum-permalloy coil has a Q at 100 kc about twice that of the air core coil, yet it occupies approximately 1/10 as much space. The third order modulation products are approximately 80 db below the level of the normal channel currents. This is considered to be tolerable from the standpoint of interchannel crosstalk on circuits used for one-way transmission. An inductance-temperature coefficient of about -20×10^{-6} per degree F. has been chosen to compensate for the positive capacity-temperature coefficient of associated condensers.

In Fig. 15 data are presented illustrating the Q-frequency characteristics that can be obtained on typical coil designs using the new material in the frequency range from 300 cycles to 200 kc. The characteristics shown apply to coils wound on three sizes of cores that

Fig. 13—Comparative size of equivalent 1939 and 1922 cases.

are suitable for use in this range. For comparison, similar characteristics are also included for coils using cores of equal size but made of permalloy and electrolytic iron powder. These data include all effects on Q resulting from winding capacities and losses, which have

been made tolerably small by suitable choice of insulating materials, stranding of conductor and configuration of winding.

Conclusion

Compressed powdered cores of 2–81 molybdenum-permalloy have properties which are superior to those of earlier powdered cores in respect to permeability range, hysteresis loss and eddy current loss. Because of the lower losses and greater permeability range, inductance coils are now possible which have greatly increased Q values for a given volume. Because of the low hysteresis losses and attendant lowering

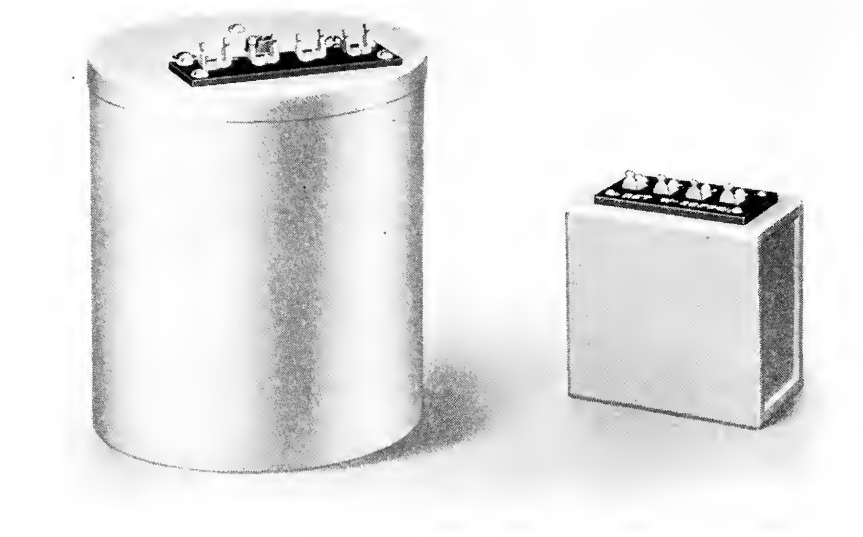


Fig. 14—Comparative size of non-magnetic core and molybdenum permalloy core filter coils.

of modulation effects to tolerable levels, magnetic core inductance coils can now be employed where non-magnetic core coils have previously been necessary, with a very great increase in Q values for a given volume. Temperature coefficients can now be obtained which are equal to the temperature coefficients of other high grade electrical elements such as mica condensers and quartz crystals. Moreover, the ability to make the temperature coefficient of coils negative or positive at will permits the attainment of remarkable stability in resonant combinations of coils and condensers. Two important applications have been made in the field of communication apparatus. For voice-frequency circuits, new loading coils of improved quality and reduced

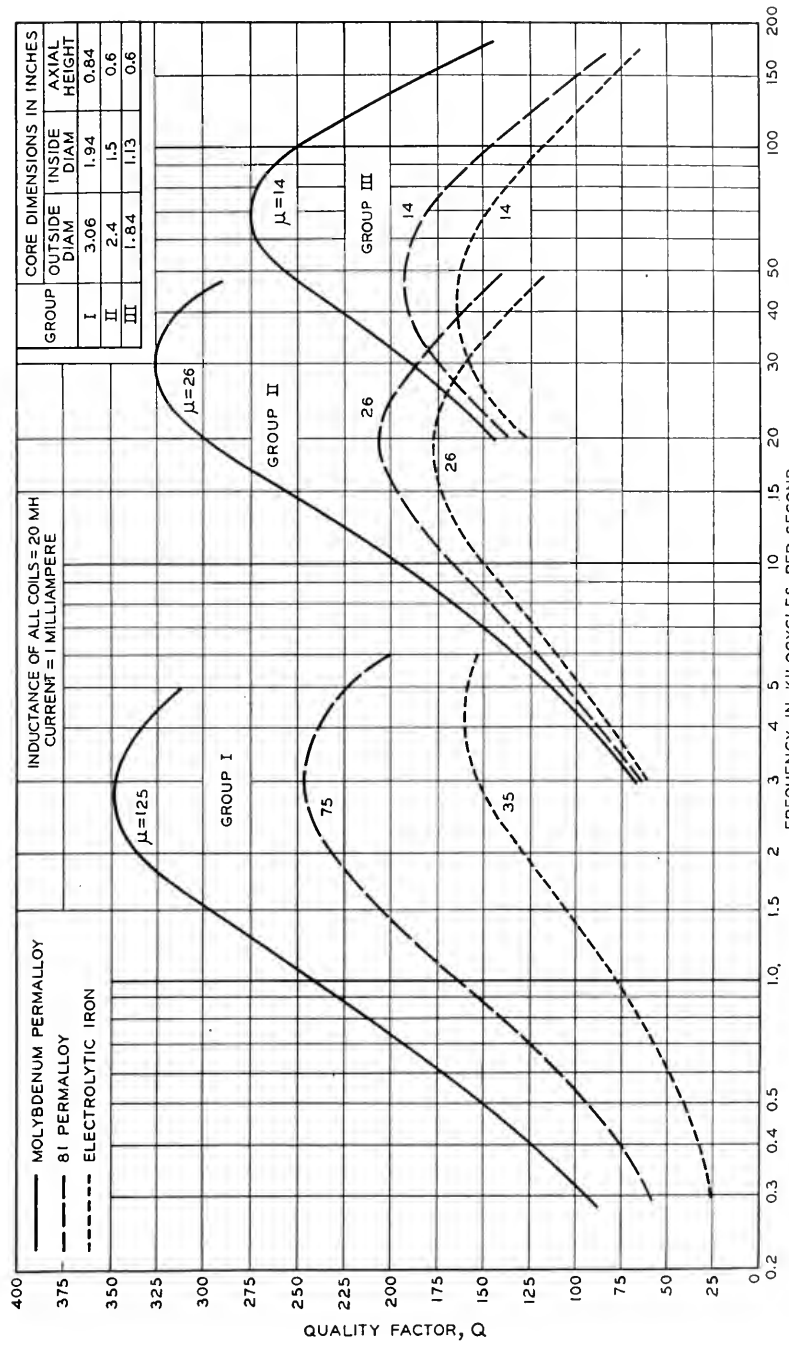


Fig. 15—Comparative Q-frequency characteristics on typical filter coils using new and superseded materials.

size have been standardized. For broad band carrier systems, the compactness of channel separating filters is largely due to the volume economies introduced by molybdenum-permalloy coils.

APPENDIX

The following formulae illustrate the essential steps in selecting the size and permeability of an annular core best suited for a coil having a desired set of characteristics. For sake of simplicity, they assume throughout a coil of arbitrarily chosen proportions, in which all dimensions bear fixed ratios to the mean core diameter, and approximating those that are practicable from a manufacturing standpoint. The core is assumed to be rectangular in section. It is assumed that wire of any diameter can be used and that the winding efficiency is independent of the wire diameter. Any inductance due to air space outside of the core is neglected. Because of these simplifications, the expressions are of somewhat restricted applicability. However, they yield solutions for optimum permeability and corresponding values of Q and core size which are sufficiently accurate for most practical purposes.

The inductance in henrys due to a core of permeability μ, and mean diameter d cm., wound with N turns of wire is

$$L = \frac{3}{8} N^2 \mu d \times 10^{-9}. \tag{1}$$

If the coil is wound with wire of resistivity ρ_c ohm-cm., with winding efficiency s (i.e., the ratio of copper area to total available winding area), the direct current resistance in ohms will be

$$R_c = \frac{19 \rho_c L \times 10^9}{s \mu d^2}. \tag{2}$$

The maximum flux density due to sine wave measuring current of effective value I amperes is

$$B_m = \frac{8I}{5} \sqrt{\frac{\mu L \times 10^9}{3d^3}}. \tag{3}$$

In terms of the hysteresis loss coefficient $k_2 = \mu a$, the modulation factor thus becomes [17]

$$m = \frac{E_3}{E_1} = \frac{3 k_2 B_m}{10\pi} = \frac{4 k_2 I}{25\pi} \sqrt{\frac{3\mu L \times 10^9}{d^3}} \tag{4}$$

or

$$d^3 = 3\mu L \times 10^9 \left(\frac{4 k_2 I}{25\pi m} \right)^2. \tag{5}$$

When the core permeability μ is reduced by dilution of a given material, the hysteresis and residual loss coefficients a and c vary so as to make the products $\mu c = k_1$ and $\mu a = k_2$ approximately constant, as may be seen by reference to Table I. The core loss resistance in ohms at frequency f cycles per second may therefore be expressed with reasonable accuracy as

$$(6) \qquad R_m = Lf(k_1 + k_2 B_m + \mu ef) = Lf\left(k_1 + \frac{10m\pi}{3} + \mu ef\right),$$

where the eddy current coefficient e depends upon the particle diameter t, and the alloy resistivity ρ being proportional [13] to t^2/ρ.

The coil quality factor is thus

$$(7) \qquad Q = \frac{2\pi fL}{R_c + R_m} = \frac{2\pi f}{\dfrac{19\rho_c \times 10^9}{s\mu d^2} + f\left(k_1 + \dfrac{10m\pi}{3} + \mu ef\right)}.$$

Case I: If the value of m is fixed and d and R_c can be freely chosen, it is desirable to know the value of μ which will yield the highest possible value of Q. By substituting in (7) the value of d^2 obtained from (5) and setting the derivative with respect to μ equal to zero, the following is obtained for the optimum permeability:

$$(8) \qquad (\mu')^8 = \frac{52.5 \times 10^{16}\rho_c{}^3 m^4}{e^3 s^3 f^6 k_2{}^4 I^4 L^2}.$$

The corresponding values of d and R_c can be obtained from equations (5) and (2). The corresponding value of Q, which is the greatest obtainable under these conditions is

$$(9) \qquad Q'_{\max} = \frac{\pi}{\dfrac{5m\pi}{3} + \dfrac{k_1}{2} + \dfrac{4}{5}\mu' ef}.$$

If a smaller value of Q than that obtained from (9) is acceptable, equations (7) and (5) can be solved simultaneously for d and μ. A smaller value of μ than that obtained from (8) and a correspondingly smaller value of d will result.

Case II: If modulation is unimportant and the hysteresis loss resistance is negligible in comparison with other component losses, then d and μ can be selected without regard to modulation. Equation (7) can be differentiated directly and solved for the permeability required to yield the maximum value of Q. This optimum permeability is

$$(10) \qquad (\mu'')^2 = \frac{19\rho_c \times 10^9}{sed^2 f^2}.$$

The corresponding value of Q is

$$Q''_{\max} = \frac{2\pi}{k_1 + 2\mu''ef}. \tag{11}$$

The values of $Q''_{\max}$ and μ'' depend on the value of d chosen. For a desired value of Q, (11) can be solved for μ'', and (10) for the corresponding diameter.

In any case, the ideal wire size has a cross-sectional area of conductor equal to

$$0.15sd^2\sqrt{\frac{\mu d \times 10^{-9}}{L}} \text{ cm}^2.$$

It is usually desirable to subdivide the wire into insulated strands to minimize eddy current losses in the coil winding.

References

1. "Commercial Loading of Telephone Circuits in the Bell System," B. Gherardi, *Trans. A.I.E.E.*, v. 30, 1911, p. 1743.
2. "Development and Application of Loading for Telephone Circuits," T. Shaw and W. Fondiller, *Jour. A.I.E.E.*, v. 45, 1926, p. 253; *B.S.T.J.*, v. 5, 1926, p. 221.
3. "Magnetic Properties of Compressed Powdered Iron," B. Speed and G. W. Elmen, *Trans. A.I.E.E.*, v. 40, 1921, p. 596.
4. "Compressed Powdered Permalloy Manufacture and Magnetic Properties," W. J. Shackelton and I. G. Barber, *Trans. A.I.E.E.*, v. 47, 1928, p. 429.
5. A. F. Bandur, U. S. Patent 1,673,790, June 19, 1928.
6. "Pupinspulen mit Kernen aus Isoperm-Blech oder -Band," H. Jordan, T. Volk and R. Goldschmidt, Europaischer Fernsprechdienst v. 31, 1933, p. 8.
7. "Permalloy, an Alloy of Remarkable Magnetic Properties," H. D. Arnold and G. W. Elmen, *Jour. Frank. Inst.*, v. 195, 1923, p. 621.
8. "Carrier Current Telephony and Telegraphy," E. H. Colpitts and O. B. Blackwell, *Trans. A.I.E.E.*, v. 40, 1921, p. 205.
9. "Radio Extension of the Telephone System to Ships at Sea," H. W. Nichols and L. Espenschied, *Proc. I.R.E.*, v. 11, 1923, p. 193; *B.S.T.J.*, v. 2, 1923, p. 141.
10. "Magnetic Alloys of Iron, Nickel, and Cobalt," G. W. Elmen, *Jour. Frank. Inst.*, v. 207, 1929, p. 583; *B.S.T.J.*, v. 8, 1929, p. 435.
11. "A Survey of Magnetic Materials in Relation to Structure," W. C. Ellis and E. E. Schumacher, *B.S.T.J.*, v. 14, 1935, p. 8.
12. "Statistical Description of the Size Properties of Non-uniform Particulate Substances," T. Hatch and S. Choate, *Jour. Frank. Inst.*, v. 207, 1929, p. 369.
13. "On the Self-induction of Wires," O. Heaviside, *Phil. Mag.*, v. 23, 1887, p. 173.
14. "Magnetostatik der Massekerne," F. Ollendorf, *Arch. f. Elektrotechn*, v. 25, 1931, p. 436.
15. "Iron Powder Compound Cores for Coils," G. W. O. Howe, *Wireless Engineer*, v. 10, 1933, p. 1.
16. "Magnetic Measurements at Low Flux Densities Using the A-C. Bridge," V. E. Legg, *B.S.T.J.*, v. 15, 1936, p. 39.
17. "Harmonic Production in Ferromagnetic Materials at Low Frequencies and Low Flux Densities," E. Peterson, *B.S.T.J.*, v. 7, 1928, p. 762.
18. "Hysteresis Effects with Varying Superposed Magnetizing Forces," W. Fondiller and W. H. Martin, *Trans. A.I.E.E.*, v. 40, 1921, p. 553.
19. "Communication by Carrier in Cable," B. W. Kendall and A. B. Clark, *Elec. Engg.*, v. 52, 1933, p. 477; *B.S.T.J.*, v. 12, 1933, p. 251.
20. "Systems for Wide Band Transmission over Coaxial Lines," L. Espenschied and M. E. Strieby, *Elect. Engg.*, v. 53, 1934, p. 1371; *B.S.T.J.*, v. 13, 1934, p. 654.
21. "The Evolution of the Crystal Wave Filter," O. E. Buckley, *Jour. Appl. Phys.*, v. 8, 1937, p. 40; *Bell Tel. Quarterly*, v. 16, 1937, p. 25.
22. "An Improved Three-Channel Carrier Telephone System," J. T. O'Leary, E. C. Blessing and J. W. Beyer, *B.S.T.J.*, v. 18, 1939, p. 49.

High Accuracy Heterodyne Oscillators

By T. SLONCZEWSKI

The accuracy of a heterodyne oscillator after the low frequency check is made is of the same order of magnitude as that of an ordinary type of oscillator in which circuit elements of the same stability are used. It depends on the constants of the variable frequency oscillator only. This accuracy can be improved by a ratio of 10 to 1 by adding another and higher check frequency. The temperature coefficient of the circuit elements can be kept down to less than 6 parts per million. Scale errors can be reduced to a value comparable with the oscillator accuracy by spreading the scale. A precision oscillator having a frequency range up to 150 kc. and an accuracy of ± 25 cycles including a scale mechanism whereby a large scale spread is obtained on a direct reading scale is described.

Introduction

THE output frequency of a heterodyne oscillator is obtained by modulating the outputs of two oscillators of appreciably higher frequency, one of the oscillators having a fixed frequency, the other being continuously variable over a band width equal to the required output frequency range.

The circuit consists essentially of the two so-called local oscillators, the modulator, where the difference frequency is generated, and an amplifier where the modulator output is raised to the desired level.

The earliest designs of heterodyne oscillator were confined to the audio frequency range, but recently carrier-frequency applications have become more numerous. As the frequency range of the oscillators has increased, their per cent accuracy requirement has increased also. The required frequency accuracy of the oscillator is determined by the maximum slope of the frequency characteristic of the apparatus being measured. If this slope is great, as in the case of a sharply tuned circuit a relatively small displacement of the frequency will result in a large error in the value to be measured. In carrier-frequency systems where the signal is displaced upwards in the frequency scale by modulation, each channel has to meet same crosstalk and transmission requirements independent of its location in the carrier band. Therefore, the maximum slope of the characteristics is independent of the frequency and an oscillator used for measuring purposes has to meet a constant frequency error requirement. In addition the accur-

acy required when expressed in cycles is comparable with that of audio-frequency oscillators so that the percentage accuracy must be much higher.

The advantages of the heterodyne oscillator have made it desirable to study its sources of error to determine whether such an oscillator can be designed to have sufficient accuracy for these applications.

Oscillators With a Single Frequency Check

The frequency of a heterodyne oscillator is given by the expression:

$$F = f' - f, \tag{1}$$

where we will assume f' to be constant and f to be variable and less than f' whence the frequency of the variable frequency oscillator is lowered as the output frequency of the heterodyne oscillator is raised.

The value of F is usually much smaller than either f' or f and relatively small frequency shifts in the local oscillators produced by aging and temperature effects upon the elements of their resonant circuits and changes in vacuum tubes and in the stray capacitances of the circuits produce large relative variations in the output frequency. Usually the stability required of F and the ratio f'/F are so high that it is impracticable to design local oscillators of sufficient stability to meet requirements. Instead, in all heterodyne oscillators an adjustment in the form of a padding condenser in the circuit of the fixed frequency oscillator is used, whereby its frequency is adjusted shortly before the measurement until the oscillator reads correctly at the bottom of its frequency range. The adjustment is made by the zero beat method or by comparison with a low-frequency standard such as a vibrating reed or the 60-cycle power supply.

At the time of the adjustment the frequency of the oscillator is

$$F_o = f' - f_o, \tag{2}$$

where f_o is the value of f at the check frequency F_o. Eliminating f' between (1) and (2) we obtain

$$F = F_o + (f_o - f). \tag{3}$$

The frequency of the variable oscillator may be expressed as

$$f = 1/(2\pi\sqrt{L(C_o + C_a)}), \tag{4}$$

where C_a is the change in the variable air condenser capacitance from the value it has at f_o, and C_o is essentially the value of the fixed con-

denser, usually a good mica unit. L is the inductance of the resonant circuit.

Combining (3) and (4) we get

$$F = F_o + 1/(2\pi\sqrt{LC_o}) - 1/(2\pi\sqrt{L(C_o + C_a)}). \tag{5}$$

The accuracy of the oscillator will depend on the variations in the values of F_o, L, C_o and C_a and is independent of the constants of the fixed frequency oscillator.

By giving increments ΔF_o, ΔC_o, ΔC_a and ΔL to the constants F_o, C_o, C_a and L we obtain after simplifying the expressions

$$\Delta F_{Fo} = \Delta F_o, \tag{6}$$

$$\Delta F_{Co} = -\frac{\Delta C_o}{2C_o} f_o \left[1 - \left(1 + \frac{F_o}{f_o} - \frac{F}{f_o} \right)^3 \right], \tag{7}$$

$$\Delta F_{Ca} = \frac{\Delta C_a}{2C_a} f_o \left[1 + \frac{F_o}{f_o} - \frac{F}{f_o} \right] \left[1 - \left(1 + \frac{F_o}{f_o} - \frac{F}{f_o} \right)^2 \right], \tag{8}$$

$$\Delta F_L = -\frac{\Delta F}{2L} F \left(1 - \frac{F_o}{F} \right) = -\frac{\Delta L}{2L} (F - F_o), \tag{9}$$

giving the corresponding frequency errors ΔF where $f_o = 1/(2\pi\sqrt{LC_o})$ is the variable oscillator frequency at the check frequency F_o.

A variation in F_o will produce an error constant over the whole frequency range. On Fig. 1 the other errors are found plotted in parametric form. To find the error ΔF corresponding to a frequency F the ordinate y corresponding to the value of $x = (F - F_o)/(f_o)$ should be found. Then

$$\Delta F_{Ca} = y_{Ca} f_o \Delta C_a / C_a; \qquad \Delta F_{Co} = y_{Co} \Delta C_o / C_o; \qquad \Delta F_L = y_L f_o \Delta L / L.$$

It is found that F_o can be neglected in all practical cases. The ratio of the ordinate to the abscissa gives the percentage error in frequency caused by a one per cent variation in the element involved.

An examination of the curves shows that they differ only slightly from straight lines which can be interpreted as meaning that the errors are fairly independent of the choice of f_o. This constant should be chosen therefore sufficiently low to require infrequent adjustment at the low-frequency end of the scale. For low values of f_o such that $x > .3$ difficulties in shaping of the air condenser plates and in designing the modulator filter begin to appear. If the errors in an ordinary type of oscillator due to capacitance and inductance variations were plotted on the same set of coordinates the curves would coincide with the line

y_L. This means that if elements of the same accuracy were used, the heterodyne oscillator would be somewhat more accurate. Its total error would be represented by $y_{ca} + y_{co} + y_L$. Since $\Delta C_a/C_a$ and $\Delta C_o/C_o$ will be both positive and of about the same order of magnitude partial compensation will obtain and the error will be of the order of

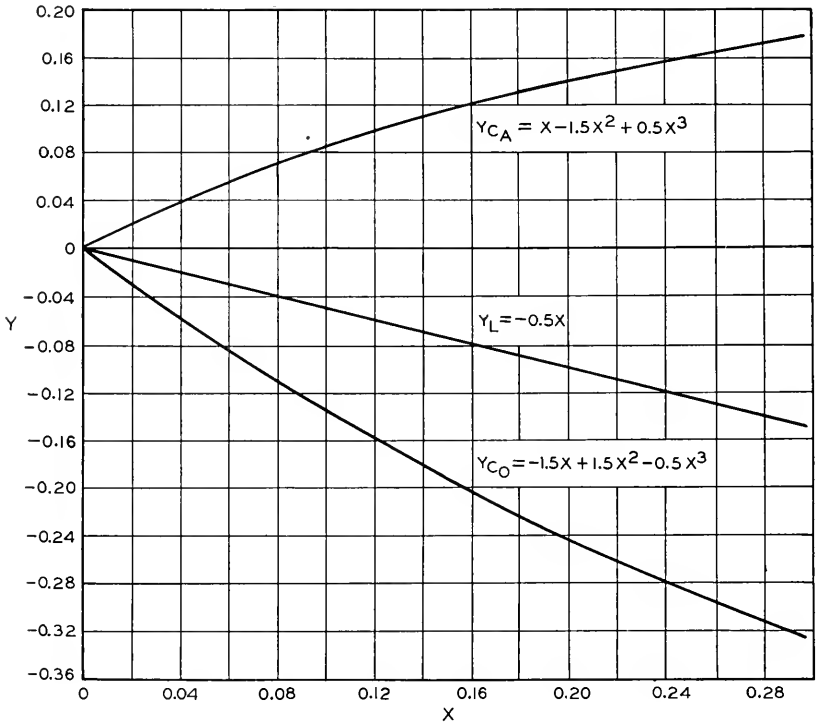


Fig. 1—The frequency errors in a heterodyne oscillator at a frequency $F = xf_0 + F_0$ after the low frequency check had been made can be obtained from the plot as follows: For a variation ΔC_0 in the fixed capacitance C_0, $\Delta F_{C_o} = y_{C_o} \frac{\Delta C_0}{C_0} f_0$; for a variation in the air condenser capacitance C_a, $\Delta F_{C_a} = y_{C_a} \frac{\Delta C_a}{C_a} f_0$; for a variation in the inductance L, $\Delta F_L = y_L \frac{\Delta L}{L} f_0$.

magnitude of ΔF_L. In the case of the ordinary type of oscillator the errors due to the capacitance and inductance variations will be equal and of the same sign so that the error will be of the order of magnitude of $2\Delta F_L$. For audio frequency applications this accuracy has been found to be adequate given sufficient care in the construction of the circuit elements.

Oscillators With a Double Frequency Check

For carrier frequency applications the tolerable error takes a constant value over the entire frequency range and it is found that if a single frequency check is used it is not possible to obtain sufficiently stable elements to maintain the required accuracy at points on the scale removed from the check frequency.

An increase in the accuracy of heterodyne oscillators has been obtained, however, by adding an adjustable condenser to C_o and checking the oscillator at two frequencies, the low frequency F_o and at another, higher, frequency F_s. Adjustment of this condenser by ΔC_o introduces a frequency change—$y_{C_o} f_o \Delta C_o / 2C_o$ adjustable in sign and magnitude and this can be made to cancel the error $\Delta F_{C_a} + \Delta F_L$ for at least one frequency, the check frequency F_s. Obviously if the adjustment is made to correct for variations in C_o no residual error remains. The residual errors which remain after correcting for ΔF_{C_a} and ΔF_L are shown on Fig. 2. The residuals of ΔF_{C_a} and ΔF_L differ from each other

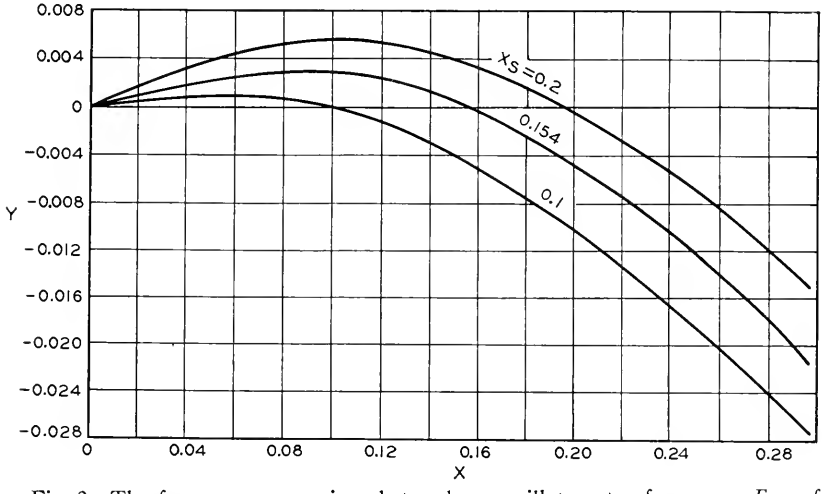


Fig. 2—The frequency errors in a heterodyne oscillator at a frequency $F = xf_0 + F_0$ after the low and high frequency checks had been made can be obtained from the plot as follows: For a variation in the air condenser capacitance C_a, $\Delta F_{C_a} = y \frac{\Delta C_a}{C_a} f_0$; for a variation in the inductance L, $\Delta F_L = y \frac{\Delta L}{L} f_0$.

so little that only one set of curves was drawn. The values of y were obtained by forming the sum $y = K y_{C_o} + y_{C_a}$ and choosing K so that $y = o$ for $x_s = (F_s - F_o)/f_o$.

For $x_s = .1$ better compensation is obtained at the lower end than at the higher. For a very wide frequency range up to $x = .25$ the best

check frequency would be $x_s = .2$. A good practical limit to x is at 1.9 and here a value of x_s around .15 is best. A further improvement of about 50 per cent could be obtained by choosing a higher value of F_o. When comparing Fig. 2 with Fig. 1 it should be borne in mind that the scale spread for y on Fig. 2 is ten times that of the Fig. 1 which shows that an improvement in accuracy of at least ten to one is obtained by the adjustment. This means, that given two frequency standards F_o and F_s of sufficient accuracy a heterodyne oscillator can be built having a much higher accuracy than an ordinary oscillator having the same frequency range and same quality of circuit elements. This is somewhat contrary to what we are accustomed to think.

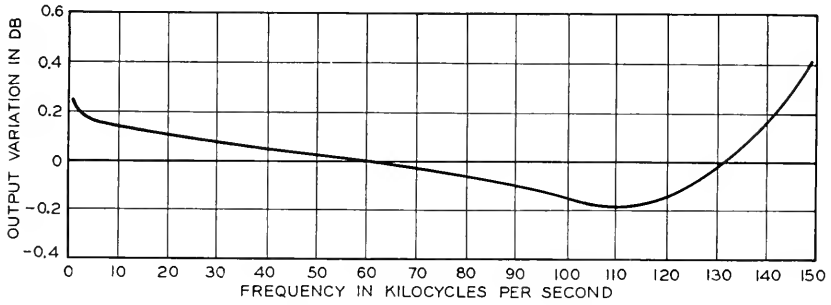


Fig. 3—Frequency-output characteristic.

One detail involved in the procedure of checking the oscillator which permits this high degree of accuracy to be obtained needs elaboration. As C_o is varied during the adjustment by the amount ΔC_o the value of f_o is changed and this destroys the low frequency adjustment at F_o. It is possible to obtain the adjustment by a process of successive approximations but the procedure is tedious. The difficulty can be overcome by the use of a mechanical device, however, as follows. The condenser ΔC_o is ganged to another condenser in the resonant circuit of the fixed oscillator, and the two condensers are so proportioned that the change in the fixed oscillator frequency is equal to the change in f_o as the condenser is adjusted. This makes the low frequency adjustment independent of the high frequency one. The oscillator is just set to the required reading at F_s and ΔC_o is adjusted until the frequency value is correct. Theoretically instead of two condensers two coupled inductometers could have been used to adjust the inductances in the resonant circuits. The net result obtained would have been the same and the mathematical treatment would be like the one given above. Condensers lend themselves better to such construction, however.

Stability of the Constants

Having determined the oscillator errors from the variations in its constants it will be of interest to inquire how large these may be.

When the zero beat method is used the error ΔF_{Fo} will depend on the value of the lowest beat frequency at which the local oscillators can operate. With a reasonable amount of shielding and some precautions in order to avoid mutual inductance in wiring loops it is quite practicable to keep this error below one cycle with local oscillators as high as 200 kc. The beat frequency may be observed on an ammeter placed in the plate circuit of the modulator. When alternating current from the power mains is used as a standard the accuracy is better than one cycle.

There are now available external frequency standards against which the high frequency check could be made which have such high accuracy that the resulting error in the heterodyne oscillator can be entirely neglected. It is desirable, however, to make the oscillator independent of external sources for its adjustment. A convenient checking circuit consists of a quartz crystal which is thrown in with a key across the grids of the output amplifier. At the series resonance frequency of the crystal the loss introduced reduces the output so sharply that the minimum output can be observed within 3 cycles at 100 kc. At any other frequency the error is therefore 30 ppm (parts per million). By using properly cut crystals the temperature variation error is made negligible.

The variations in L are chiefly due to temperature variations. Ordinary potted coils having a large number of layers have temperature coefficients up to 20 parts per million per degree Fahrenheit. The variation is chiefly due to the expansion of the wire.

This error is tolerable in audio frequency oscillators for most purposes. For carrier frequency oscillators unpotted coils having a single layer bank winding wound on a phenol plastic form may be used. Here the lengthwise expansion of the form, which tends to decrease the inductance partly compensates for the expansion of the winding which tends to increase the inductance. Coefficients from 0 to + 6 ppm per °F. are obtained.

The capacitance C_a in commercial air condensers has temperature coefficients of up to 25 ppm per °F. This, again, gives sufficient accuracy for audio frequency oscillators but is not satisfactory for carrier applications. The variations in capacitance with temperature are produced by increase in area of the plates with their expansion which increases by an amount equal to twice the linear coefficient of expansion of the material used. This change is partly compensated by the length-

ening of the air-gaps. When, as usual, several materials are used in the construction, bending of the stator plates due to strains introduced by unequal expansions of the members produce unpredictable changes in capacitance. This is particularly true in the most common construction where the stators are held in place by rods of insulating material. The insulator having a different temperature coefficient of expansion than the plates, the difference in the expansion causes the plates to buckle.

Better stability can be obtained in a condenser built as follows. All parts determining the length of the condenser, including the stator supports and the stator plates are of aluminum. The ends of the stator supports are held in place by insulating bushings of sufficiently small dimensions to make the difference in expansion negligible. The bushings are made of Alsimag, a ceramic material which has a small dielectric constant and coefficient of dielectric constant.

With such a construction, the temperature coefficient of the condenser is equal to twice the temperature coefficient of expansion of the material of which the rotor plates are made minus the temperature coefficient of linear expansion of aluminum determining the length of the air-gaps. One half of the rotor plates are made of invar and one half of aluminum. The average expansion of the area of the rotor plates equals then the temperature coefficient of linear expansion of aluminum and the temperature coefficient of capacitance of the condenser should be equal to the temperature coefficient of air dielectric constant which is about 1 ppm per °F. negative. Measurements show that the temperature coefficient of the condenser varies from -3 to $+4$ ppm per °F., a quite acceptable value. The capacitance change due to a variation in the atmospheric pressure of one inch, a large variation, is 20 ppm.

Temperature coefficients, of paraffined mica condensers, can be adjusted by special manufacturing methods to 10 ppm negative. For the sake of increasing the instantaneous stability the two condensers used in each oscillator are paired within 3 ppm. As mentioned before, no residual error due to ΔC_o remains after the frequency check is made. The low temperature coefficients are desirable only to improve the stability of the oscillator.

By using high Q circuits and suitable corrective reactances, the variations in the frequency due to power line variations may be readily kept smaller than any one of the other errors discussed above.

Scale Errors

A heterodyne oscillator cannot be classified as a purely electrical circuit for it is used to translate a mechanical coordinate, the scale

setting, into an electric coordinate, the output frequency. In planning the oscillator design, therefore, it is necessary to give as much attention to the construction of the scale as to the construction of the circuit elements.

For maximum scale length economy the scale should be so subdivided that a frequency interval equal to the tolerable frequency error ΔF could be read. The scale interval Δl corresponding to this frequency interval, will vary with the measuring conditions. For well illuminated scales on panel mounted equipment to be read conveniently at arm's length an interval Δl of at least .05″ is needed. For portable apparatus, intervals as small as .02″ have been used. With the aid of a vernier it can be brought down to .001″. Scale spreads such that a frequency interval much smaller than ΔF can be read are not only uneconomical but are also objectionable because they encourage the use of the instrument beyond its accuracy limits.

Having chosen Δl and the frequency error ΔF at all points of the scale, the scale shape $l = f(F)$ can be determined by the approximation

$$l = \int_0^F \frac{\Delta l}{\Delta F}\, dF.$$

As an example, in audio frequency applications the most common form of frequency accuracy desired is that having a constant percentage value $\Delta F/F = \rho$ at the upper part of the scale. At lower frequencies this accuracy is higher than necessary and the requirement is changed to a constant ΔF_o. A smooth shape is obtained by making the transition point F_T at such a frequency that $\Delta F_o/F_T = \rho$. The scale shape is then approximately

$$l = \int_0^F \frac{\Delta l}{\Delta F_o}\, dF = \frac{\Delta l}{\Delta F_o} F \qquad \text{for} \qquad F < F_T$$

and

$$l = \int_0^{F_T} \frac{\Delta l}{\Delta F_o}\, dF + \int_{F_T}^F \frac{\Delta l}{\rho F}\, dF = \frac{\Delta l}{\Delta F_o} F_T + \frac{\Delta l}{\rho} \log_e \frac{F}{F_T}, \quad \text{for} \quad F > F_T$$

The scale of common type of audio frequency oscillator can be spread over a ten inch dial giving a satisfactory accuracy.

For carrier applications, where the spread of any voice band is independent of its position in the frequency range the error function takes the form of a constant and the scale should be linear. Usually the scale lengths involved are much larger than in audio oscillators. To obtain sufficient scale length a precision worm and gear mechanism has to be used to drive the tuning condenser of the heterodyne oscilla-

tor. It gives a scale length of 300 inches, the equivalent of a 5-foot dial and can duplicate settings to better than one part in 10,000.

One detail of construction of such long scales deserves mention. Commercial worm driven air condensers carry on the worm shaft a drum or a dial on which fractions of a revolution of the worm shaft are recorded, while the number of revolutions is recorded on a main dial fixed on the rotor shaft. The effective scale length is then equal to the total displacement of the periphery of the small dial or drum. Using such a construction the oscillator has to be set by consulting a calibration chart where the position of the main dial and of the worm shaft is recorded against the oscillator frequency. Thus one of the most valuable properties of the short scale audio frequency oscillator, its direct reading, is lost.

To remedy this situation a special scale mechanism has been developed for carrier frequency oscillators which combines great scale length with good spread and compactness. It consists of a long motion picture film strip engaged by a two inch film sprocket mounted in place of the conventional drum on the worm shaft. The rotation of the shaft determines the displacement of the film against an index which reads the frequency directly in kilocycles. The loose ends of the strip are wound up on two spools interconnected by a spring mechanism which takes up the slack. The whole mechanism is confined in a space about 4″ by 5″ by 5″ accommodating a scale length up to 450 inches with a scale spread corresponding to $\Delta l = .05''$.

Specific Application

An example of application of these methods in the design of a heterodyne oscillator is furnished by an oscillator built for use in connection with the installation and maintenance of broad band transmission systems. It is shown on Fig. 4.

It has a frequency range of from 1 to 150 kc. Its variable frequency oscillator covers a range of from 500 to 650 kc. This was chosen as low as possible to obtain good instantaneous stability, but high enough not to introduce difficulties in designing the filter following the modulator. The capacitance of the air condenser is about 800 $\mu\mu$f. From the circuit design standpoint a larger capacitance would be desirable, but for the stability required the overall size of the condenser sets an upper limit to the capacitance. The frequency and air condenser capacitance values determine the value of the fixed condenser at 1000 mmf and the coil inductance at 50 microhenries. The fixed oscillator is similar to the variable oscillator except for the omission of the variable air condenser.

The frequency setting is recorded on a 300-inch film scale such as described above. This gives a spread of two inches per kilocycle. With the 50-cycle divisions marked directly the mechanism can be readily set to an accuracy better than 25 cycles. The visibility of the scale is greatly enhanced by a pilot lamp placed in back of the scale

Fig. 4—Front view of the oscillator.

window with an intervening opal glass. A crank on the front of the panel is used to set the oscillator, the range being covered in 47 revolutions. When changing the frequency setting even at a moderate rate the speed with which the film moves prevents the operator from observing the frequency setting. To make the adjustment more convenient, a coarse scale is recorded on a dial which can be read easily to one

kilocycle while the mechanism is in motion. It can be seen under the hood in the center of the panel.

Below the coarse frequency dial is seen a small dial connected to a variable condenser which permits the operator to vary the frequency of the oscillator up to ± 50 cycles from the frequency to which it is set and to read the frequency change with an accuracy of about 3 cycles. This feature is found to be useful in locating peaks of frequency characteristics of sharply resonant circuits.

The frequency checks are made by operating a key which throws the oscillator output across a telephone switchboard lamp and a 100 kc crystal across the grid of the output stage. For the low frequency check another key superposes the 60 cycles power main frequency on the oscillator output and a screwdriver adjustment operating a condenser in the fixed oscillator adjusts the oscillator frequency to synchronism with the scale set at 60 cycles. For the high frequency check a minimum signal is obtained on the lamp with the scale set to 100 kc by adjusting a padding condenser in the variable oscillator.

In the modulator a pentode type vacuum tube is used, which has a control grid-plate current characteristic which over nearly the entire region from zero bias to cut-off approaches a parabola so closely, that where modulation products lower than 40 db down on the useful output can be neglected, only first and second modulation products need be considered. The bias is placed in the middle of the parabolic range and the two input signals are adjusted to equality and to a value covering the entire parabolic range. This gives the maximum useful modulation output necessitating the smallest amount of gain in the output stage at little sacrifice in efficiency. The modulator being parabolic, the only products of modulation other than the useful output are the two high frequency input signals, their harmonics and sum frequencies. These are eliminated from the output by inserting a filter between the modulator and the output stage. Advantage is taken of phase discrimination since the circuit is arranged in push-pull to decrease the filter requirements for some of the products, which are generated in phase.

The plate supply is obtained from a rectifier operating on the 60-cycle main supply. It is provided with a vacuum tube regulator circuit which keeps the plate and screen voltages constant over a ± 5 volt variation of the power line voltage. The output control is obtained by means of a potentiometer in the output amplifier input. Two output impedances, 600 and 135 ohms, may be selected by operating a key.

The apparatus is mounted on a standard 19-inch panel 28 inches high. The bottom, the coolest part, is occupied by the oscillators; the middle

by the modulator and amplifier; and the top by the power pack. Perforations in the oscillator cover provide ventilation to reduce warming-up effects. A close-up giving the details of the scale mechanism and the shielding is shown on Fig. 5.

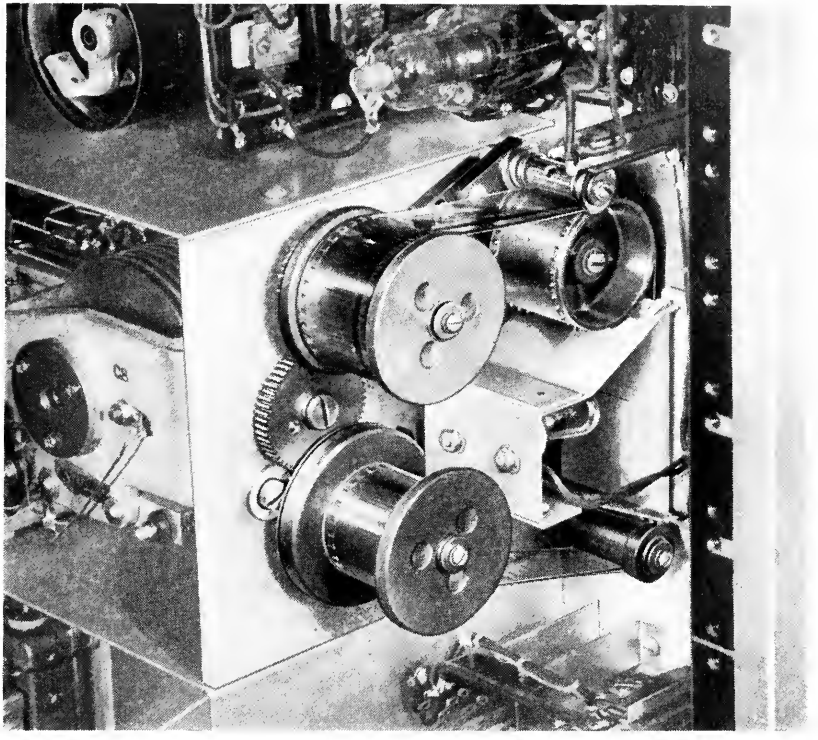


Fig. 5—Details of the scale mechanism.

Tests on the oscillator show that the overall frequency accuracy throughout its range can be maintained to ± 25 cycles. The harmonics are down 40 db from the fundamental at 100-milliwatt output. With the full output of one watt the harmonics are 30 db down. The total output variation with frequency is shown on Fig. 3.

This oscillator has found a wide range of applications as an accurate source of frequency in the communications field.

List of Symbols

F output frequency of the heterodyne oscillator

F_o standard frequency used to check the oscillator at the low end of the scale

F_s standard frequency used to check the oscillator at the high end of the scale

F_T frequency at which the scale changes from linear to logarithmic

l length of the scale interval from 0 to F

f frequency of the variable oscillator

f' frequency of the fixed oscillator

f_o frequency of the variable oscillator at the setting $F = F_o$

L inductance in the resonant circuit of variable oscillator

C total capacitance in the resonant circuit of variable oscillator

C_o total capacitance in the resonant circuit of variable oscillator when set to $F = F_o$

$C_a = C - C_o$ capacitance change in the air condenser

ΔF_o variation in the standard frequency F_o

ΔC_a variation in C_a

ΔC_o variation in C_o

ΔL variation in L

Δl smallest readable scale interval

ρ relative frequency error $\Delta F/F$

ΔF error in F

ΔF_{Fo} error in F caused by ΔF_o

ΔF_{Co} error in F caused by ΔC_o

ΔF_{Ca} error in F caused by ΔC_a

References

"Beat-Frequency Oscillators," M. S. Mead, Jr. *G. E. Rev.*, v. 32, pp. 521–529, Oct., 1929.

"Beat-Frequency Oscillator," M. F. Cooper and L. G. Page, *Wireless Engineer*, v. 10, pp. 469–476, Sept., 1933.

"Precision Heterodyne Oscillators," D. A. Bell, *Wireless Engineer*, v. 11, p. 308, June, 1934.

"Precision Heterodyne Oscillators," W. H. F. Griffiths, *Wireless Engineer*, v. 11, pp. 234–244, May, 1934.

"A Precision Heterodyne Oscillator," L. E. Ryall, *P. O. E. E., Jl.*, v. 27, pp. 213–221, Oct., 1934.

"The Beat-Frequency Oscillator," M. Slaffer, *Wireless Engineer*, v. 11, pp. 25–26, Jan., 1934. (Letter to the Editor.)

"Further Notes on Precision Heterodyne Oscillators," W. H. F. Griffiths, *Wireless Engineer*, v. 12, pp. 357–362, July, 1935.

"Beat Frequency Oscillators," A. W. Barber, *Radio Engg.*, v. 16, pp. 13–16, July, 1936.

"A Beat-Frequency Oscillator," L. B. Hallman, *Commun. & Broadcast Engg.*, v. 3, pp. 10–13, May, 1936.

Relations Between Attenuation and Phase in Feedback Amplifier Design

By H. W. BODE

Introduction

THE engineer who embarks upon the design of a feedback amplifier must be a creature of mixed emotions. On the one hand, he can rejoice in the improvements in the characteristics of the structure which feedback promises to secure him.[1] On the other hand, he knows that unless he can finally adjust the phase and attenuation characteristics around the feedback loop so the amplifier will not spontaneously burst into uncontrollable singing, none of these advantages can actually be realized. The emotional situation is much like that of an impecunious young man who has impetuously invited the lady of his heart to see a play, unmindful, for the moment, of the limitations of the $2.65 in his pockets. The rapturous comments of the girl on the way to the theater would be very pleasant if they were not shadowed by his private speculation about the cost of the tickets.

In many designs, particularly those requiring only moderate amounts of feedback, the bogy of instability turns out not to be serious after all. In others, however, the situation is like that of the young man who has just arrived at the box office and finds that his worst fears are realized. But the young man at least knows where he stands. The engineer's experience is more tantalizing. In typical designs the loop characteristic is always satisfactory—except for one little point. When the engineer changes the circuit to correct that point, however, difficulties appear somewhere else, and so on ad infinitum. The solution is always just around the corner.

Although the engineer absorbed in chasing this rainbow may not realize it, such an experience is almost as strong an indication of the existence of some fundamental physical limitation as the census which the young man takes of his pockets. It reminds one of the experience of the inventor of a perpetual motion machine. The perpetual motion machine, likewise, always works—except for one little factor. Evidently, this sort of frustration and lost motion is inevitable in

[1] A general acquaintance with feedback circuits and the uses of feedback is assumed in this paper. As a broad reference, see H. S. Black, "Stabilized Feedback Amplifiers," *B. S. T. J.*, January, 1934.

feedback amplifier design as long as the problem is attacked blindly. To avoid it, we must have some way of determining in advance when we are either attempting something which is beyond our resources, like the young man on the way to the theater, or something which is literally impossible, like the perpetual motion enthusiast.

This paper is written to call attention to several simple relations between the gain around an amplifier loop, and the phase change around the loop, which impose limits to what can and cannot be done in a feedback design. The relations are mathematical laws, which in their sphere have the same inviolable character as the physical law which forbids the building of a perpetual motion machine. They show that the attempt to build amplifiers with certain types of loop characteristics *must* fail. They permit other types of characteristic, but only at the cost of certain consequences which can be calculated. In particular, they show that the loop gain cannot be reduced too abruptly outside the frequency range which is to be transmitted if we wish to secure an unconditionally stable amplifier. It is necessary to allow at least a certain minimum interval before the loop gain can be reduced to zero.

The question of the rate at which the loop gain is reduced is an important one, because it measures the actual magnitude of the problem confronting both the designer and the manufacturer of the feedback structure. Until the loop gain is zero, the amplifier will sing unless the loop phase shift is of a prescribed type. The cutoff interval as well as the useful transmission band is therefore a region in which the characteristics of the apparatus must be controlled. The interval represents, in engineering terms, the price of the ticket.

The price turns out to be surprisingly high. It can be minimized by accepting an amplifier which is only conditionally stable.[2] For the customary absolutely stable amplifier, with ordinary margins against singing, however, the price in terms of cutoff interval is roughly one octave for each ten db of feedback in the useful band. In practice, an additional allowance of an octave or so, which can perhaps be regarded as the tip to the hat check girl, must be made to insure that the amplifier, having once cut off, will stay put. Thus in an amplifier with 30 db feedback, the frequency interval over which effective control of the loop transmission characteristics is necessary is at least four octaves, or sixteen times, broader than the useful band. If we raise the feedback to 60 db, the effective range must be more than a hundred times the useful range. If the useful band is itself large these factors

[2] Definitions of conditionally and unconditionally stable amplifiers are given on page 432.

may lead to enormous effective ranges. For example, in a 4 megacycle amplifier they indicate an effective range of about 60 megacycles for 30 db feedback, or of more than 400 megacycles if the feedback is 60 db.

The general engineering implications of this result are obvious. It evidently places a burden upon the designer far in excess of that which one might anticipate from a consideration of the useful band alone. In fact, if the required total range exceeds the band over which effective control of the amplifier loop characteristics is physically possible, because of parasitic effects, he is helpless. Like the young man, he simply can't pay for his ticket. The manufacturer, who must construct and test the apparatus to realize a prescribed characteristic over such wide bands, has perhaps a still more difficult problem. Unfortunately, the situation appears to be an inevitable one. The mathematical laws are inexorable.

Aside from sounding this warning, the relations between loop gain and loop phase can also be used to establish a definite method of design. The method depends upon the development of overall loop characteristics which give the optimum result, in a certain sense, consistent with the general laws. This reduces actual design procedure to the simulation of these characteristics by processes which are essentially equivalent to routine equalizer design. The laws may also be used to show how the characteristics should be modified when the cutoff interval approaches the limiting band width established by the parasitic elements of the circuit, and to determine how the maximum realizable feedback in any given situation can be calculated. These methods are developed at some length in the writer's U. S. Patent No. 2,123,178 and are explained in somewhat briefer terms here.

Relations Between Attenuation and Phase in Physical Networks [3]

The amplifier design theory advanced here depends upon a study of the transmission around the feedback loop in terms of a number of general laws relating the attenuation and phase characteristics of physical networks. In attacking this problem an immediate difficulty presents itself. It is apparent that no entirely definite and universal

[3] Network literature includes a long list of relations between attenuation and phase discovered by a variety of authors. They are derived typically from a Fourier analysis of the transient response of assumed structures and are frequently ambiguous, because of failure to recognize the minimum phase shift condition. No attempt is made to review this work here, although special mention should be made of Y. W. Lee's paper in the *Journal for Mathematics and Physics* for June, 1932. The proof of the relations given in the present paper depends upon a contour integration in the complex frequency plane and can be understood from the disclosure in the patent referred to previously.

relation between the attenuation and the phase shift of a physical structure can exist. For example, we can always change the phase shift of a circuit without affecting its loss by adding either an ideal transmission line or an all-pass section. Any attenuation characteristic can thus correspond to a vast variety of phase characteristics.

For the purposes of amplifier design this ambiguity is fortunately unimportant. While no unique relation between attenuation and phase can be stated for a general circuit, a unique relation does exist between any given loss characteristic and the *minimum* phase shift which must be associated with it. In other words, we can always add a line or all-pass network to the circuit but we can never subtract such a structure, unless, of course, it happens to be part of the circuit originally. If the circuit includes no surplus lines or all-pass sections, it will have at every frequency the least phase shift (algebraically) which can be obtained from any physical structure having the given attenuation characteristic. The least condition, since it is the most favorable one, is, of course, of particular interest in feedback amplifier design.

For the sake of precision it may be desirable to restate the situations in which this minimum condition fails to occur. The first situation is found when the circuit includes an all-pass network either as an individual structure or as a portion of a network which can be replaced by an all-pass section in combination with some other physical structure.[4] The second situation is found when the circuit includes a transmission line. The third situation occurs when the frequency is so high that the tubes, network elements and wiring cannot be considered to obey a lumped constant analysis. This situation may be found, for example, at frequencies for which the transit time of the tubes is important or for which the distance around the feedback loop is an appreciable part of a wave-length. The third situation is, in many respects, substantially the same as the second, but it is mentioned separately here as a matter of emphasis. Since the effective band of a feedback amplifier is much greater than its useful band, as the introduction pointed out, the considerations it reflects may be worth taking into account even when they would be trivial in the useful band alone.

It will be assumed here that none of these exceptional situations is found. For the minimum phase condition, then, it is possible to derive

[4] Analytically this condition can be stated as follows: Let it be supposed that the transmission takes place between mesh 1 and mesh 2. The circuit will include an all-pass network, explicit or concealed, if any of the roots of the minor Δ_{12} of the principal circuit determinant lie below the real axis in the complex frequency plane. This can happen in bridge configurations, but not in series-shunt configurations, so that all ladder networks are automatically of minimum phase type.

a large number of relations between the attenuation and phase characteristics of a physical network. One of the simplest is

$$\int_{-\infty}^{\infty} B du = \frac{\pi}{2}(A_\infty - A_0), \tag{1}$$

where u represents $\log f/f_0$, f_0 being an arbitrary reference frequency, B is the phase shift in radians, and A_0 and A_∞ are the attenuations in nepers at zero and infinite frequency, respectively. The theorem states, in effect, that the total area under the phase characteristic plotted on a logarithmic frequency scale depends only upon the difference between the attenuations at zero and infinite frequency, and not upon the course of the attenuation between these limits. Nor does it depend upon the physical configuration of the network unless a non-minimum phase structure is chosen, in which case the area is necessarily increased. The equality of phase areas for attenuation characteristics of different types is illustrated by the sketches of Fig. 1.

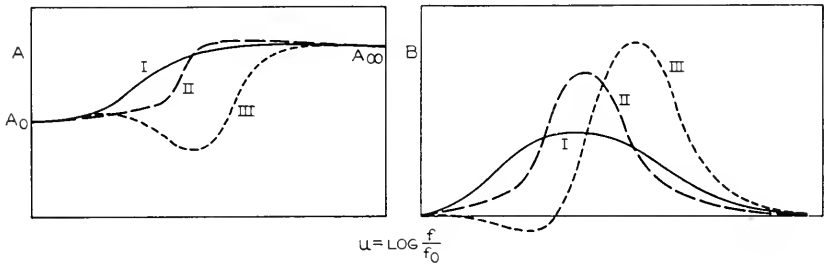


Fig. 1—Diagram to illustrate relation between phase area and change in attenuation.

The significance of the phase area relation for feedback amplifier design can be understood by supposing that the practical transmission range of the amplifier extends from zero to some given finite frequency. The quantity $A_0 - A_\infty$ can then be identified with the change in gain around the feedback loop required to secure a cut-off. Associated with it must be a certain definite phase area. If we suppose that the maximum phase shift at any frequency is limited to some rather low value the total area must be spread out over a proportionately broad interval on the frequency scale. This must correspond roughly to the cut-off region, although the possibility that some of the area may be found above or below the cut-off range prevents us from determining the necessary interval with precision.

A more detailed statement of the relationship between phase shift and change in attenuation can be obtained by turning to a second

theorem. It reads as follows:

$$B(f_c) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{dA}{du} \log \coth \frac{|u|}{2}\, du, \tag{2}$$

where $B(f_c)$ represents the phase shift at any arbitrarily chosen frequency f_c and $u = \log f/f_c$. This equation, like (1), holds only for the minimum phase shift case.

Although equation (2) is somewhat more complicated than its predecessor, it lends itself to an equally simple physical interpretation. It is clear, to begin with, that the equation implies broadly that the

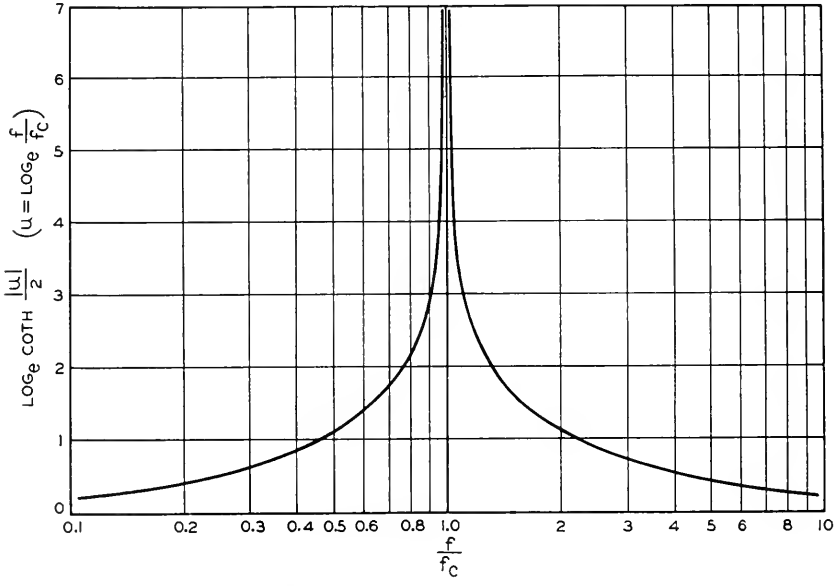


Fig. 2—Weighting function in loss-phase formula.

phase shift at any frequency is proportional to the derivative of the attenuation on a logarithmic frequency scale. For example, if dA/du is doubled B will also be doubled. The phase shift at any particular frequency, however, does not depend upon the derivative of attenuation at that frequency alone, but upon the derivative at all frequencies, since it involves a summing up, or integration, of contributions from the complete frequency spectrum. Finally, we notice that the contributions to the total phase shift from the various portions of the frequency spectrum do not add up equally, but rather in accordance with the function log coth $|u|/2$. This quantity, therefore, acts as a weighting function. It is plotted in Fig. 2. As we might expect physically

it is much larger near the point $u = 0$ than it is in other regions. We can, therefore, conclude that while the derivative of attenuation at all frequencies enters into the phase shift at any particular frequency $f = f_c$ the derivative in the neighborhood of f_c is relatively much more important than the derivative in remote parts of the spectrum.

As an illustration of (2), let it be supposed that $A = ku$, which corresponds to an attenuation having a constant slope of 6 k db per octave. The associated phase shift is easily evaluated. It turns out, as we might expect, to be constant, and is equal numerically to $k\pi/2$ radians. This is illustrated by Fig. 3. As a second example, we may consider

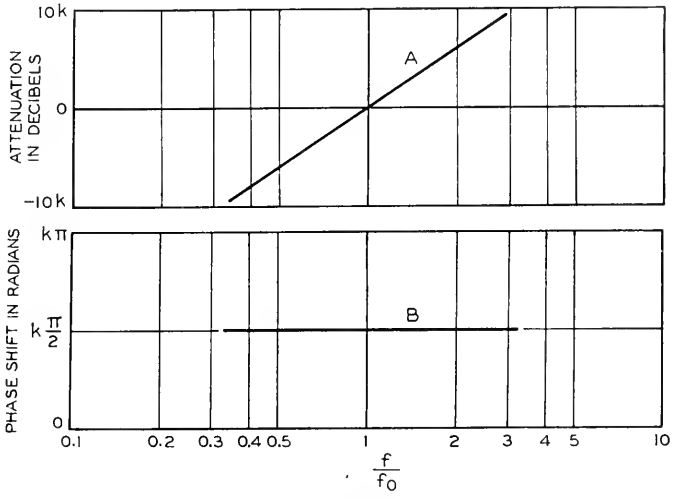


Fig. 3—Phase characteristic corresponding to a constant slope attenuation.

a discontinuous attenuation characteristic such as that shown in Fig. 4. The associated phase characteristic, also shown in Fig. 4, is proportional to the weighting function of Fig. 2.

The final example is shown by Fig. 5. It consists of an attenuation characteristic which is constant below a specified frequency f_b and has a constant slope of 6 k db per octave above f_b. The associated phase characteristic is symmetrical about the transition point between the two ranges. At sufficiently high frequencies, the phase shift approaches the limiting $k\pi/2$ radians which would be realized if the constant slope were maintained over the complete spectrum. At low frequencies the phase shift is substantially proportional to frequency and is given by the equation

$$B = \frac{2k}{\pi} \frac{f}{f_b} . \tag{3}$$

Solutions developed in this way can be added together, since it is apparent from the general relation upon which they are based that the phase characteristic corresponding to the sum of two attenuation

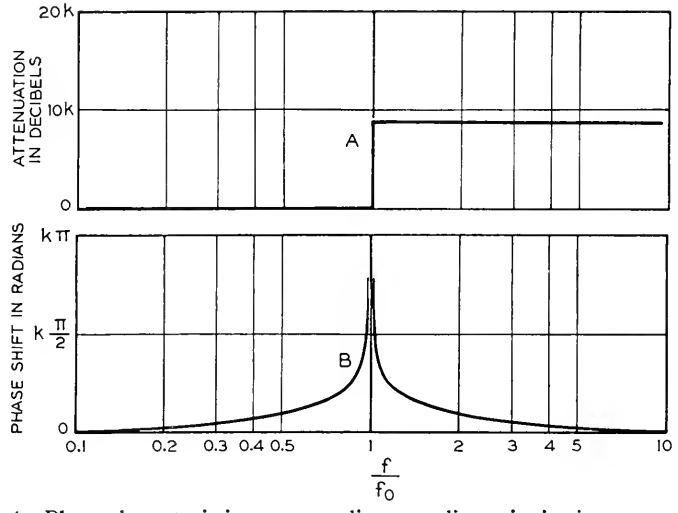


Fig. 4—Phase characteristic corresponding to a discontinuity in attenuation.

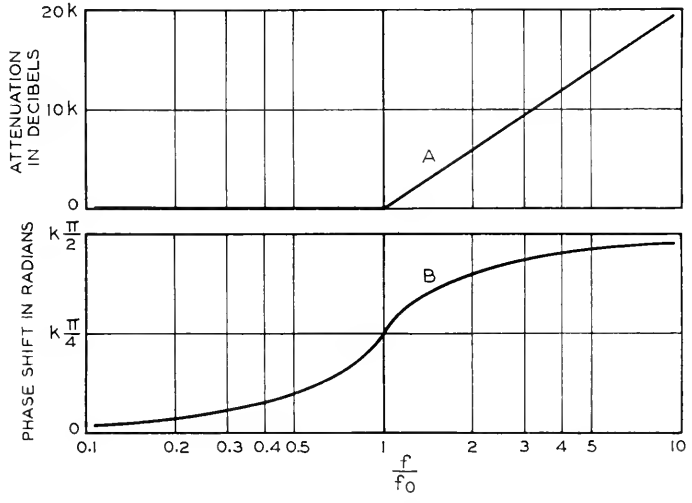


Fig. 5—Phase characteristic corresponding to an attenuation which is constant below a prescribed frequency and has a constant slope above it.

characteristics will be equal to the sum of the phase characteristics corresponding to the two attenuation characteristics separately. We can therefore combine elementary solutions to secure more complicated

characteristics. An example is furnished by Fig. 6, which is built up from three solutions of the type shown by Fig. 5. By proceeding sufficiently far in this way, an approximate computation of the phase characteristic associated with almost any attenuation characteristic can be made, without the labor of actually performing the integration in (2).

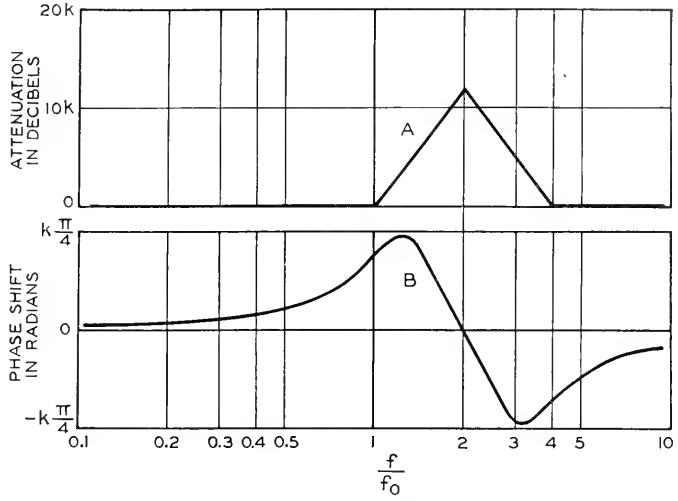


Fig. 6—Diagram to illustrate addition of elementary attenuation and phase characteristics to produce more elaborate solutions of the loss-phase formula.

Equations (1) and (2) are the most satisfactory expressions to use in studying the relation between loss and phase in a broad physical sense. The mechanics of constructing detailed loop cut-off characteristics, however, are simplified by the inclusion of one other, somewhat more complicated, formula. It appears as

$$\int_0^{f_0} \frac{A\,df}{\sqrt{f_0^2 - f^2}\,(f^2 - f_c^2)} + \int_{f_0}^{\infty} \frac{B\,df}{\sqrt{f^2 - f_0^2}\,(f^2 - f_c^2)}$$

$$= \frac{\pi}{2 f_c \sqrt{f_0^2 - f_c^2}}, \qquad f_c < f_0$$

$$= -\frac{\pi}{2 f_c \sqrt{f_c^2 - f_0^2}}, \qquad f_c > f_0, \quad (4)$$

where f_0 is some arbitrarily chosen frequency and the other symbols have their previous significance.

The meaning of (4) can be understood if it is recalled that (2) implies that the minimum phase shift at any frequency can be computed if the

attenuation is prescribed at all frequencies. In the same way (4) shows how the complete attenuation and phase characteristics can be determined if we begin by prescribing the attenuation below f_0 and the phase shift above f_0. Since f_0 can be chosen arbitrarily large or small this is evidently a more general formula than either (1) or (2), while it can itself be generalized, by the introduction of additional irrational factors, to provide for more elaborate patterns of bands in which A and B are specified alternately.

As an example of this formula, let it be assumed that $A = K$ for $f < f_0$ and that $B = k\pi/2$ for $f > f_0$. These are shown by the solid lines in Fig. 7. Substitution in (4) gives the A and B characteristics in the rest of the spectrum as

$$B = k \sin^{-1} \frac{f}{f_0}, \qquad f < f_0$$

$$A = K + k \log \left[\sqrt{\frac{f^2}{f_0^2} - 1} + \frac{f}{f_0} \right], \qquad f > f_0. \tag{5}$$

These are indicated by broken lines in Fig. 7. In this particularly

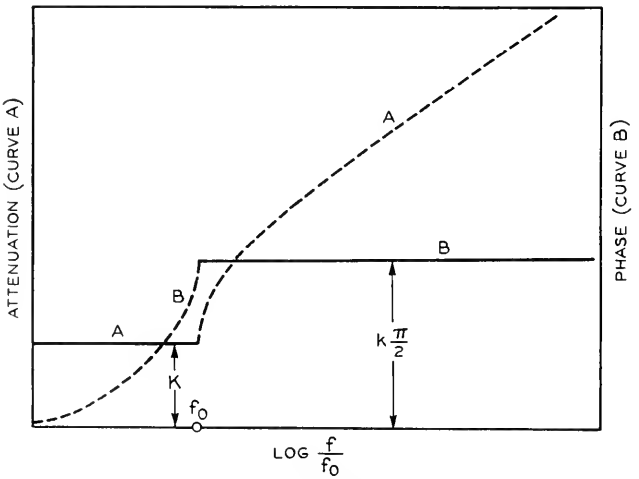


Fig. 7—Construction of complete characteristics from an attenuation characteristic specified below a certain frequency and a phase characteristic above it. The solid lines represent the specified attenuation and phase characteristics, and the broken lines their computed extensions to the rest of the spectrum.

simple case all four fragments can be combined into the single analytic formula

$$A + iB = K + k \log \left[\sqrt{1 - \frac{f^2}{f_0^2}} + i \frac{f}{f_0} \right]. \tag{6}$$

This expression will be used as the fundamental formula for the loop cut-off characteristic in the next section.

Overall Feedback Loop Characteristics

The survey just concluded shows what combinations of attenuation and phase characteristics are physically possible. We have next to determine which of the available combinations is to be regarded as representing the transmission around the overall feedback loop. The choice will naturally depend somewhat upon exactly what we assume that the amplifier ought to do, but with any given set of assumptions it is possible, at least in theory, to determine what combination is most appropriate.

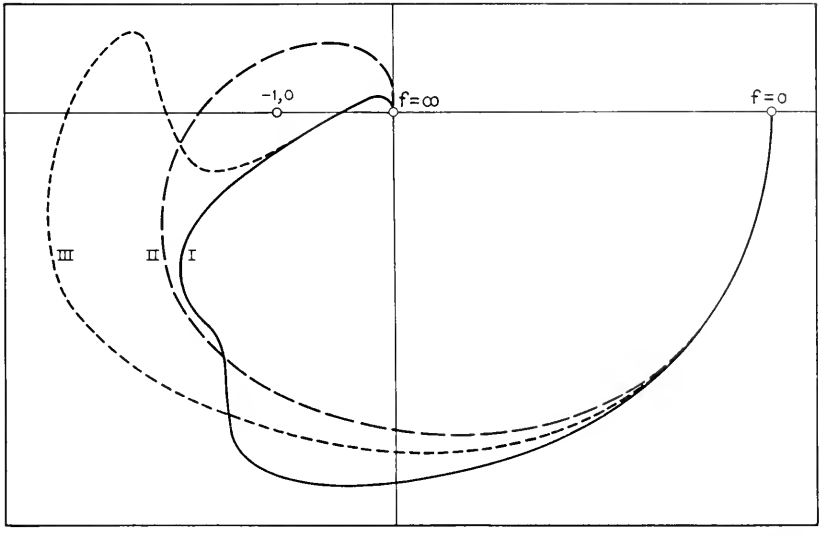


Fig. 8—Nyquist stability diagrams for various amplifiers. Curve I represents "absolute" stability, Curve II instability, and Curve III "conditional" stability. In accordance with the convention used in this paper the diagram is rotated through 180° from its normal position so that the critical point occurs at − 1, 0 rather than + 1, 0.

The situation is conveniently investigated by means of the Nyquist stability diagram [5] illustrated by Fig. 8. The diagram gives the path

[5] *Bell System Technical Journal*, July, 1932. See also Peterson, Kreer, and Ware, *Bell System Technical Journal*, October, 1934. The Nyquist diagrams in the present paper are rotated through 180° from the positions in which they are usually drawn, turning the diagrams in reality into plots of $-\mu\beta$. In a normal amplifier there is one net phase reversal due to the tubes in addition to any phase shifts chargeable directly to the passive networks in the circuit. The rotation of the diagram allows this phase reversal to be ignored, so that the phase shifts actually shown are the same as those which are directly of design interest.

traced by the vector representing the transmission around the feedback loop as the frequency is assigned all possible real values. In accordance with Nyquist's results a path such as II, which encircles the point $-1, 0$, indicates an unstable circuit and must be avoided. A stable amplifier is obtained if the path resembles either I or III, neither of which encircles $-1, 0$. The stability represented by Curve III, however, is only "Nyquist" or "conditional." The path will enclose the critical point if it is merely reduced in scale, which may correspond physically to a reduction in tube gain. Thus the circuit may sing when the tubes begin to lose their gain because of age, and it may also sing, instead of behaving as it should, when the tube gain increases from zero as power is first applied to the circuit. Because of these possibilities conditional stability is usually regarded as undesirable and the present discussion will consequently be restricted to "absolutely" or "unconditionally" stable amplifiers having Nyquist diagrams of the type resembling Curve I.

The condition that the amplifier be absolutely stable is evidently that the loop phase shift should not exceed 180° until the gain around the loop has been reduced to zero or less. A theoretical characteristic which just met this requirement, however, would be unsatisfactory, since it is inevitable that the limiting phase would be exceeded in fact by minor deviations introduced either in the detailed design of the amplifier or in its construction. It will therefore be assumed that the limiting phase is taken as 180° less some definite margin. This is illustrated by Fig. 9, the phase margin being indicated as $y\pi$ radians. At frequencies remote from the band it is physically impossible, in most circuits, to restrict the phase within these limits. As a supplement, therefore, it will be assumed that larger phase shifts are permissible if the loop gain is x db below zero. This is illustrated by the broken circular arc in Fig. 9. A theoretical loop characteristic meeting both requirements will be developed for an amplifier transmitting between zero and some prescribed limiting frequency with a constant feedback, and cutting off thereafter as rapidly as possible. This basic characteristic can be adapted to amplifiers with varying feedback in the useful range or with useful ranges lying in other parts of the spectrum by comparatively simple modifications which are described at a later point. It is, of course, contemplated that the gain and phase margins x and y will be chosen arbitrarily in advance. If we choose large values we can permit correspondingly large tolerances in the detailed design and construction of the apparatus without risk of instability. It turns out, however, that with a prescribed width of cutoff interval the amount of feedback which can be realized in the

useful range is decreased as the assumed margins are increased, so that it is generally desirable to choose as small margins as is safe.

The essential feature in this situation is the requirement that the diminution of the loop gain in the cutoff region should not be accompanied by a phase shift exceeding some prescribed amount. In view of the close connection between phase shift and the slope of the attenuation characteristic evidenced by (2) this evidently demands that the amplifier should cut off, on the whole, at a well defined rate which is not too fast. As a first approximation, in fact, we can choose the cutoff

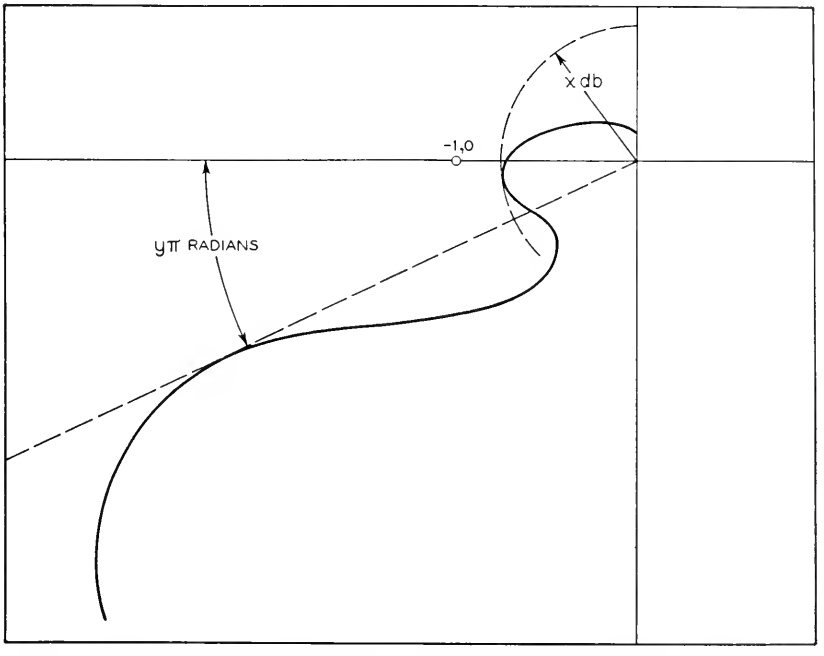


Fig. 9—Diagram to illustrate definitions of phase and gain margins for the feedback loop.

characteristic as an exactly constant slope from the edge of the useful band outward. Such a characteristic has already been illustrated by Fig. 5 and is shown, replotted,[6] by the broken lines in Fig. 10. If we choose the parameter corresponding to k in Fig. 5 as 2 the cutoff rate is 12 db per octave and the phase shift is substantially 180° at high frequencies. This choice thus leads to zero phase margin. By choosing a somewhat smaller k on the other hand, we can provide a definite

[6] To prevent confusion it should be noticed that the general attenuation-phase diagrams are plotted in terms of relative loss while loop cutoff characteristics, here and at later points, are plotted in terms of relative gain.

margin against singing, at the cost of a less rapid cutoff. For example, if we choose $k = 1.5$ the limiting phase shift in the $\mu\beta$ loop becomes 135°, which provides a margin of 45° against instability, while the rate of cutoff is reduced to 9 db per octave. The value $k = 1.67$, which corresponds to a cutoff rate of 10 db per octave and a phase margin of 30°, has been chosen for illustrative purposes in preparing Fig. 10. The loss margin depends upon considerations which will appear at a later point.

Although characteristics of the type shown by Fig. 5 are reasonably satisfactory as amplifier cutoffs they evidently provide a greater phase

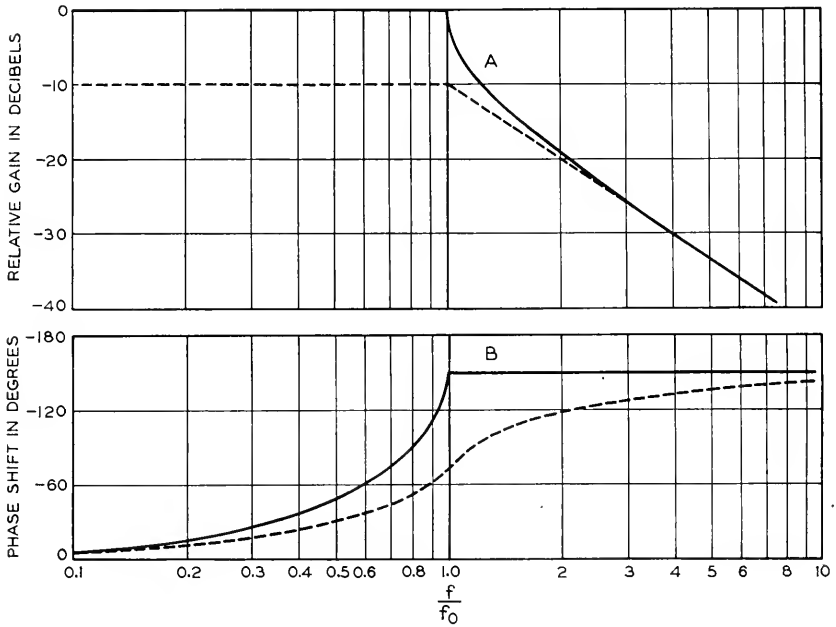


Fig. 10—Ideal loop cutoff characteristics. Drawn for a 30° phase margin.

margin against instability in the region just beyond the useful band than they do at high frequencies. In virtue of the phase area law this must be inefficient if, as is supposed here, the optimum characteristic is one which would provide a constant margin throughout the cutoff interval. The relation between the phase and the slope of the attenuation suggests that a constant phase margin can be obtained by increasing the slope of the cutoff characteristic near the edge of the band, leaving its slope at more remote frequencies unchanged, as shown by the solid lines in Fig. 10. The exact expression for the required curve can be found from (6), where the problem of determining such a characteristic appeared as an example of the use of the general formula (4).

At high frequencies the new phase and attenuation characteristics merge with those obtained from the preceding straight line cutoff, as Fig. 10 indicates. In this region the relation between phase margin and cutoff slope is fixed by the k in the equation (6) in the manner already described for the more elementary cutoff. At low frequencies, however, the increased slope near the edge of the band permits 6 k db more feedback.

It is worth while to pause here to consider what may be said, on the basis of these characteristics, concerning the breadth of cutoff interval required for a given feedback, or the "price of the ticket," as it was expressed in the introduction. If we adopt the straight line cutoff and assume the k used in Fig. 10 the interval between the edge of the useful band and the intersection of the characteristic with the zero gain axis is evidently exactly 1 octave for each 10 db of low frequency feedback. The increased efficiency of the solid line characteristic saves one octave of this total if the feedback is reasonably large to begin with. This apparently leads to a net interval one or two octaves narrower than the estimates made in the introduction. The additional interval is required to bridge the gap between a purely mathematical formula such as (6), which implies that the loop characteristics follow a prescribed law up to indefinitely high frequencies, and a physical amplifier, whose ultimate loop characteristics vary in some uncontrollable way. This will be discussed later. It is evident, of course, that the cutoff interval will depend slightly upon the margins assumed. For example, if the phase margin is allowed to vanish the cutoff rate can be increased from 10 to 12 db per octave. This, however, is not sufficient to affect the order of magnitude of the result. Since the diminished margin is accompanied by a corresponding increase in the precision with which the apparatus must be manufactured such an economy is, in fact, a Pyrrhic victory unless it is dictated by some such compelling consideration as that described in the next section.

Maximum Obtainable Feedback

A particularly interesting consequence of the relation between feedback and cutoff interval is the fact that it shows why we cannot obtain unconditionally stable amplifiers with as much feedback as we please. So far as the purely theoretical construction of curves such as those in Fig. 10 is concerned, there is clearly no limit to the feedback which can be postulated. As the feedback is increased, however, the cutoff interval extends to higher and higher frequencies. The process reaches a physical limit when the frequency becomes so high that parasitic effects in the circuit are controlling and do not permit the prescribed cutoff

characteristic to be simulated with sufficient precision. For example, we are obviously in physical difficulties if the cutoff characteristic specifies a net gain around the loop at a frequency so high that the tubes themselves working into their own parasitic capacitances do not give a gain.

This limitation is studied most easily if the effects of the parasitic elements are lumped together by representing them in terms of the asymptotic characteristic of the loop as a whole at extremely high frequencies. An example is shown by Fig. 11. The structure is a

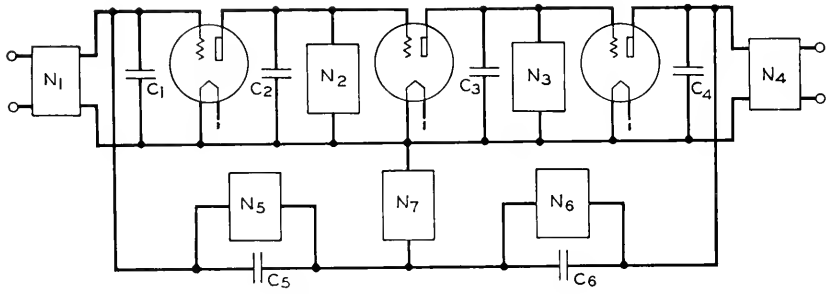


Fig. 11—Elements which determine the asymptotic loop transmission characteristic in a typical amplifier.

shunt feedback amplifier. The β circuit is represented by the T composed of networks N_5, N_6 and N_7. The input and output circuits are represented by N_1 and N_4 and the interstage impedances by N_2 and N_3. The C's are parasitic capacitances with the exception of C_5 and C_6, which may be regarded as design elements added deliberately to N_5 and N_6 to obtain an efficient high frequency transmission path from output to input. At sufficiently high frequencies the loop transmission will depend only upon these various capacitances, without regard to the N's. Thus, if the transconductances of the tubes are represented by G_1, G_2, and G_3 the asymptotic gains of the first two tubes are $G_1/\omega C_1$ and $G_2/\omega C_3$. The rest of the loop includes the third tube and the potentiometer formed by the capacitances C_1, C_4, C_5 and C_6. Its asymptotic transmission can be written as $G_3/\omega C$, where

$$C = C_1 + C_4 + \frac{C_1 C_4}{C_5 C_6}(C_5 + C_6).$$

Each of these terms diminishes at a rate of 6 db per octave. The complete asymptote is $G_1G_2G_3/\omega^3 CC_2C_3$. It appears as a straight line with a slope of 18 db per octave when plotted on logarithmic frequency paper.

A similar analysis can evidently be made for any amplifier. In the particular circuit shown by Fig. 11 the slope of the asymptote, in units of 6 db per octave, is the same as the number of tubes in the circuit. The slope can evidently not be less than the number of tubes but it may be greater in some circuits. For example if C_5 and C_6 were omitted in Fig. 11 and N_5 and N_6 were regarded as degenerating into resistances the asymptote would have a slope of 24 db per octave and would lie below the present asymptote at any reasonably high frequency. In any event the asymptote will depend only upon the parasitic elements of the circuit and perhaps a few of the most significant design elements. It can thus be determined from a skeletonized version of the final structure. If waste of time in false starts is to be avoided such a determination should be made as early as possible, and certainly in advance of any detailed design.

The effect of the asymptote on the overall feedback characteristic is illustrated by Fig. 12. The curve $ABEF$ is a reproduction of the ideal

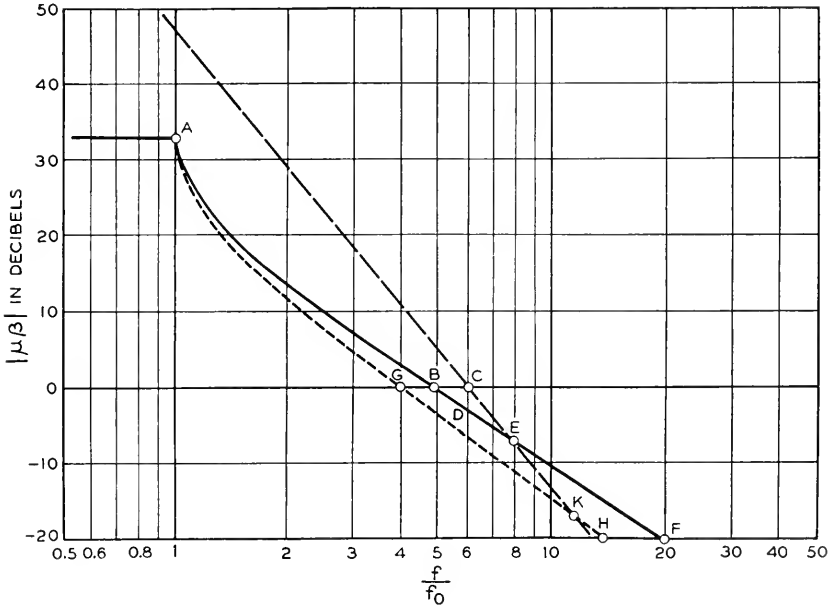


Fig. 12—Combination of asymptotic characteristic and ideal cutoff characteristic.

cutoff characteristic originally given by the solid lines in Fig. 10. It will be recalled that the curve was drawn for the choice $k = 5/3$, which corresponds to a phase margin of 30° and an almost constant slope, for the portion DEF of the characteristic, of about 10 db per octave. The

straight line *CEK* represents an asymptote of the type just described, with a slope of 18 db per octave. Since the asymptote may be assumed to represent the practical upper limit of gain in the high-frequency region, the effect of the parasitic elements can be obtained by replacing the theoretical cutoff by the broken line characteristic *ABDEK*. In an actual circuit the corner at *E* would, of course, be rounded off, but this is of negligible quantitative importance. Since *EF* and *EK* diverge by 8 db per octave the effect can be studied by adding curves of the type shown by Fig. 5 to the original cutoff characteristic.

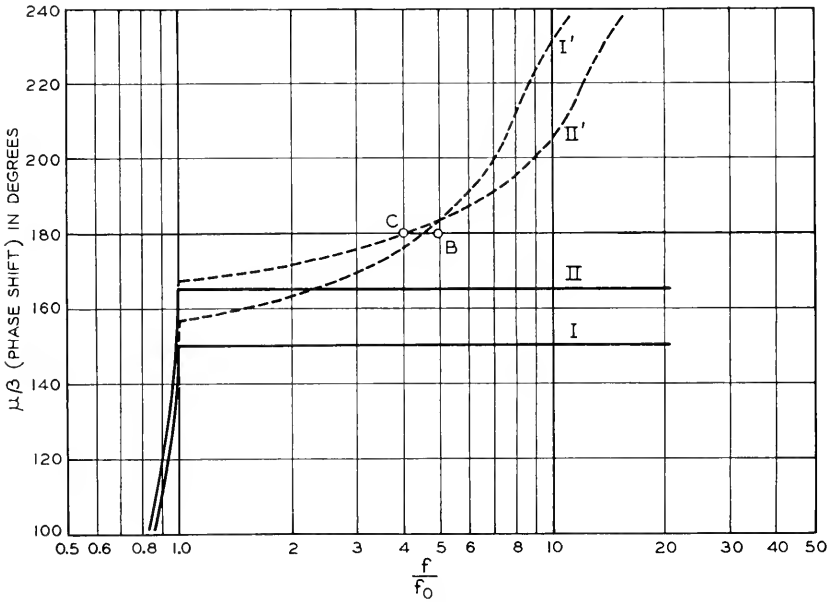


Fig. 13—Phase characteristics corresponding to gain characteristics of Fig. 12.

The phase shift in the ideal case is shown by Curve I of Fig. 13. The addition of the phase corresponding to the extra slope of 8 db per octave at high frequencies produces the total phase characteristic shown by Curve I′. At the point *B* where $|\mu\beta| = 1$, the additional phase shift amounts to 35 degrees. Since this is greater than the original phase margin of 30 degrees the amplifier is unstable when parasitic elements are considered. In the present instance stability can be regained by increasing the coefficient *k* to 1-5/6, which leads to the broken line characteristic *AGKH* in Fig. 12. This reduces the nominal phase margin to 15 degrees, but the frequency interval between *G* and *K* is so much greater than that between *B* and *E* that the added phase is reduced still more and is just less than 15° at the new

cross over point G. This is illustrated by II and II′ in Fig. 13. On the other hand, if the zero gain intercept of the asymptote CEK had occurred at a slightly lower frequency, no change in k alone would have been sufficient. It would have been necessary to reduce the amount of feedback in the transmitted range in order to secure stability.

The final characteristic in Fig. 13 reaches the limiting phase shift of 180° only at the crossover point. It is evident that a somewhat more efficient solution for the extreme case is obtained if the limiting 180° is approximated throughout the cutoff interval. This result is attained by the cutoff characteristic shown in Fig. 14. The characteristic con-

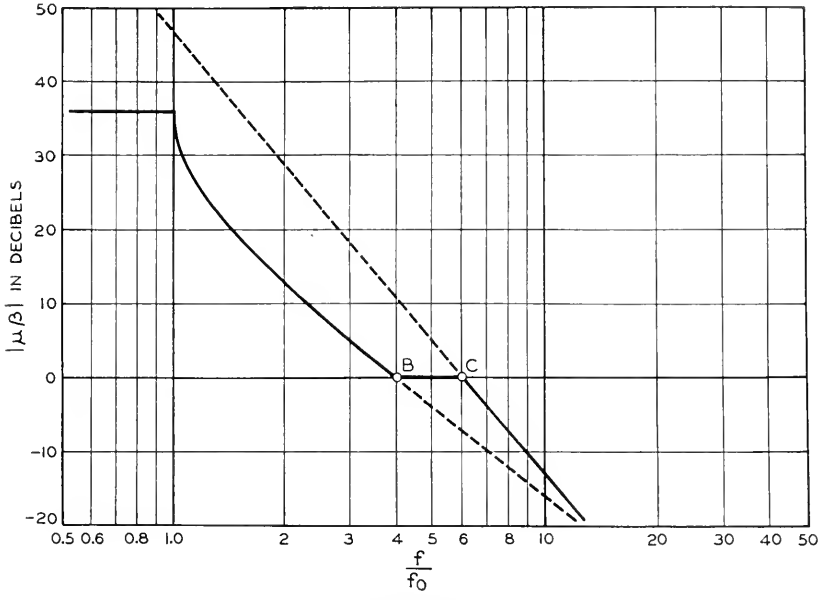


Fig. 14—Ideal cutoff modified to take account of asymptotic characteristic. Drawn for zero gain and phase margins.

sists of the original theoretical characteristic, drawn for $k = 2$, from the edge of the useful band to its intercept with the zero gain axis, the zero gain axis from this frequency to the intercept with the high-frequency asymptote, and the asymptote thereafter. It can be regarded as a combination of the ideal cutoff characteristic and two characteristics of the type shown by Fig. 5. One of the added characteristics starts at B and has a positive slope of 12 db per octave, since the ideal cutoff was drawn for the limiting value of k. The other starts at C and has the negative slope, − 18 db per octave, of the asymptote itself. As (3) shows, the added slopes correspond at lower frequencies to ap-

proximately linear phase characteristics of opposite sign. If the frequencies B and C at which the slopes begin are in the same ratio, 12 : 18, as the slopes themselves the contributions of the added slopes will substantially cancel each other and the net phase shift throughout the cutoff interval will be almost the same as that of the ideal curve alone. The exact phase characteristic is shown by Fig. 15. It dips

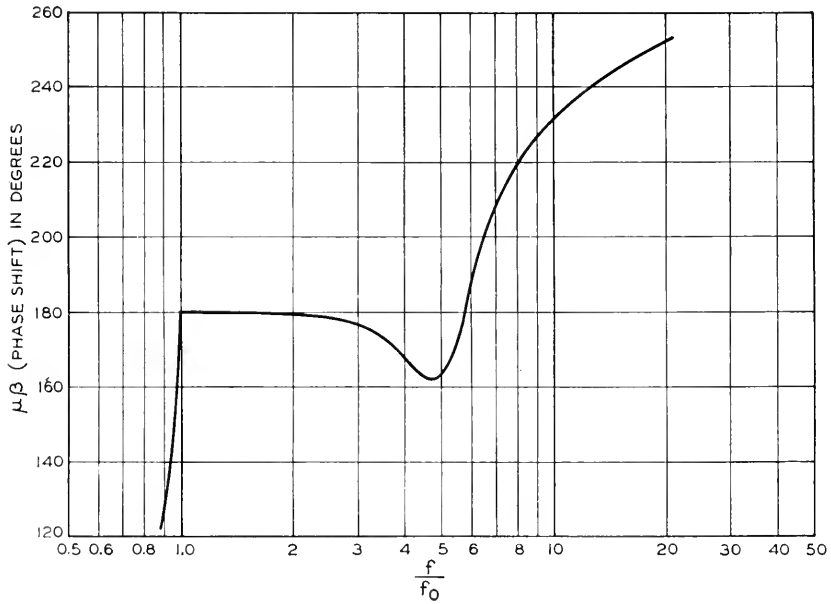


Fig. 15—Phase characteristic corresponding to gain characteristic of Fig. 14.

slightly below 180° at the point at which the characteristic reaches the zero gain axis, so that the circuit is in fact stable.

The same analysis can evidently be applied to asymptotes of any other slope. This makes it easy to compute the maximum feedback obtainable under any asymptotic conditions. If f_0 and f_a are respectively the edge of the useful band and the intercept (C in Figs. 12 and 14) of the asymptote with the zero gain axis, and n is the asymptotic slope, in units of 6 db per octave, the result appears as

$$A_m = 40 \log_{10} \frac{4f_a}{f_0}, \tag{7}$$

where A_m is the maximum feedback in db.[7]

[7] The formulæ for maximum feedback given here and in the later equation (8) are slightly conservative. It follows from the phase area law that more feedback should be obtained if the phase shift were exactly 180° below the crossover and rose

For the sake of generality it is convenient to extend this formula to include also situations in which there exists some further linear phase characteristic in addition to those already taken into account. In exceptional circuits, the final asymptotic characteristic may not be completely established by the time the curve reaches the zero gain axis and the additional phase characteristic may be used to represent the effect of subsequent changes in the asymptotic slope. Such a situation might occur in the circuit of Fig. 11, for example, if C_5 or C_6 were made extremely small. The additional term may also be used to represent departures from a lumped constant analysis in high-frequency amplifiers, as discussed earlier, If we specify the added phase characteristic, from whatever source, by means of the frequency f_d at which it would equal $2n/\pi$ radians, if extrapolated, the general formula corresponding to (7) becomes

$$A_n = 40 \log_{10} \frac{4}{nf_0} \frac{f_a f_d}{f_a + f_d} \cdot \tag{8}$$

It is interesting to notice that equations (7) and (8) take no explicit account of the final external gain of the amplifier. Naturally, if the external gain is too high the available μ circuit gain may not be sufficient to provide it and also the feedback which these formulæ promise. This, however, is an elementary question which requires no further discussion. In other circumstances, the external gain may enter the situation indirectly, by affecting the asymptotic characteristics of the β path, but in a well chosen β circuit this is usually a minor consideration. The external gain does, however, affect the parts of the circuit upon which reliance must be placed in controlling the overall loop characteristic. For example, if the external gain is high the μ circuit will ordinarily be sharply tuned and will drop off rapidly in gain beyond the useful band. The β circuit must therefore provide a decreasing loss to bring the overall cutoff rate within the required limit. Since the β circuit must have initially a high loss to correspond to the high final gain of the complete amplifier, this is possible. Conversely, if the gain of the amplifier is low the μ circuit will be relatively flexible and the β circuit relatively inflexible.

rapidly to its ultimate value thereafter. These possibilities can be exploited approximately by various slight changes in the slope of the cutoff characteristic in the neighborhood of the crossover region, or a theoretical solution can be obtained by introducing a prescribed phase shift of this type in the general formula (4). The theoretical solution gives a Nyquist path which, after dropping below the critical point with a phase shift slightly less than 180°, rises again with a phase shift slightly greater than 180° and continues for some time with a large amplitude and increasing phase before it finally approaches the origin. These possibilities are not considered seriously here because they lead to only a few db increase in feedback, at least for moderate n's, and the degree of design control which they envisage is scarcely feasible in a frequency region where, by definition, parasitic effects are almost controlling.

In setting up (7) and (8) it has been assumed that the amplifier will, if necessary, be built with zero margins against singing. Any surplus which the equations indicate over the actual feedback required can, of course, be used to provide a cutoff characteristic having definite phase and gain margins. For example, if we begin with a lower feedback in the useful band the derivative of the attenuation between this region and the crossover can be proportionately reduced, with a corresponding decrease in phase shift. We can also carry the flat portion of the characteristic below the zero gain axis, thus providing a gain margin when the phase characteristic crosses 180°. In reproportioning the characteristic to suit these conditions, use may be made of the approximate formula

$$A_m - A = (A_m + 17.4)y + \frac{n-2}{n}x + \frac{2}{n}xy, \tag{9}$$

where A_m is the maximum obtainable feedback (in db), A is the actual feedback, and x and y are the gain and phase margins in the notation of Fig. 9. Once the available margin has been divided between the x and

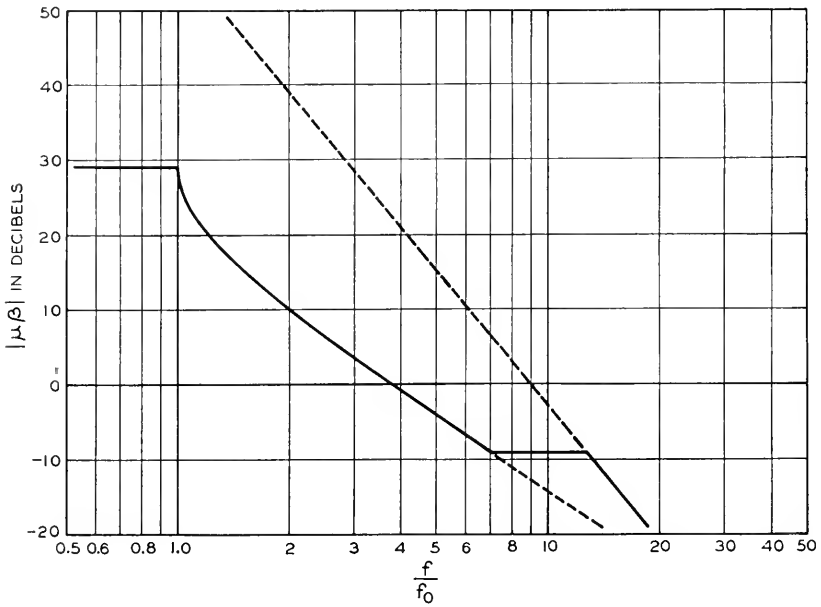


Fig. 16—Modified cutoff permitting 30° phase margin and 9 db gain margin.

y components by means of this formula the cutoff characteristic is, of course, readily drawn in. An example is furnished by Figs. 16 and 17,

where it is assumed that $A_m = 43$ db, $A = 29$ db, $x = 9$ db, $n = 3$ and $y = 1/6$. The Nyquist diagram for the structure is shown by Fig. 18. It evidently coincides almost exactly with the diagram postulated originally in Fig. 9.

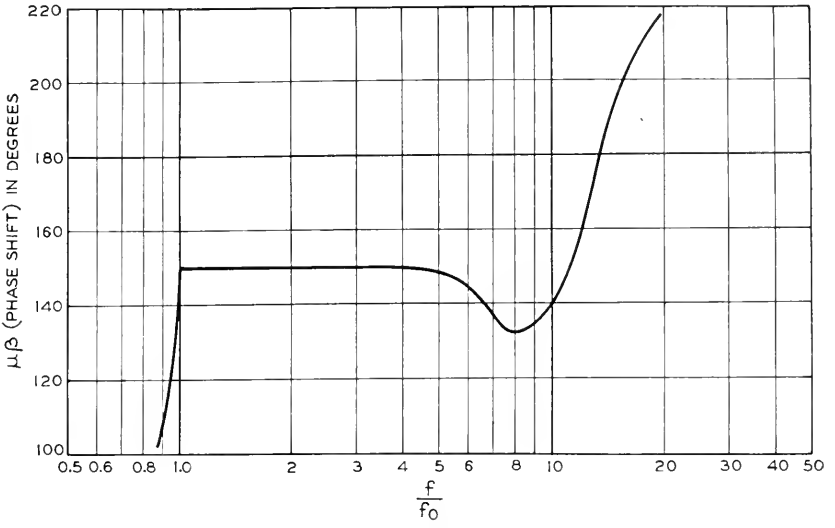


Fig. 17—Phase characteristic corresponding to gain characteristic of Fig. 16.

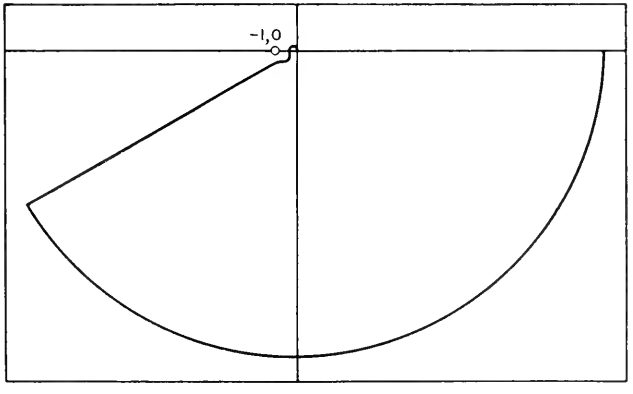


Fig. 18—Nyquist diagram corresponding to gain and phase characteristics of Figs. 16 and 17. As in Fig. 8 the diagram is rotated to place the critical point at $-1, 0$ rather than $+1, 0$.

With the characteristic of Fig. 16 at hand, we can return once more to the calculation of the total design range corresponding to any given feedback. From the useful band to the intersection of the cutoff

characteristic with the zero gain axis the calculation is the same as that made previously in connection with Fig. 10. From the zero gain intercept to the junction with the asymptote, where we can say that design control is finally relaxed, there is, however, an additional interval of nearly two octaves. Although Fig. 16 is fairly typical, the exact breadth of the additional interval will depend somewhat on circumstances. It is increased by an increase in the asymptotic slope and reduced by decreasing the gain margin.

Relative Importance of Tubes and Circuit in Limiting Feedback [8]

The discussion just finished leads to the general conclusion that the feedback which can be obtained in any given amplifier depends ultimately upon the high-frequency asymptote of the feedback loop. It is a matter of some importance, then, to determine what fixes the asymptote and how it can be improved. Evidently, the asymptote is finally restricted by the gains of the tubes alone. We can scarcely improve upon the result secured by connecting the output plate directly to the input grid. Within this limit, however, the actual asymptotic characteristic will depend upon the configuration and type of feedback employed, since a given distribution of parasitic elements may evidently affect one arrangement more than another. The salient circuit problem is therefore that of choosing a general configuration for the feedback circuit which will allow the maximum efficiency of transmission at high frequencies.

The relative importance of tube limitations and circuit limitations is most easily studied if we replace (7) by

$$A_m = 40 \log_{10} \frac{4f_t}{nf_0} - \frac{2A_t}{n}, \tag{10}$$

where f_t is the frequency at which the tubes themselves working into their own parasitic capacitances have zero gain [9] and A_t is the asymptotic loss of the complete feedback loop in db at $f = f_t$. The first term

[8] The material of this section was largely inspired by comments due to Messrs. G. H. Stevenson and J. M. West.

[9] I.e., $f_t = \frac{G_m}{2\pi C}$, where G_m and C are respectively the transconductance and capacitance of a representative tube. The ratio $\frac{G_m}{C}$ is the so-called "figure of merit" of the tube. The analysis assumes that the interstage network is a simple shunt impedance, so that the parasitic capacitance does correctly represent its asymptotic behavior. More complicated four-terminal interstage networks, such as transformer coupling circuits and the like, are generally inadmissible in a feedback amplifier because of the high asymptotic losses and consequent high-phase shifts which they introduce.

of (10) shows how the feedback depends upon the intrinsic band width of the available tubes. In low-power tubes especially designed for the purpose f_t may be 50 mc or more, but if f_0 is small the first term will be substantial even if tubes with much lower values of f_t are selected. The second term gives the loss in feedback which can be ascribed to the rest of the circuit. It is evidently not possible to provide input and output circuits and a β-path without making some contribution to the asymptotic loss, so that A_t cannot be zero. In an amplifier designed with particular attention to this question, however, it is frequently possible to assign A_t a comparatively low value, of the order of 20 to 30 db or less. Without such special attention, on the other hand, A_t is likely to be very much larger, with a consequent diminution in available feedback.

In addition to f_t and A_t, (10) includes the quantity n, which represents the final asymptotic slope in multiples of 6 db per octave. Since the tubes make no contribution to the asymptotic loss at $f = f_t$ we can vary n without affecting A_t by changing the number of tubes in the circuit. This makes it possible to compute the optimum number of tubes which should be used in any given situation in order to provide the maximum possible feedback. If A_t is small the first term of (10) will be the dominant one and it is evidently desirable to have a small number of stages. The limit may be taken as $n = 2$ since with only one stage the feedback is restricted by the available forward gain, which is not taken into account in this analysis. On the other hand since the second term varies more rapidly than the first with n, the optimum number of stages will increase as A_t is increased. It is given generally by

$$n = \frac{A_t}{8.68} \tag{11}$$

or in other words the optimum n is equal to the asymptotic loss at the tube crossover in nepers.

This relation is of particular interest for high-power circuits, such as radio transmitters, where circuit limitations are usually severe but the cost of additional tubes, at least in low-power stages, is relatively unimportant. As an extreme example, we may consider the problem of providing envelope feedback around a transmitter. With the relatively sharp tuning ordinarily used in the high-frequency circuits of a transmitter the asymptotic characteristics of the feedback path will be comparatively unfavorable. For illustrative purposes we may assume that $f_a = 40$ kc. and $n = 6$. In accordance with (7) this would provide a maximum available feedback over a 10 kc. voice band of 17 db. It

will also be assumed that the additional tubes for the low-power portions of the circuit have an f_t of 10 mc.[10] The corresponding A_t is 33 nepers [11] so that equation (11) would say that the feedback would be increased by the addition of as many as 27 tubes to the circuit. Naturally in such an extreme case this result can be looked upon only as a qualitative indication of the direction in which to proceed. If we add only 4 tubes, however, the available feedback becomes 46 db while if we add 10 tubes it reaches 60 db. It is to be observed that only a small part of the available gain of the added tubes is used in directly increasing the feedback. The remainder is consumed in compensating for the unfortunate phase shifts introduced by the rest of the circuit.

Amplifiers of Other Types

The amplifier considered thus far is of a rather special type. It has a useful band extending from zero up to some prescribed frequency f_0, constant feedback in the useful band, and it is absolutely stable. Departures from absolute stability are rather unusual in practical amplifiers and will not be considered here. It is apparent from the phase area relation that a conditionally stable amplifier may be expected to have a greater feedback for a cut-off interval of given breadth than a structure which is unconditionally stable, but a detailed discussion of the problem is beyond the scope of this paper.

Departures from the other assumptions are easily treated. For example, if a varying feedback in the useful band is desired, as it may be in occasional amplifiers, an appropriate cut-off characteristic can be constructed by returning to the general formula (4), performing the integrations graphically, if necessary. If the phase requirement in the cut-off region is left unchanged only the first integral need be modified. The most important question, for ordinary purposes, is that of determining how high the varying feedback can be, in comparison with a corresponding constant feedback characteristic, for any given asymptote. This can be answered by observing the form to which the first integral in (4) reduces when f_c is made very large. It is easily seen that the asymptotic conditions will remain the same provided the

[10] In tubes operating at a high-power level f_t may, of course, be quite low. It is evident, however, that only the tubes added to the circuit are significant in interpreting (11). The additional tubes may be inserted directly in the feedback path if they are made substantially linear in the voice range by subsidiary feedback of their own. This will not affect the essential result of the present analysis.

[11] It is, of course, not to be expected that the actual asymptotic slope will be constant from 40 kc. to 10 mc. Since only the region extending a few octaves above 40 kc. is of interest in the final design, however, the apparent A_t can be obtained by extrapolating the slope in this region.

feedback in the useful band satisfies a relation of the form

$$\int_0^{\pi/2} A d\phi = \text{constant}, \tag{12}$$

where $\phi = \sin^{-1} f/f_0$. Thus the area under the varying characteristic, when plotted against ϕ, should be the same as that under a corresponding constant characteristic having the same phase and gain margins and the same final asymptote. This is exemplified by Fig. 19, the

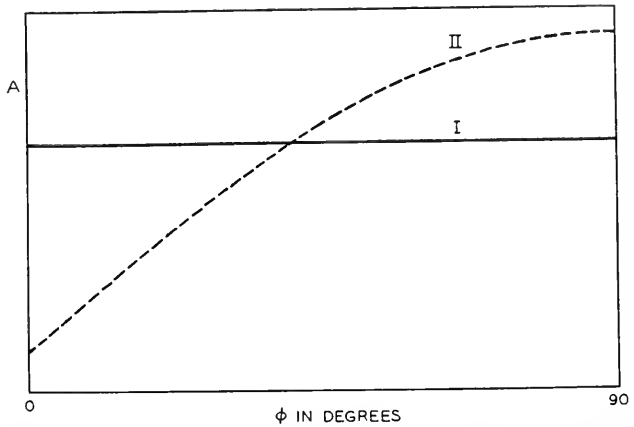


Fig. 19—Diagram to illustrate the computation of available feedback when the required feedback in the useful band is not the same at all frequencies.

varying characteristic being chosen for illustrative purposes as a straight line on an arithmetic frequency scale.

The most important question has to do with the assumption that the useful transmission band extends down to zero frequency. In most amplifiers, of course, this is not true. It is consequently necessary to provide a cut-off characteristic on the lower as well as the upper side of the band. The requisite characteristics are easily obtained from the ones which have been described by means of frequency transformations of a type familiar in filter theory. Thus if the cut-off characteristics studied thus far are regarded as being of the "low-pass" type the characteristics obtained from them by replacing f/f_0 by its reciprocal may be regarded as being of the "high-pass" type. If the band width of the amplifier is relatively broad it is usually simplest to treat the upper and lower cut-offs as independent characteristics of low-pass and high-pass types. In this event, the asymptote for the lower cut-off is furnished by such elements as blocking condensers and choke coils in the plate supply leads. The low-frequency asymptote is usually not so

serious a problem as the high-frequency asymptote since it can be placed as far from the band as we need by using large enough elements in the power supply circuits. The superposition of a low-frequency cutoff on the idealized loop gain and phase characteristics of a "low-pass" circuit is illustrated by the broken lines in Fig. 20.

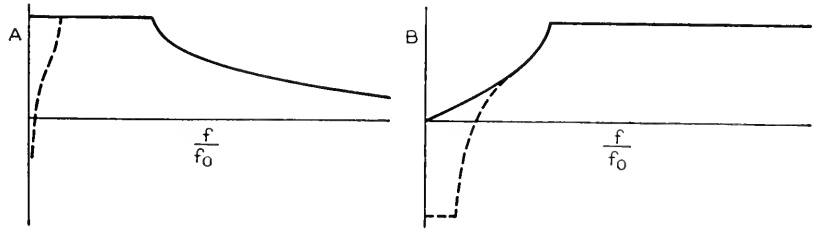


Fig. 20—Modification of loop characteristics to provide a lower cutoff in a broad-band amplifier.

If the band width is relatively narrow it is more efficient to use the transformation in filter theory which relates a low-pass to a symmetrical band-pass structure. The transformation is obtained by replacing f/f_0 in the low-pass case by $(f^2 - f_1f_2/f(f_2 - f_1))$, where f_1 and f_2 are the edges of the prescribed band. It substitutes resonant and anti-resonant circuits tuned to the center of the band for the coils and condensers in the low-pass circuit. In particular each parasitic inductance is tuned by the addition of a series condenser and each parasitic capacity is tuned by a shunt coil. The parameters of the transformation must, of course, be so chosen that the parasitic elements have the correct values for use in the new branches. .

This leads to a simple but important result. If the inductance of a series resonant circuit is fixed, the interval represented by $f_b - f_a$ in Fig. 21, between the frequencies at which the absolute value of the reactance reaches some prescribed limit X_0, is always constant and equal to the frequency at which the untuned inductance would exhibit the reactance X_0, whatever the tuning frequency may be. The same relation holds for the capacity in an anti-resonant circuit. Thus the frequency range over which the branches containing parasitic elements exhibit comparable impedance variations is the same in the band-pass structure and in the prototype low-pass structure. But since the transformation does not affect the relative impedance levels of the various branches in the circuit, this result can be extended to the complete $\mu\beta$ characteristic. We can therefore conclude that *the feedback which is obtainable in an amplifier of given general configuration and with given parasitic elements and given margins depends only upon the breadth*

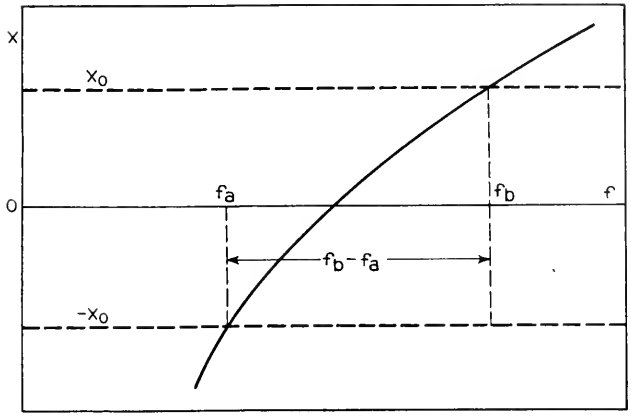


Fig. 21—Frequency interval between prescribed reactances of opposite sign in a resonant circuit with fixed inductance.

of the band in cycles and is independent of the location of the band in the frequency spectrum.

These relations are exemplified by the plots of a low-pass cutoff characteristic and the equivalent band-pass characteristic shown by Fig. 22. The equality of corresponding frequency intervals is indicated by the horizontal lines *A*, *B* and *C*.

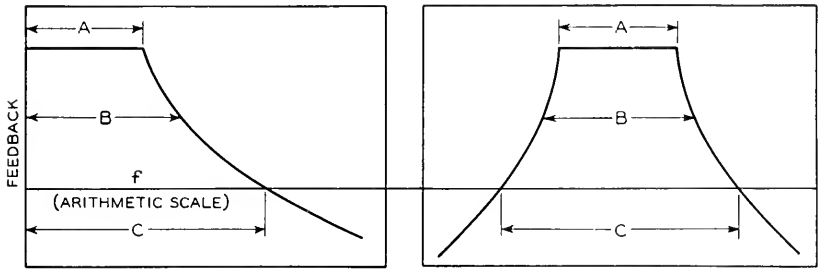


Fig. 22—Diagram to illustrate the conservation of band width in the low-pass to band-pass transformation with fixed parasitic elements. *A*, *B* and *C* represent typical corresponding intervals of equal breadth.

Example

An example showing the application of the method in an actual design is furnished by Fig. 23. The structure is a feedback amplifier intended to serve as a repeater in a 72-ohm coaxial line.[12] The useful frequency range extends from 60 to 2,000 kc. Coupling to the line is

[12] The author's personal contact with this amplifier was limited to the evolution of a paper schematic for the high frequency design. The other aspects of the problem are the work of Messrs. K. C. Black, J. M. West and C. H. Elmendorf.

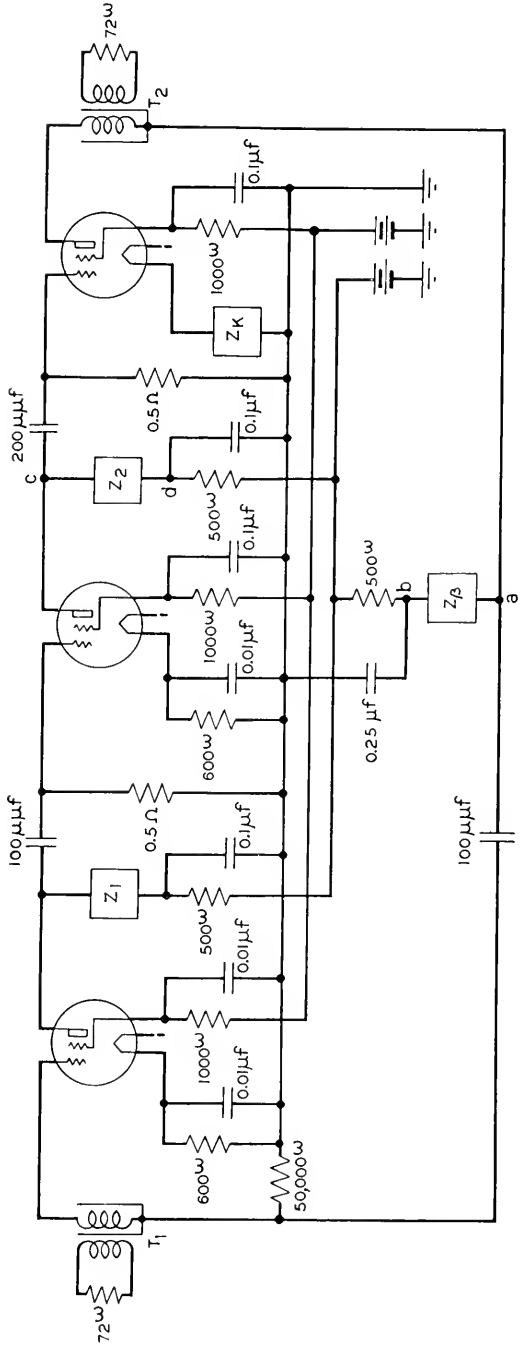


Fig. 23—General schematic of a 2 mc. feedback amplifier.

obtained through the shielded input and output transformers T_1 and T_2. The three stages in the μ circuit are represented in Fig. 23 as single tubes. Physically, however, each stage employes two tubes in parallel, the transconductances of the individual tubes being about 2000 micromhos. The principal feedback is obtained through the impedance Z_β. There is in addition a subsidiary local feedback on the power stage through the impedance Z_K. This is advantageous in producing a further reduction in the effects of modulation in this stage but it does not materially affect the feedback available around the principal loop.

The elements shown explicitly include resistance-capacitance filters in the power supply leads to the plates and screens, cathode resistances and by-pass condensers to provide grid bias potentials, and blocking condenser-grid-leak combinations for the several tubes. In addition to serving these functions, the various resistance-capacitance combinations are also used to provide the cutoff characteristic below the useful band. The low-frequency asymptote is established by the grid leak resistances and the associated coupling condensers and the approach of the feedback characteristic to the asymptote is controlled mainly by the cathode impedances and the resistance-capacitance filters in the power supply leads to the plates. The principal parts of the circuit entering into the $\mu\beta$ characteristic at high frequencies are the interstage impedances Z_1 and Z_2, the feedback impedance Z_β,[13] the cathode impedance Z_K, and the two transformers. The four network designs are shown in detail in Figs. 24, 25, 26, and 27.

The joint transconductance, 4000 micromhos, of two tubes in parallel operating into an average interstage capacity of 14 mmf, as indicated by Figs. 24 and 25, gives an f_t of about 50 mc. The parasitic capacities (chiefly transformer high side and ground capacities) in the other parts of the feedback loop provide a net loss, A_t, of about 18 db at this frequency. Since the asymptotic slope is 18 db per octave the intercept of the complete asymptote with the zero gain axis occurs about one octave lower, at slightly less than 25 mc. This is a relatively high intercept and may be attributed in part to the high gain of the vacuum tubes. The care used in minimizing parasitic capacities in the construction of the amplifier and the general circuit arrangement, including in particular the use of single shunt impedances for the coupling and feedback networks, are also helpful.

[13] The relative complexity of this network is explained by the fact that it actually serves as a regulator to compensate for the effects of changes in the line temperature. (See H. W. Bode, "Variable Equalizers," *Bell System Technical Journal*, April, 1938.) The present discussion assumes that the controlling element is at its normal setting. For this setting the network is approximately equal to a resistance in series with a small inductance. The fact that the amplifier must remain stable over a regulation range may serve to explain why the design includes such large stability margins.

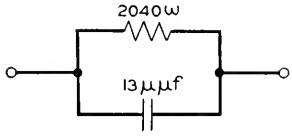


Fig. 24—First interstage for the amplifier of Fig. 23.

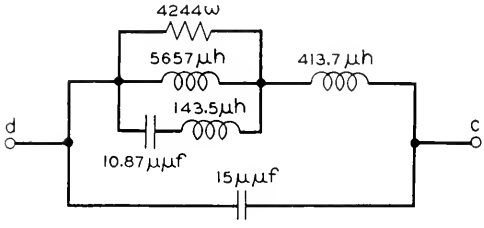


Fig. 25—Second interstage for the amplifier of Fig. 23.

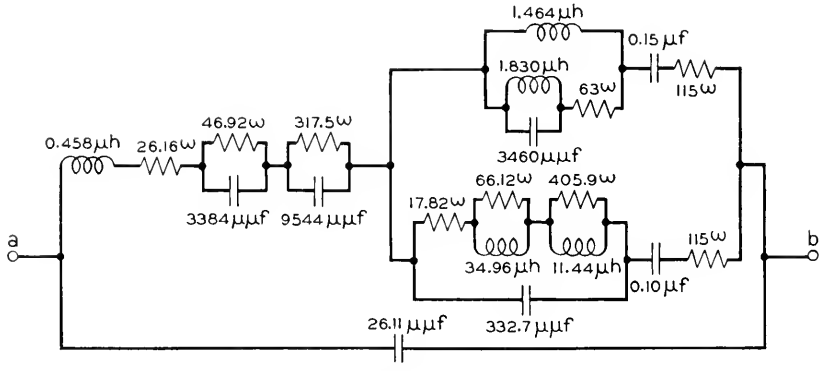


Fig. 26—β circuit impedance for the amplifier of Fig. 23.

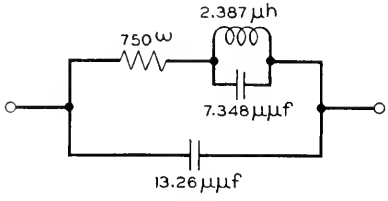


Fig. 27—Cathode impedance for the amplifier of Fig. 23.

In accordance with (7) the maximum available feedback A_m is 48 db. For design purposes, however, x and y in (9) were chosen as 15 db and 1/5 respectively. This reduces the actual feedback A to about 28 db. The theoretical cutoff characteristic corresponding to

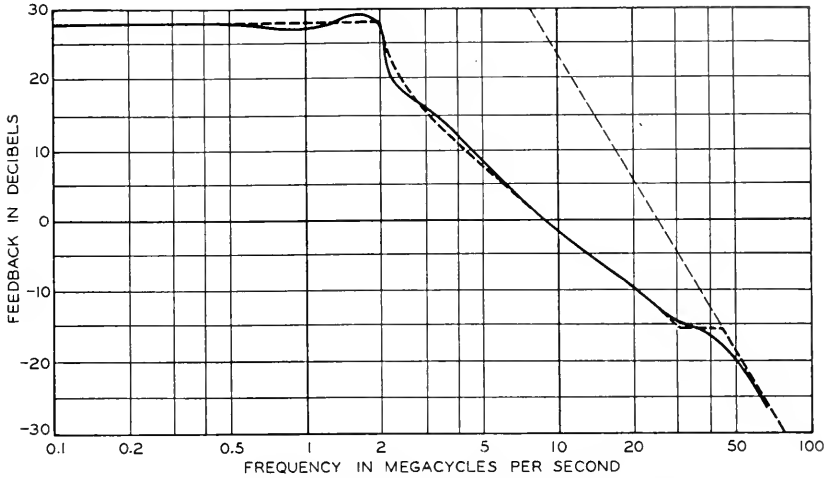


Fig. 28—Loop gain characteristic for the amplifier of Fig. 23.

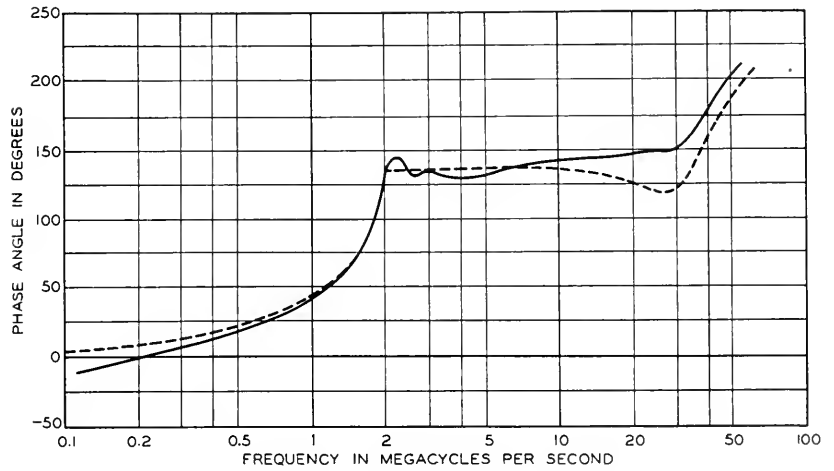


Fig. 29—Loop phase characteristic for the amplifier of Fig. 23.

these parameters is shown by the broken lines in Figs. 28 and 29, and the actual design characteristic by the solid lines. Since this is a structure in which the required forward gain is considerably less than the maximum available gain, the general course of the cutoff character-

istic is controlled, in accordance with the procedure outlined previously, by the elements in the μ circuit. The sharp slope just beyond the edge of the useful band is obtained from a transformer anti-resonance. The relatively flat portion of the characteristic near its intersection with the asymptote is due partly to an anti-resonance of the β circuit with its distributed capacitance and partly to an increase in the gain of the third tube because of the filter-like action of the elements of Z_K in cutting out the local feedback on the tube in this region.

The large margins in the design made it possible to secure a substantial increase in feedback without instability. For example, with a loss margin as great as 15 db the feedback can be increased by adjusting the screen and plate voltages to increase the tube gains. A higher feedback can also be obtained by adjusting the resistance in the first interstage. As this interstage was designed, an increase in the resistance results in an increased amplifier gain and a correspondingly increased feedback which follows a new theoretical characteristic with a somewhat reduced phase margin. The adjustment, in effect, produces a change in the value of the constant k in equation (6). With this adjustment the feedback can be increased to about 40 db before the amplifier sings.

Analysis of the Ionosphere *

By KARL K. DARROW

The ionosphere is a region in the very high atmosphere from which radio signals are reflected, a fact which is adequately explained by assuming that region to be populated with free electrons. In exploring the ionosphere, signals of a wide range of frequencies are successively sent upward, and the time elapsing before the return of the echo is measured. The delay of the echo mulitplied by $\frac{1}{2}c$ is called the virtual height of the ceiling for the signal. The data appear in the form of curves relating virtual height of ceiling to frequency of signal. These curves are peculiar in shape and vary remarkably with time of day, time of year and epoch of the solar cycle. By theory they can be translated into curves relating electron-density to true height above ground. The theory is approximative, but the results are accurate enough to be of value. The magnetic field of the earth affects the data remarkably, making it possible to test the theory and to evaluate the field-strength at great heights. The free electrons are supposed to be liberated from the air-molecules by ionizing agents, of which the chief but not the only one is ultra-violet light from the sun.

THE very title of this article embodies the assumption that in the upper reaches of the atmosphere there is a host of ions. By "upper reaches" here is meant, a region of the atmosphere so high that no man ever entered it, nor even a balloon with instruments. The ions therefore have never been observed by normal electrical means. They are postulated as the explanation of two things mainly: the echoing of radio signals from the sky, and that small portion of the earth's magnetic field which fluctuates with time.

The idea that these things require explanation, and the idea of the sort of postulate that is required to explain them, can both be followed back for many years. What was lacking in the early days was the notion of mobile electrified particles, that is to say, of "ions," in the air. That notion did not even exist, when in the eighties Balfour Stewart desired to imagine a conducting layer in the upper air for explaining magnetic fluctuations. It was only just being formed, when in 1902 Kennelly and Heaviside independently desired to imagine a conducting layer in the upper air for explaining why wireless signals can travel around the world. To speak of a conducting layer in the

* This paper, in abbreviated form, appears in the current issue of *Electrical Engineering*.

eighties, when air was regarded as an almost perfect insulator and nothing was known that could make it conductive, was certainly audacious. To speak of it in 1902 was still ingenious but no longer daring, for by then it could reasonably be expected that the researches on ions lately begun by Thomson and so many others would justify the notion.

Never was an expectation better founded. Within a few years those researches had made it sure that the upper atmosphere must be conductive, because of containing the raw material required for making ions and one at least among the agents capable of making them: to wit, atoms and molecules, and ultra-violet light from the sun capable of ionizing them. The problem then became: what distribution and what kinds of ions must be postulated for the upper atmosphere, to explain (for instance) the reflection of radio signals?

This problem could not even be attacked, without great forward strides in both the art of experimentation and the mathematical theory. These strides were rapidly made in the middle and late twenties. Had theory alone gone ahead, it would have been little more than a pretty exercise in mathematics. Had the art of experimentation progressed by itself, the experimenters would at least have found some interesting correlations of the data with such variables as time of day and epoch of the solar cycle and presence of magnetic storms; but the lack of theory would have been sorely felt. But theory and experiment advanced together, and the interplay between the two has seldom been so well exemplified.

The advance in the art of experiment lay not so much in the invention of new apparatus (though this has not been wanting) as in turning away from the practical problem of sending signals to great distances, and instead designing the experiments for the purpose, first of proving the ionosphere and then of "sounding" it. Three methods were invented for this purpose, all based upon the fact that wireless waves when sent into the sky come bouncing back from it. Two of these will be scarcely more than mentioned in the pages to follow, since an already great and ever-increasing proportion of the data is obtained by the third. In this third a sharply-delimited signal or pulse or wave-group is sent up, and a short time (a few milliseconds) later it is detected coming back, like an echo from a cliff: the delay of the echo is measured. This is done for many signals, and the delay is plotted against the mean frequency of the wireless waves composing the signals; and curves so plotted constitute the ultimate data. Usually the signal is sent vertically upward, the echo comes vertically downward; and there is the quaint situation, that wireless telegraphy

is chiefly famed for bridging great distances over the earth, but its foundations are best studied with sender and receiver side by side.

Electromagnetic signals thus find a mirror or a ceiling overhead; and the theory interprets this mirror as consisting of the ions, and especially the free electrons, diffused in the upper air. This perhaps seems singular, in view of the tenuity of the air and the lightness of the individual electrons. It might have seemed better, at least in the days of the Greeks, to propose that the dome of the sky is a hard metallic mirror—of well-polished silver, for instance. Well, in effect that is what *is* proposed. A mirror of silver reflects not by virtue of its hardness, but because of electrons diffused like a gas through the pores of the metal. A metal is a container for an electron-gas, and in the upper air there is an electron-gas without a container; and both of them reflect.

The theory is strictly classical, in the sense of the word prevailing in physics. No relativity, no quantum theory, no suggested revision of the concepts of space and time, afflict the student thereof. It is the working-out of the basic principle of Maxwell and Lorentz, that the passage of electromagnetic waves through a medium is controlled by the electric current which the waves themselves evoke in the medium. Under the influence of the electric field in the waves, the ions swing in sympathetic vibration, and form a part of that current. They thus react upon the waves, alter the speed thereof, and bring about the reflection. The motion of the ions is simple-harmonic, so that the mathematics of the theory is simple and familiar—so long, at least, as no account is taken of any forces acting on the ions other than that due to the field of the waves themselves. Here is the explanation of the echoing of radio signals, and hence follows the procedure for translating the data of echoes into statements about the distribution of the ions in the atmosphere. It is not difficult to describe or explain, and will be carried through in this article.

From this point the theory ramifies in two directions. Two things modify the sympathetic vibrations of the ions: the collisions between the ions and the neutral molecules of the air, and the earth's magnetic field. By their influence on the vibrations, they modify the speed of the waves, and therefore the conditions of the echoing. The theory extended in either direction continues to be easy in one sense, for neither the physical concepts nor the mathematical operations are unfamiliar; but becomes very hard in another, for the algebraic expressions are often of fearful complexity, impossible to remember and hard even to keep straight when written out. When it is extended in both ways at once the expressions become so intricate, that nearly

every investigator when taking account of either influence simply ignores the other. On the whole, the mathematical developments have far outrun the data. Yet there are important connections between experiment and theory, including for instance the proof that the ions which principally reflect the signals are free electrons.

After the theory come what I will call, for contrast, the speculations. The analysis of the ionosphere being made and accepted, a host of questions arise. Must we assume additional agents of ionization, other than the ultra-violet light of the sun? The answer to this question being certainly "yes," one must inquire how to distinguish that part of the ionization which is due to sunlight from the rest, and what are the causes of the rest. Why the distinctive distribution-in-height of the ions, amounting to what is called "the stratification of the ionosphere"? What assumptions must we make about the composition of the atmosphere in its dependence on height? or (as the question is more commonly put) what information can we derive about the composition of the atmosphere? How far can we go in interpreting the fluctuations of terrestrial magnetism, and (as later will be apparent) in mapping out the earth's magnetic field? The possible questions even rise to the realm of astronomy, and the suggested answers form a part of the theory of the sun as a potent source of radiations of all kinds, luminous and electrical and material. The implications of the ionosphere seem to be almost limitless, but a severe limit will nevertheless be set by space and time upon this article.

Methods of Experiment, and a Simplified Picture of the Ionosphere Adduced for Illustrating Them

The ionosphere is a canopy of ions overarching the earth, and in Fig. 1 it is represented by a model, very simplified indeed and yet instructive. Here it is shown as consisting of two "layers" marked E and F, with an ion-density which is uniform in each, and greater in F than in E. It is time to become familiar with the symbol N used for number of ions per unit volume: this picture shows N having the constant values N_E and N_F ($> N_E$) in E-layer and F-layer respectively, and the value zero between.

The lines which are broken at the layer-edges are paths of wireless signals or waves sent out from the source at S—sent out obliquely, for transmission over long distances. There is a path reflected from E, a path reflected from F and a path which penetrates both. These correspond to relatively low, medium, and high frequencies respectively: as examples I will give the values 1, 10 and 100 mc. (megacycles,

i.e. millions of cycles per second) corresponding to wave-lengths of 300, 30 and 3 metres. Here already the reader meets the fact that the height at which such a signal is reflected, or the question whether it shall be reflected at all, depends on the frequency of the waves and the density of the ions. For every frequency there is what I shall call a "mirror-density": signals are reflected as soon as they reach the lowest level in the ionosphere where that mirror-density is attained. The higher the frequency, the higher the mirror-density. The formula

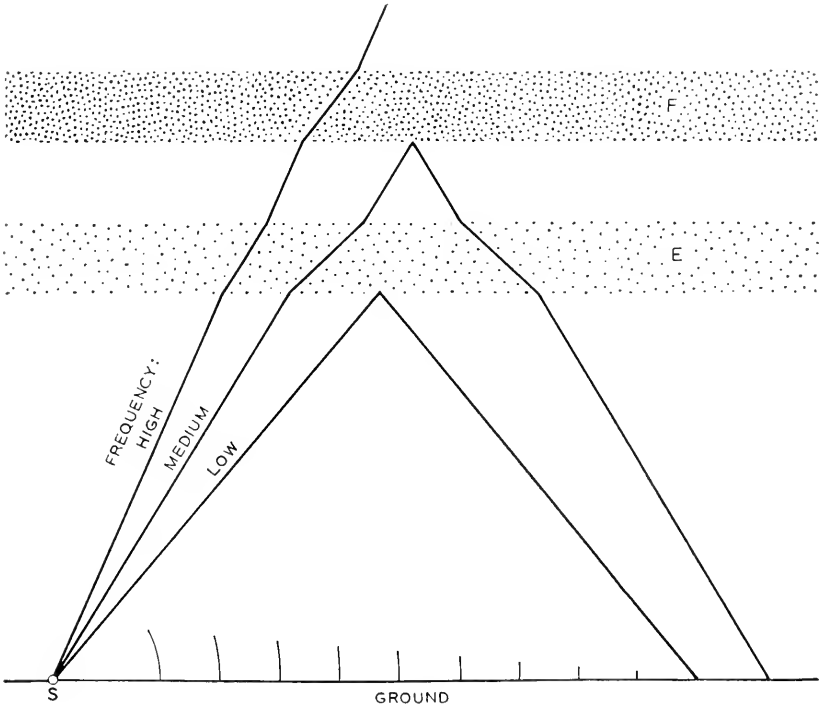


Fig. 1

will soon be derived and shown; but for the moment, let us inquire how the altitude of either layer can be measured, in terms of the simple model of Fig. 1.

One way of measuring the altitude is very obvious. Suppose the observer to go a known distance away from the aerial, and measure there the angle which the reflected wave or "sky wave" makes with the horizontal as it comes down to him. If this can be done, then clearly he can get the altitude by the simplest trigonometry—the altitude of the *E*-layer or the *F*-layer, according to the frequency

which he uses. It *can* be done, and it was done by Appleton and Barnett in 1925. What they measured was the angle between the directions of propagation of the sky wave and the "ground wave," of which the wave-fronts are shown in Fig. 1 creeping along the ground with a rapid attenuation. This ground wave, by the way, is the only one by which radio transmission could be effected but for the ionosphere; and it is seldom detectable beyond a few hundred miles.

Another scheme is much more complicated, and owing to its supersession I may be excused for giving only the merest outline of it. It is a clever way of putting to useful service the very great inconvenience known as "fading." This term refers to a train of signals which dies out and revives and keeps on fluctuating over and over again, in a most irregular fashion. This sort of thing occurs in the region where the sky wave and the ground wave both arrive and overlap one another, and it has been traced to what in optics is called the "interference" of the two. If conditions were absolutely stable, then in taking a walk in the region of overlapping one would pass through several maxima and minima of intensity. Since conditions are never absolutely stable, the observer need not take the walk; while he stands at any fixed point, the maxima and the minima float past him while the ionosphere wavers in the sky and this is "fading." But imagine the conditions relatively stable for a time, and the observer standing still; and suppose that the engineer at the sending aerial changes the wave-length by a small and known amount—then, several maxima and minima will float past the observer, and by counting them he can (though this is not at all obvious!) get a datum which enables him to figure out the altitude of the reflecting layer in the sky. This method also was invented by Appleton, and can be found explained in the literature under the name "wave-length-change method."

The third of the methods has crowded out the others, and henceforth will figure alone in these pages. It is the "echo-method," still sometimes called by the clumsy name of "group-retardation method." Anticipated by Swann, it was realized by Breit and Tuve at the Carnegie Institution of Washington.

What is sent up to the sky is here a short sharp signal; if it could be heard, it would be called a click. What comes back is the echo of the signal. Passing over the receiving device, it produces a short sharp kick on an oscillograph-record. Some of these are shown in Fig. 2. The kicks marked E_1 and F_1 are due to signals echoed from the layers E and F respectively. Those marked $F_2 \cdots F_5$ are due to

"multiple echoes"; the signal has traveled two to five times the entire journey from ground to ionosphere to ground again, the surface of the earth being itself a good reflector. Those marked G are due to the signal spreading along the ground itself. If the sender and the receiver are practically side by side, as usually is the case, the kicks G occur at the instants of departure of the signals. The record is moving laterally with the speed intimated by the wavy line beneath, and accordingly the distance along it from a G-kick to the following echo-kick is a measure of the "delay of the echo."

The delay of the echo is an indication of the altitude of the mirror where it was reflected—the layer E or F, as the case may be. Signals of relatively low frequency being reflected from E while those of medium frequency are echoed at F, one adjusts the frequency according

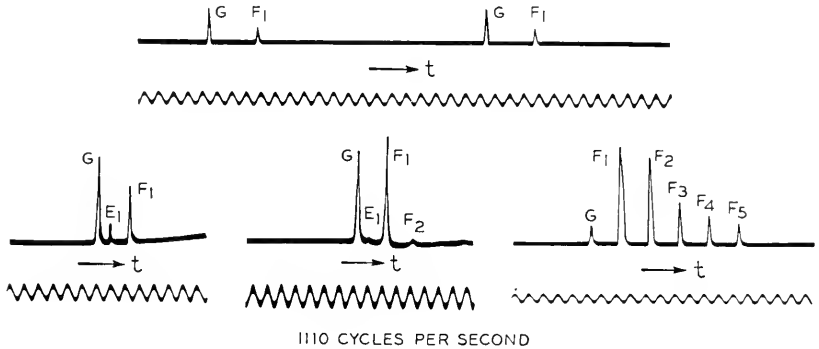


Fig. 2—Echoes. G, original signal; E_1, F_1, echoes returning after a single reflection from E and F respectively; $F_2 \cdots F_5$, echoes which have suffered two to five reflections at F-layer. (Appleton and Builder.)

to the layer which one wishes to locate. If there should be not two but several layers of the ionosphere, each having a greater N-value than the one beneath it, one would locate them all with appropriate frequencies. If there is a continuously-rising distribution of N with height in the ionosphere, one may plumb it by varying the frequency continuously. Now we are at the principle of the echo-method; but before it is used, there are many details to clear up.

First as to the "signals," a term which (it must have been noticed) replaced the term "waves" in the foregoing paragraphs. These signals are wave-trains indeed, but not the long continuous uniform trains tacitly assumed in the description of the other two methods. Those methods are adapted to trains of indefinite length, but not the echo-method, which for an obvious reason requires wave-trains of limited length—and the more limited, the better.

Now, a limited wave-train is equivalent to an infinity of infinitely long wave-trains, with an infinite variety of frequencies. The amplitudes and the frequencies of these constituent waves are chosen so that the waves reinforce one another over the length of the signal, counteract one another for all time before and after the signal. To choose them thus is always mathematically feasible, whatever shape of signal be prescribed. Whether these constituent waves should be regarded as "physically real" is a question that was discussed long before the days of quantum theory and other modern puzzles. Anyhow, by taking them as such, one arrives at verifiable statements about the signals, and this is all that matters.

Let us now conceive the signal as a chopped-off segment of a sine-wave-train of frequency f_0; and let us compare its travel with the travel of a limitless wave-train of frequency f put equal to f_0. One feels that the signal ought to follow the same path through the ionosphere as would the limitless train, and ought to move along that path with the same speed as would the wave-crests of the limitless train. This is true if, but only if, the speed u of the wave-crests in unlimited trains is independent of f. But when wireless waves are traveling through the ionosphere, u varies very much with f, according to a law which will later be worked out; and this makes a remarkable difference.

The difference as to path is not serious. The infinite wave-trains which form the signal are most intense at frequencies very close to f_0, and this is sufficient to make the signal follow nearly (though not without some deviation and distortion) the path which the infinite train of frequency f_0 would follow by itself. We may therefore regard the broken lines of Fig. 1 as the paths of signals or of waves, indifferently.

The difference as to speed *is* serious. It is not reduced by the preponderance of component wave-trains very close to f_0, and it does not tend to vanish as this preponderance is increased by lengthening the signal. It remains serious in the artificially-simplified case of just two component wave-trains of small frequency-difference Δf, where the signals become the "beats" well known in acoustics and in radio. In this case the beat-speed or signal-speed approaches a limit as Δf approaches zero. This limit, the "group-speed" denoted by v, is always used for the signal-speed, though for actual signals it is but an approximation, and the signals themselves become distorted as they travel. Contrasted with v is the "wave-speed" or the "phase-speed" of the wave-crests of the unlimited wave-trains, already denoted by u.

Since it is entirely the signal-speed which determines the delay of the echo, the wave-speed may seem a pointless side-issue. This quantity u is, however, essential in the theory from which are derived first the conditions of echoing and then the dependence of v upon f_0. Postponing the topics of signal-speed and echo-delay in order to return to them later with better preparation, we now take up the theory.

Theory of Wave-Speed, Total Reflection and Group-Speed in the Ionosphere

It will now be proved from Maxwell's theory, combined with the concept of mobile ions, that total reflection of wireless waves must occur in the ionosphere at the level where the ion-density attains a certain value depending on the wave-frequency.

The famous equations of Maxwell melt together into a wave-equation. The waves which it describes consist of an oscillating electric field which I will denote by $E_0 \sin nt$, and an oscillating magnetic field which we are permitted to ignore. When of high enough frequency these are the waves of light, as Maxwell knew; when of the frequency-range with which we are now concerned they are the waves of radio, as Maxwell was never to know because of his premature death. In the wave-equation there figures of course the wave-speed u. Here then is a paraphrase of the great idea of Maxwell: *the square of the wave-speed varies inversely as the current-density provoked by unit amplitude* [1] *of the oscillating field.*

Now we see at once that in the ionosphere the wave-speed must be affected by the presence of the free electrons, since they are set into oscillation by the waves and therefore make a contribution to the current-density.

At this point those who were educated in the electronic era (an ever-increasing fraction of the population) are in some danger of falling into a serious error. One may in fact assume that the electrons form the whole of the current, and deduce that in vacuo the oscillating field provokes no current at all, and the wave-speed must therefore be infinite—an absurd conclusion! Maxwell was wiser. He understood, and made it a part of his theory, that wherever there is an electric field which is changing in time the rate-of-change of that field is equivalent to a current. This he called the "displacement-current," and for the case of vacuum he said that the displacement-current-density is precisely equal to the rate-of-change of the field, multiplied by $1/4\pi$.

[1] I introduce the words "unit amplitude" to shield the reader from drawing the false inference that wave-speed depends upon wave-amplitude.

In vacuo, therefore, there is a current-density $(1/4\pi)$ times the rate-of-change of $E_0 \sin nt$, and it has an amplitude of $(nE_0/4\pi)$ and is 90° ahead of the field in phase. To this current-density corresponds the speed of light in vacuo, the well-known constant c. The speed of light in the non-ionized lower regions of the atmosphere differs so little from c that we need never bother with the difference, which henceforth will be ignored.

When the waves pass out of ordinary air into the ionosphere, there is still the displacement-current but now in addition there is the current borne by moving electrons. Here is a second pitfall. It may seem obvious that the electron-current must add on to the displacement-current, creating a total current-density greater than that in vacuo and therefore lowering the wave-speed. Not so at all! The point is, that when the electrons are truly free, the field sets them into oscillation in such a curious way that when they become adjusted, they are oscillating with their velocities 90° behind the field in phase. Their contribution to the current, being proportional to their velocity, is also 90° behind the field, and hence in perfect opposition of phase to the displacement-current.

Therefore the electron-current density—call it I_e—is to be *subtracted* from the displacement current-density! Accordingly I write,

$$\frac{u^2}{c^2} = \frac{n(1/4\pi)E_0}{n(1/4\pi)E_0 - I_e}. \tag{1}$$

The reader may suppose that the factor $\cos nt$, common to both currents, has been divided out.[2] The quantity I_e is clearly proportional to E_0 and also to our familiar N the density of electrons, and in fact the reader can undoubtedly work out with ease that it is equal to NE_0e^2/mn. Here e and m stand for the charge and mass of the ion, as is customary. Therefore we find:

$$\frac{u^2}{c^2} = \frac{1}{1 - 4\pi Ne^2/mn^2} = \frac{1}{1 - Ne^2/\pi mf^2}. \tag{2}$$

The wave-speed is greater in the ionosphere than it is in vacuo or ordinary air. I now recall from the most elementary optics the principle that when two media adjoin in which light has different wave-speeds, and light passes through their common boundary into the medium where its speed is greater, it is refracted away from the normal to the boundary. Accepting for the moment the over-simplified model of the

[2] Actually Maxwell's theorem does refer to the amplitudes of the currents—but if the currents are not exactly 0° or 180° apart in phase, the amplitude of one must be taken as a complex quantity.

ionosphere in Fig. 1, and considering the lower frequencies, we have the non-ionized lower atmosphere and the E-layer for these media. The paths of the waves are drawn accordingly. Now I further recall that total reflection occurs for all values of the angle of incidence i greater than that given by the equation:

$$\sin i = c^2/u^2. \tag{3}$$

Thus we see that for any frequency whatever, total reflection must occur when the waves impinge with sufficient obliqueness upon the ionosphere; but (so long as c/u does not sink to zero) total reflection will not occur if the waves rise vertically, or in a direction sufficiently near to the vertical.

The waves thus penetrate or are reflected back from the ionosphere, according as their angle of incidence thereon is less or greater than a certain critical value. Here is the explanation of what is called "skip-distance": the sky-wave is perceived beyond a certain distance from the source, but not within that certain distance.[3]

But all this seems to have nothing to do with the usual conditions of experiment, in which, as I intimated, the signals are sent up vertically! It is indeed a fact that in optics, no case is known in which total reflection occurs at vertical incidence. Yet equations (2) and (3) predict that if ever c^2/u^2 should vanish, total reflection would extend even to vertical incidence. Now there is nothing mathematically impossible or physically unplausible about the condition for the vanishment of c^2/u^2, which is simply that f should be equal to f_c given thus:

$$f_c^2 = Ne^2/\pi m \tag{4}$$

or alternatively that N should be equal to N_c given thus:

$$N_c = \pi m f^2/e^2. \tag{5}$$

Here we have the basic formula of the analysis of the ionosphere; for it is assumed that vertically-rising waves or signals of any frequency f climb until they reach the lowest level at which N is equal to N_c, and there they find their mirror or their ceiling, and are converted into echoes which return. Equation (5) is the formula for the "mirror-density" for signals of frequency c, to which I above referred.

It sounds all right to say that c^2/u^2 is zero when $N = N_c$, and negative when $N > N_c$; but it is disconcerting to notice that this

[3] Notice incidentally that owing to the curvature of the earth and its overhanging ionosphere, the angle of incidence can never rise to 90°; it follows that waves of frequency beyond a certain value (ordinarily around 30 mc.) never suffer total reflection.

amounts to saying that the phase-speed is infinite when $N = N_c$, imaginary when $N > N_c$. However, the concept of phase-speed is of such a quality of abstractness, that even these statements imply nothing absurd in the physical situation. The signal-speed itself remains safely finite and real.

The signal-speed is strictly indefinite, since the signal distorts itself as it proceeds. However, the practice is to identify it with the group-speed v, which, as I intimated (page 462), is the speed of the beats formed by two superposed wave-trains differing infinitesimally in wave-length, each such beat being a very special type of signal. The formula is,

$$v = u - \lambda(du/d\lambda) = u \Big/ \left(1 - \frac{n}{u}\frac{du}{dn}\right). \tag{6}$$

It is difficult to visualize or derive without a diagram,[4] but the derivation may be summarized as follows. Imagine two superposed wave-trains of phase-speeds u and $u + du$, wave-lengths λ and $\lambda + d\lambda$; consider two consecutive wave-crests A, A' of one and two consecutive wave-crests B, B' of the other; transpose temporarily to a frame of reference in which the former wave-train is stationary. At a certain place and time A and A' will coincide, and the maximum of one of the beats will be right there. Let the time $d\lambda/du$ elapse; when it has elapsed, the crests B and B' will be coinciding and the maximum of the beat will have moved on by one entire wave-length. The beat therefore travels with speed $\lambda du/d\lambda$ in the temporary and with speed $u - \lambda(du/d\lambda)$ in the original frame of reference (the minus sign is evident when the reasoning is gone through in detail).

Combining (6) with (2) one finds:

$$v = c^2/u; \tag{7}$$

the greater the phase-speed, the slower the signal! Relativists will be pleased to observe that according to this formula, the signal never attains any speed greater than c; students of quantum mechanics may be misled by its superficial resemblance to a formula relating phase-speed to group-speed for de Broglie waves, with which it has nothing to do. Students of the ionosphere should remember its approximative character. Almost all that needs to be known for the purposes of this article is, that as a signal climbs into the ionosphere it goes more and more slowly, the nearer N approaches to that value N_c where the signal finds its ceiling.

[4] Cf. this journal, 9, 173 (1930), or my *Introduction to Contemporary Physics*, 2nd edition, p. 147.

Characteristic Curves of the Ionosphere: the (h', f) Curves

Now that we have the concept of a signal ascending until it reaches the ceiling where $N = N_c$, we will consider first the to-be-expected relation between true height of ceiling and frequency of signal, then the relation between delay of echo and frequency of signal. By following this order we pass from the unobservable to the observed, which is the reverse of the customary way, but nevertheless has its advantages. Let z stand for height over ground when used as independent variable, with N depending on it; h for the height of the mirror or ceiling for signals, when expressed as a function of the signal-frequency f.

Take first the oversimplified model of the ionosphere appearing in Fig. 1: N having the values N_E over one range of heights and N_F ($> N_E$) over another range at a higher elevation, and the value zero elsewhere. It is evident that the (h, f) curve for such an ionosphere would consist of two horizontal lines or "branches," extending respectively from abscissa 0 to abscissa $f_E = \sqrt{N_E e^2/\pi m}$ and from abscissa f_E to abscissa $f_F = \sqrt{N_F e^2/\pi m}$ respectively. The latter would lie higher than the former; there would be a jump or gap between the branches. The names "E-branch" and "F-branch" for these last are obvious, and so is the usage "penetration-frequency of the E (or F) layer" for f_E or f_F; "critical frequency" is also used.

Next we approach closer to the truth by supposing that N rises continuously with increase of z across the E-layer and also across the F-layer, N_E and N_F now representing the highest N-values found in the respective layers. The two branches of the (h, f) curve would then be no longer horizontal, but slanting or probably curving upwards toward the right.

In the foregoing paragraph it was tacitly assumed that N still vanishes between the layers; but now let us approach still closer to the truth by postulating the sort of dependence of N on z shown in Fig. 3A. Here N drops with further increase of height after the "crown" of the E-layer is reached, but it does not fall to zero. It might, however, just as well fall to zero so far as reflections are concerned, for the signals which could be reflected from these regions never reach them. Regarding the curve of Fig. 3A as a sequence of hills and valleys, we see that the valleys contribute no echoes. The E-branch of the (h, f) curve refers to the left-hand side of the first hill; the F-branch refers to the left-hand side of the second hill, and not even to all of that, but only to the portion which rises above the first hill. Thus Fig. 3B, with its upturning branches and its gap, represents the (h, f) curve for the ionosphere of Fig. 3A, without in

the least depending on the dashed parts of the $N(z)$ curve of Fig. 3*A* or indicating anything whatever about those parts except that they do not rise above the ordinate N_E.

Now if the signal and the echo traveled fro and to with the speed c, the delay T of the echo multiplied by $\frac{1}{2}c$ would be the height of the ceiling. This, however, is not the case, since the signal-speed depends on N. We must therefore denote the product $\frac{1}{2}cT$ by another symbol h', and make an inquiry into the probable dependence of h' on f, taking into account our vague knowledge as to the dependence of signal-speed on N.

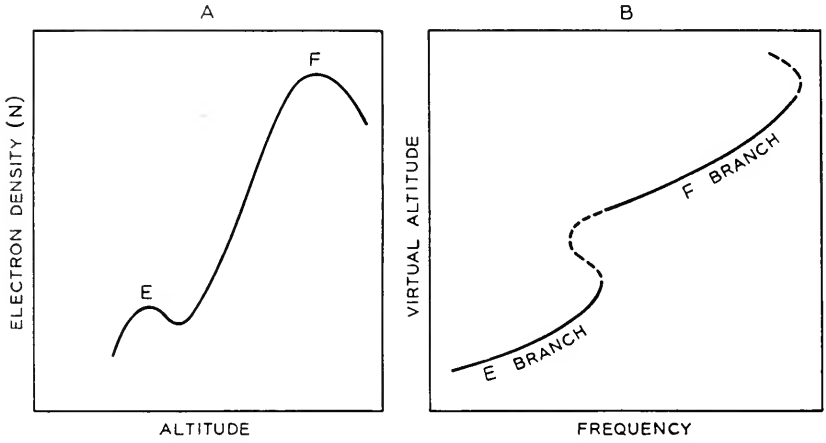


Fig. 3—*A*. The "curve of inference": conjectural dependence of number N of electrons per unit volume on true altitude h. *B*. The "curve of data": dependence of virtual altitude h' of ceiling (one-half the delay of the echo, multiplied by c) on frequency f.

It is easily seen that h' must be greater (or at least no less) than h, and that the excess of h' over h must be larger, the farther the signal travels through regions where N is almost but not quite equal to N_c. The (h', f) curve must therefore lie above the (h, f) curve, and farthest above it in the immediate neighborhood of the gap on both sides. There will still be an E-branch and an F-branch, but the upturns toward the right-hand ends of these branches will be exaggerated, and an upturn running to the left will be introduced into the left-hand end of the F-branch. It is conceivable that these upturns may become so large, that the (h', f) curve will appear to show a peak where the (h, f) curve would show a gap.

With the remark that h' is known as "virtual altitude," "virtual height," "equivalent height," or "effective height," I turn now to examples of the characteristic (h', f) curves of the ionosphere.

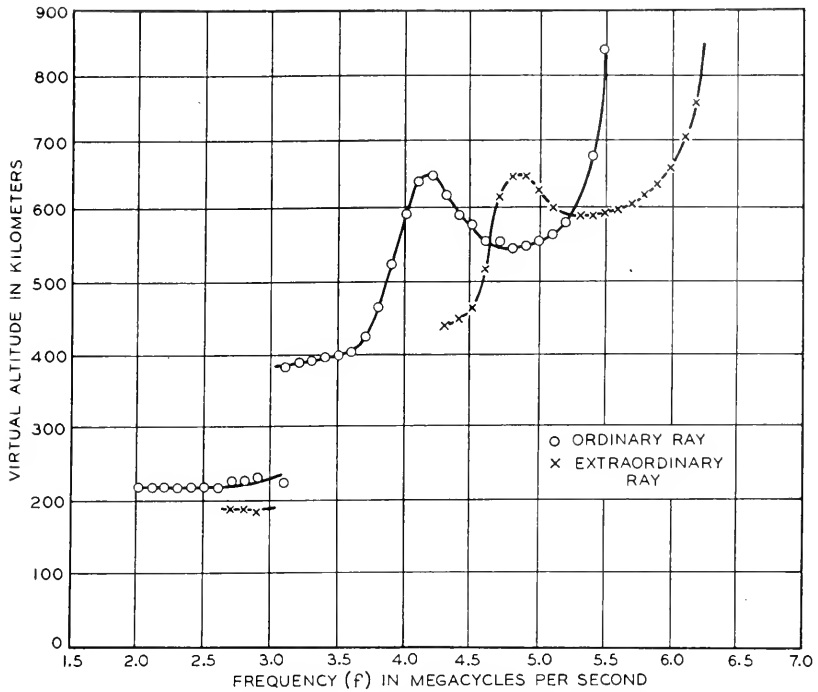


Fig. 4—Example of characteristic or (h', f) curves, showing gap between E-branch and F-branch, and crinkle in F-branch indicating presence of F_1-layer. Duplication of curve due to earth's magnetic field (page 479). (Appleton.)

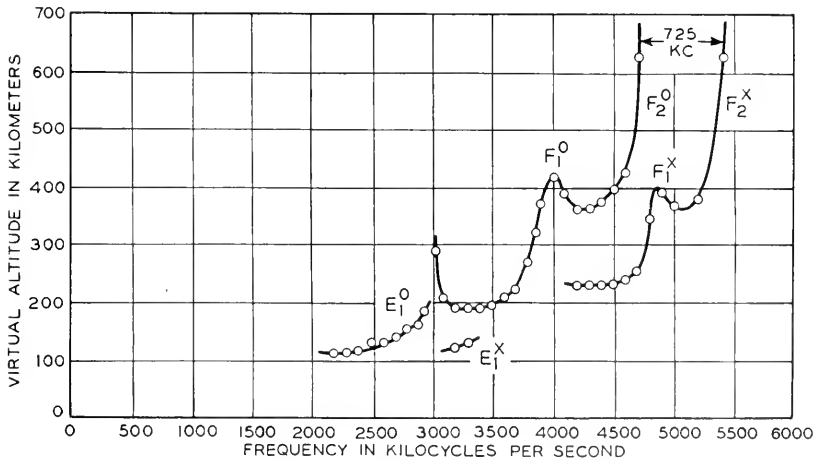


Fig. 5—Another example of (h', f) curve, showing gap and crinkle. (Schafer and Goodall.)

Figures 4 and 5 show two examples of these curves, from data obtained while the sun was high in the sky. Actually there are two curves in each of the figures; the appearance is that of a single curve, repeated with a sidewise shift. I mention that this repetition is due to the earth's magnetic field, but ask the reader to ignore for the present the right-hand curve and fix his attention on the left-hand one. Here he will see the E-branch, the gap, and the F-branch. The

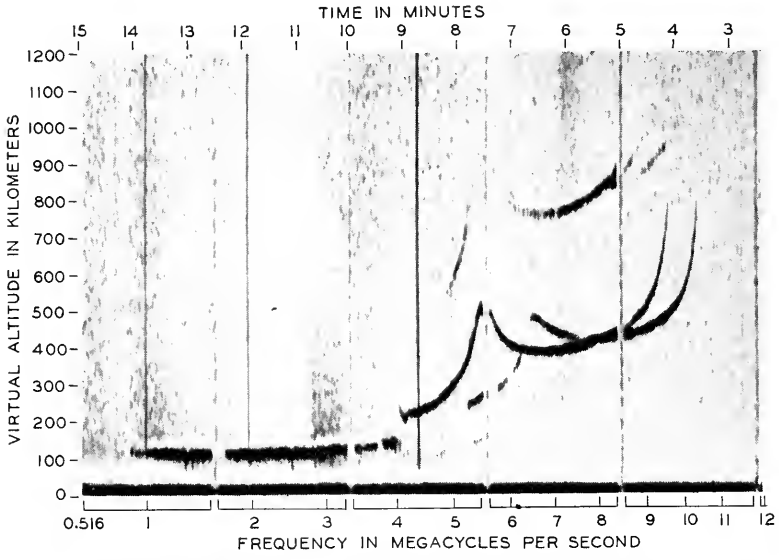


Fig. 6—Characteristic (h', f) curves obtained with the multi-frequency apparatus; sun high in sky, F_1 crinkle apparent. (Carnegie Institution of Washington.)

upturns to right and left of the gap are striking on Fig. 5, insignificant in Fig. 4. The F-branch is deformed by an enormous hump or crinkle. This is supposed to correspond to a second gap, the upturns on right and left being so pronounced as to give a perfect semblance of a peak; indeed one sees in Fig. 5 how readily the gap between E and F might have been drawn as a peak. Curves of this sort are therefore taken as evidence for three layers in the ionosphere, denoted by E and F_1 and F_2. Sometimes there are signs of a fourth, lying between E and F_2, and denoted by M or E_2.

So great is the interest in curves like these, and so much do they vary from time to time and from place to place, that lately there have been more than a score of stations over the world engaged in making them. At some of these the tracing of the curves is speeded up and made incessant by a remarkable machine developed at the

Bureau of Standards and the Carnegie Institution. Automatically sending out the signals ten times in a second, and changing the frequency by (on the average) 1600 cycles between each signal and the next while the photographic film is moved a tiny bit from left to right, this "multi-frequency apparatus" traces the (h', f) curve over the

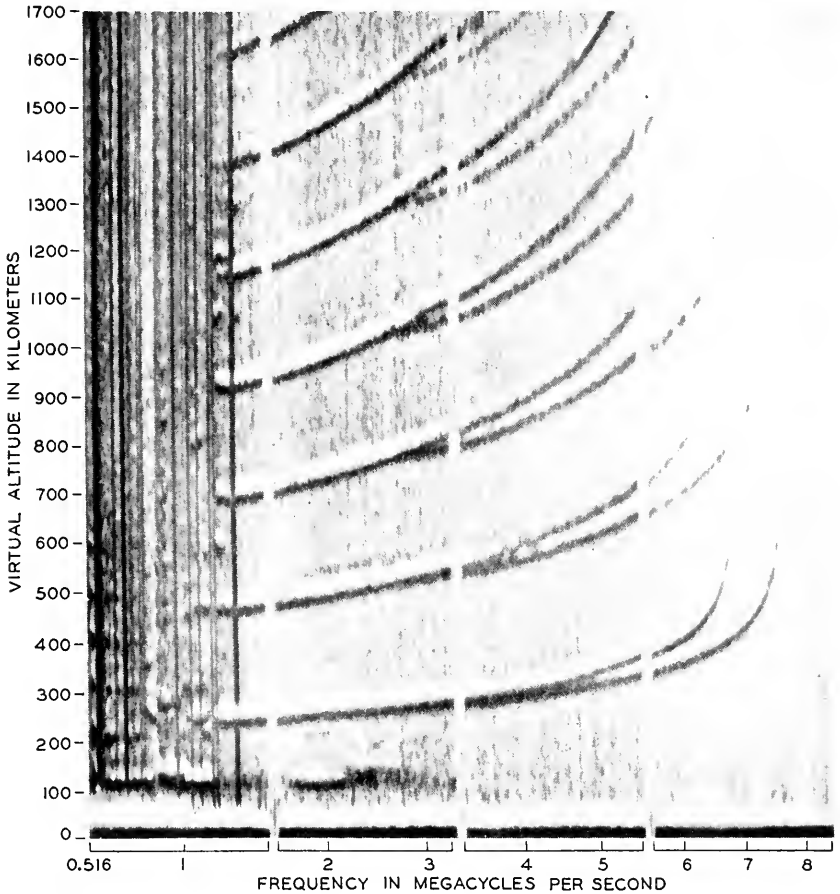


Fig. 7—Characteristic (h', f) curves obtained with the multi-frequency apparatus; sun low in sky, F_1 crinkle missing. (Carnegie Institution of Washington)

frequency-range between 0.516 mc. and 16 mc. in fifteen minutes, and then goes right back and does it over and over again. Figures 6, 7 and 8 show individual curves thus automatically taken, and Fig. 15 a sequence of them spanning several hours of the day.

In Fig. 6 are curves with crinkles in the F-branch, similar to those of Figs. 4 and 5. In Fig. 7, however, the crinkle is missing, and the

F-branch sweeps smoothly and slowly upward from its commencement. (The forking signifies that here are two similar curves lying side by side as in the previous figures, but overlapping so much that over a large part of their course they are not distinct.) Many observations have concurred in showing that the crinkle is present only when the

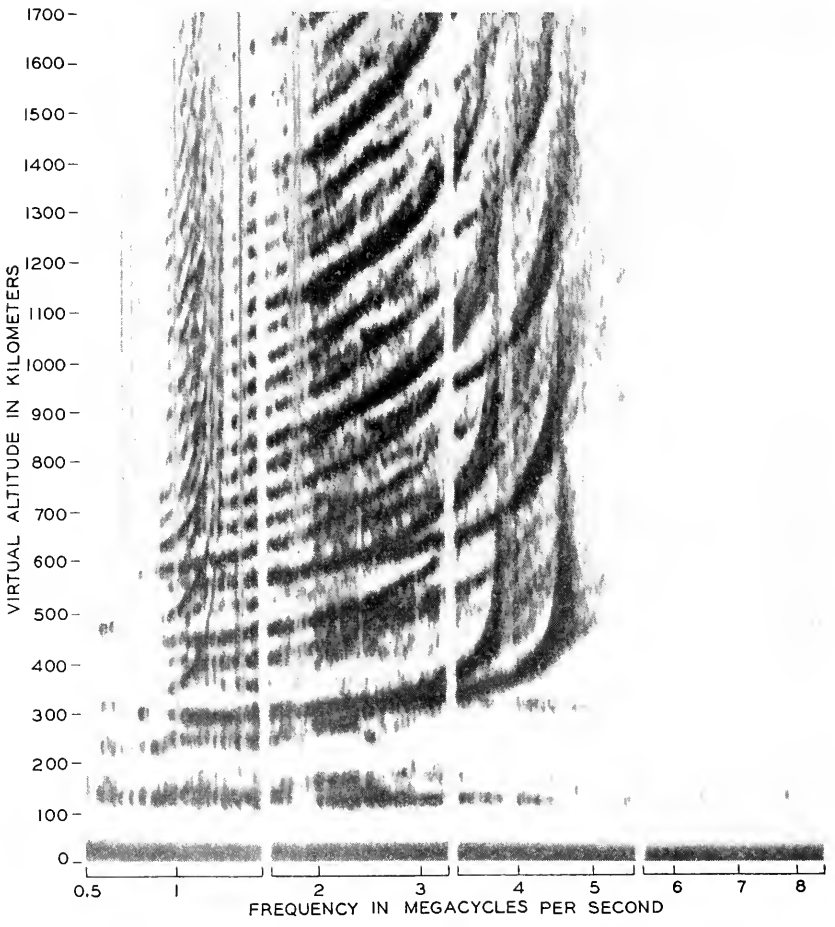


Fig. 8—Multiple echoes. (Carnegie Institution of Washington.)

sun is high in the sky (within some 40° of the zenith)—therefore absent by night and at the beginning and end of day, and indeed absent all day in winter where the latitude is high. This is our first example of the dependence of the ionosphere on sunlight, a very important feature.

In all of these photographs the curves are repeated several times along the vertical direction. This signifies echoes which have traveled four, six, eight or more times between ground and ionosphere, being reflected by both. Figure 8 shows a wonderful multitude of such echoes.

The curves of Fig. 9 are sketches generalized from many data. Contrasting those on the left with those on the right, we see the

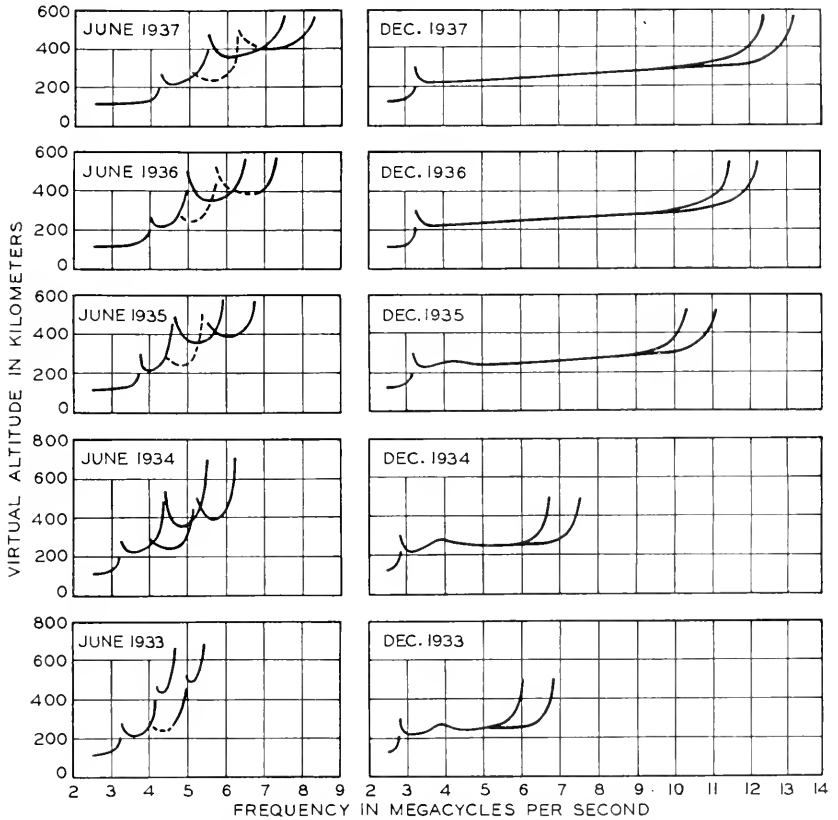


Fig. 9—Dependence of (h', f) curve on season and on the sunspot cycle (Smith, Gilliland and Kirby: National Bureau of Standards).

crinkle prominent in the ones, missing or feeble in the others. This is the difference between summer and winter. All of the data were taken near noon, but though the District of Columbia is not exactly in polar latitudes, the sun in December does not rise far enough in Washington's sky to bring out that feature of the ionosphere of which the crinkle is the sign. Running the eye along either column, one

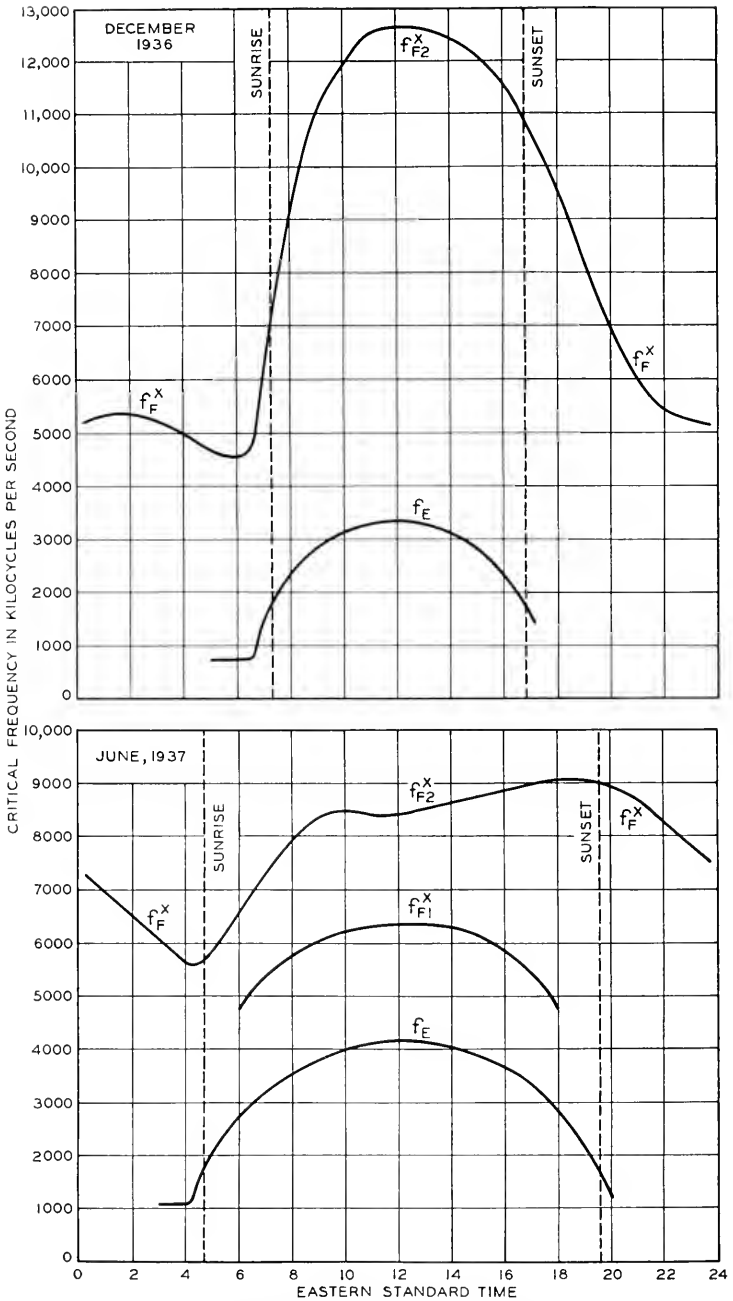


Fig. 10—Dependence of critical frequencies on season and hour of day. (Smith, Gilliland and Kirby.)

sees the curve lengthening out to the right as year follows year. During this series of years the sun spots were growing more frequent: the sun was in the ascending part of that eleven-year cycle of its fever, of which the sun spots are one of the manifestations, while the form of these curves is another. Do not, however, misread this statement as meaning that the curve lengthens out, when and only when there are sun spots on the solar disc! there is no such correlation.

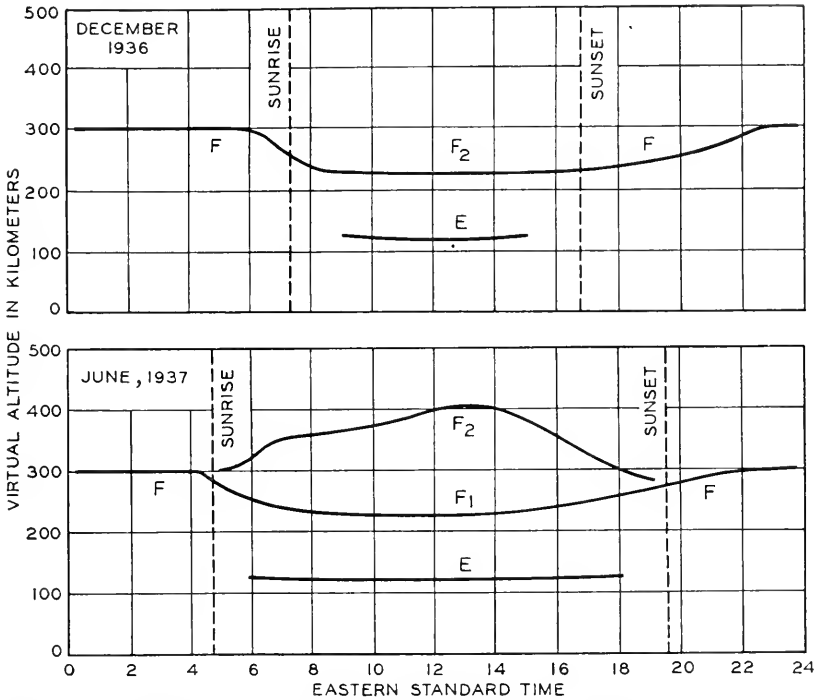


Fig. 11—Dependence of virtual altitude of the crowns of the principal layers on season and hour of day. (Smith, Gilliland and Kirby.)

Another way to study dependence on the sun is to pick out salient features of the (h', f) curves, and see how they vary. Such features are the abscissae and the ordinates of the right-hand ends of the branches and of the top of the F_1 crinkle. The abscissae are the penetration-frequencies for the layers E, F_1 and F_2; the ordinates are the virtual heights (not the true heights) of what I have called the crowns of these layers. All are shown, in their dependence on season and hour of day, in Figs. 10 and 11.

It may have struck the reader that for the last five paragraphs there has been no allusion to the theory, the correlation of the (h', f)

curves with the sun having been presented as if for its own sake. Much work in the field does stop at this point, and is not without value in spite of its stopping there. The theorist indeed may care for an (h', f) curve only as material for deducing the (N, z) curve—the distribution-in-height of the ions—to which he aspires. It is, however, fortunate that this is not the only value of the (h', f) curves, for as we now shall see, the derivation of the (N, z) curves from them is full of difficulties.

Perhaps the greatest of these difficulties springs from the dependence of the signal-speed on N, which is to blame for the difference between virtual height h' and true height h. Even if it is fully justifiable to identify signal-speed with group-speed v, the difficulty is not banished. It resides in the fact that $(h' - h)$ depends not on things already known but on the very thing one is striving to find out, to wit, the distribution of ion-density in the ionosphere. The value of h for any particular h' depends indeed not on the value of N at that height alone, but on the values of N at all inferior heights. The problem is somewhat like having to solve for x an equation in which x appears badly entangled on both sides of the equals sign. The mathematical technique is difficult and approximative.

Another major difficulty resides in the fact that when N varies by an appreciable fraction over a distance equal to a wave-length of the waves, the consequences of the theory become a good deal more complex than those embodied in the simple equations (4) and (5). For instance, partial reflection may occur at a level where N is rising rapidly, though as yet far below the mirror-density N_c. One cannot therefore say that whenever an echo is observed on a frequency f, there *must* somewhere exist an electron-density related to f by (5). The literature is full of allusions to mystifying echoes, some of which are ascribed to partial reflection. Figure 2 shows an E-echo and an F-echo received from the same signal; and many a (h', f) curve shows the E-branch running along for quite a distance underneath the F-branch, instead of stopping at just the abscissa where the F-branch begins.[5] Yet on the other hand, the E-layer may be denser at some places than it is at others of equal altitude, and parts of a signal may be reflected from the places of high density while other parts slip between these and go on to the F-layer. The assumption that N depends on z only, which hitherto has been taken for granted in this paper as it is in most theory, is in fact very assailable; and people are

[5] This is so well-known a phenomenon that lengthy papers have been written about it under the name of "abnormal E-ionization," though it seems too common to deserve the adjective "abnormal."

beginning to study the distribution of N in the horizontal plane, e.g. by using obliquely-sent as well as vertically-emitted signals.

Now we turn briefly to a difficulty affecting not the relation between h' and h, but the relation between N_e and f presented as equation (5). This equation was based on the tacit assumption that the electric force on a single electron is the same as though there were no other electrons at all in the ionosphere. The assumption has been doubted, and quite a polemic has ranged about it. The question is in fact a special case of one of the most pestiferous questions of all mathematical physics, occurring for instance in the theory of magnetized bodies and of bodies polarized electrically: when a great many similar atoms side by side are exposed together to an external field, how is the force suffered by any one of them modified by the presence of its equally-affected neighbors? One strongly-held position is, that there *is* such a modification which manifests itself in a factor 3/2, to be multiplied into the right-hand member of equation (5). A test experiment has been devised, and the early results have favored this theory. The presence or absence of this factor alters in equal proportion all the ordinates, but does not modify in the least the trend of the $N(z)$ curve; but the student specially interested in numerical values of N must discover, from each paper wherein such are given, which formula was used in computing them.

After uttering all these warnings about the theory underlying the (N, z) curves, I will risk a few statements about the curves themselves.

The shape of the (N, z) curve, when the sun is low in the sky and there is no crinkle in the $(h'f)$ curve, is roughly that of Fig. 3*A*. If the sun is within some 40° of the zenith and the crinkle is present in the $(h'f)$ curve, the theory indicates not that the F-peak of Fig. 3*A* has split into two, but rather that a bulge has appeared on the left-hand side of the F-peak. The letters F_1 and F_2 are then applied to the bulge and the peak, respectively. If the shape of the (h', f) curve indicates yet another layer between E and F_1, it appears as a small hump in the valley between the peaks of Fig. 3*A*.

As for the N-values, those of most interest are those corresponding to the crests of the peaks; or to define them better by staying closer to the data, they are the ones corresponding to the points on $(h'f)$ curves which adjoin the gaps or lie at the tops of the crinkles. These may be called the values corresponding to the "crowns" of the several layers.

At Huancayo in the Peruvian Andes, at a typical summer noon, N has the values $1.8 \cdot 10^5 - 3.3 \cdot 10^5 - 1 \cdot 10^6$ at the crowns of the layers E, F_1, F_2: so says Berkner. At Slough near London, at noon

on a certain day of early spring (1933) Appleton found $1.2 \cdot 10^5$ at the crown of E and $3.8 \cdot 10^5$ at the crown of F, the F-layer being at that time and place not differentiated into F_1 and F_2. These figures are not far apart, if E be compared with E and F_1 with F; but with a little search I could have found plenty of values differing much more greatly, as is attested by Figs. 9 and 10. From the former of these we have already deduced that critical frequencies vary as the sunspot cycle proceeds: I now add that from minimum to maximum of the cycle just ending, N at the crown of the E-layer increased by three-fifths while N at the crown of F_2 went up no less than fourfold! From Fig. 10 we infer, by squaring the values of critical frequencies, how great is the change of these N-values with hour of day. Sudden unaccountable changes also occur; one evening over Cambridge (Massachusetts) the N-value for the E-layer was more than tenfold the values given above, being ascertained by Mimno as $2.8 \cdot 10^6$!

I therefore summarize, as precisely as seems justifiable: the N-values at the crowns of the layers vary with hour of day and time of year and year of the sunspot-cycle very markedly, not to speak of sudden unexplained fluctuations; and 10^5 to 10^6 electrons per cc. is a good figure to keep in mind for the order of magnitude thereof.

To terminate this section I show Fig. 12, in which the delay of the echo for a certain frequency (2 mc.) is plotted against time during the hours preceding and following dawn. Interpreting with the aid of Fig. 3*A*: during the night the E-peak was too low to echo back the signals of this frequency, which accordingly climbed farther and found their mirror in F; but at 6:35 A.M. very sharply, the E-peak increased in height to just the extent needed to intercept them. Or since confusion may arise from using the word "height" in two senses, I express what went on in an exacter way: during the night the electron-density at the crown of the E-layer was inferior to the mirror-density for $f = 2$ mc., but with the oncoming of day it rose, and at 6:35 A.M. very sharply it attained and overpassed that mirror-density.

I recall that Fig. 11 exhibits how the virtual altitudes of the layer-crowns vary with hour of day and season of the year in the sky over Washington. It is evident that E is a fixture of the ionosphere with a virtual height surprisingly steady at the close neighborhood of 120 km, while F rises and falls in the course of a winter day, rises and falls and divides and merges again during a day of summer.

Now we must take brief notice of a difficult subject: the forkings and the doublings of the (P', f) curves, and the theory which finds their source in the earth's magnetic field.

A magnetic field, the earth's or any other, should have no effect whatsoever on radio waves so long as these are traveling in air composed entirely of neutral molecules. When, however, the waves are setting electrons into motion, the moving electrons are affected by the field, which has a twisting action on their paths. We have seen already that the moving electrons react, so to speak, upon the waves, raising the wave-speed thereof. By altering the motions of the electrons, the magnetic field will influence at second hand the waves themselves. But will the result be perceptible? In view of the fact

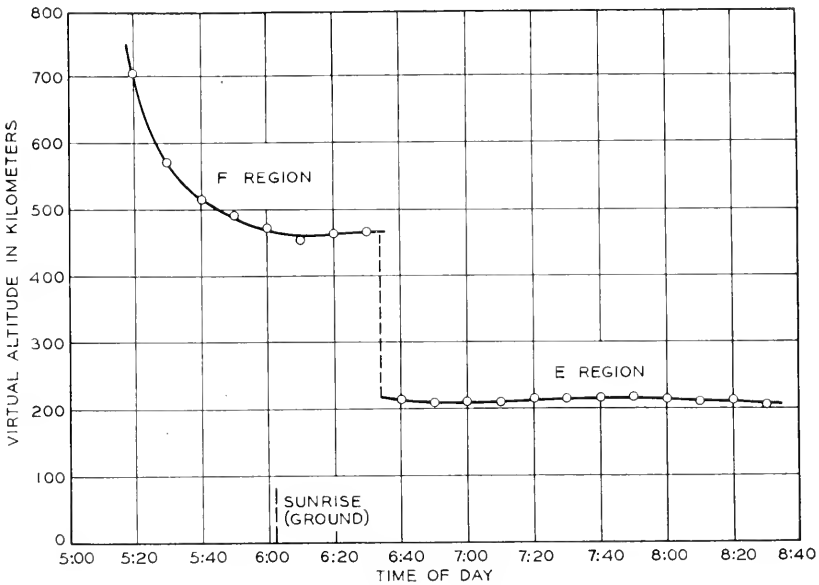


Fig. 12—Ceiling for signal of fixed frequency shifting near sunrise from F to E as the electron-density of the E-layer increases with increase of light. (Appleton.)

that the earth's magnetic field is very feeble by comparison with the fields between the poles of our electromagnets great or small, or even with those around the horseshoe magnets which are playthings, one might well think the influence not worth the trouble of computing. But those who first undertook to compute it—Nichols and Schelleng in America, Appleton independently in England, in the winter of 1924–25—found it a serious influence, and very well worth the trouble.

The problem is one of those which are not very hard to state, but can be very tedious to solve except in special cases which may or may not be of practical importance. For this problem it happens that two of the special cases can be solved with relative ease, and one at

least is realizable in practice. I cannot venture to give the theory of even this one, but at least I will attempt to describe what happens.

The special case occurs when the waves are traveling at right angles to the field. Since they travel vertically,[6] it is necessary to find a place on the earth where the field is horizontal. Such places are found in the equatorial regions only, and these are not precisely crowded with universities or engineering experiment stations. However, the Carnegie Institution of Washington was inspired, several years ago, to set up a station in just such a place: Huancayo, in the Andes of Peru. Here they established long straight horizontal antennae, one running north-and-south, another east-and-west, and yet another northeast-and-southwest. In the waves which mount from these to the ionosphere and then come bouncing back, the electric field $E_0 \sin nt$ —henceforth to be called "the electric vector"—is faithful to the direction of the antenna. They are called "plane-polarized waves."

When the north-south antenna is used, the electrons are impelled to and fro in the north-south direction which is that of the magnetic meridian. Now as is well known, an electron moving parallel to a magnetic field behaves just as it would if there were no such field at all. These waves ought therefore to behave according to the theory which we set up while we were still disregarding the magnetic field. They are the so-called "ordinary waves" or "o-waves."

When the east-west antenna is used, the electrons of the ionosphere are impelled to and fro in the east-west direction, which is transverse to the earth's magnetic field. This is just the condition for the maximum amount of meddling by the field in the motion of the electrons. The meddling consists in bending the electron-paths into curiously twisted arcs. *The action of the magnetic field is tantamount to strengthening the electron-current-density I_e parallel to the electric vector.* It will be recalled (from page 464) that it is I_e which for small N-values cancels a part of the displacement-current and so speeds up the waves, and for a certain critical N-value cancels the whole of the displacement-current and so brings about total reflection. So, for these "extraordinary waves" or "x-waves," a given N-value produces a greater augmentation of the wave-speed, and the critical N-value for total reflection is smaller, than for the ordinary waves. The signal composed of x-waves, mounting into the ionosphere, finds its appropriate mirror at a lesser altitude than does the signal composed of o-waves, and it gets earlier back to earth. It may indeed come

[6] In addition to the other advantages of sending the waves up vertically, there is the feature that the angle between their line of motion and the field is the same when they are going up and when they are returning.

back from E while the o-signal goes on to F or from F_1 while the o-signal goes on to F_2, or from F_2 while the other goes irretrievably forth into space.

When the northwest-southwest antenna is used, the ionosphere takes charge of the signals, and separates them into an o-component and an x-component. Each travels according to its proper law, and the x-component reaches its lower-down mirror earlier and beats the o-component back to earth. Two echoes return instead of one. The earlier is plane-polarized with electric vector east-and-west; the laggard is plane-polarized with electric vector north-and-south. Suppose a long straight horizontal antenna is used to respond to the returning signals. It will respond to both, if pointed north-west-southeast; only to the earlier, if pointed east-and-west; only to the later, if pointed north-and-south.

All the foregoing were statements of theory at first, but thanks to the experiments of Wells and Berkner at Huancayo, they now are statements of data as well.[7] Figure 13 exhibits a small selection from the data.

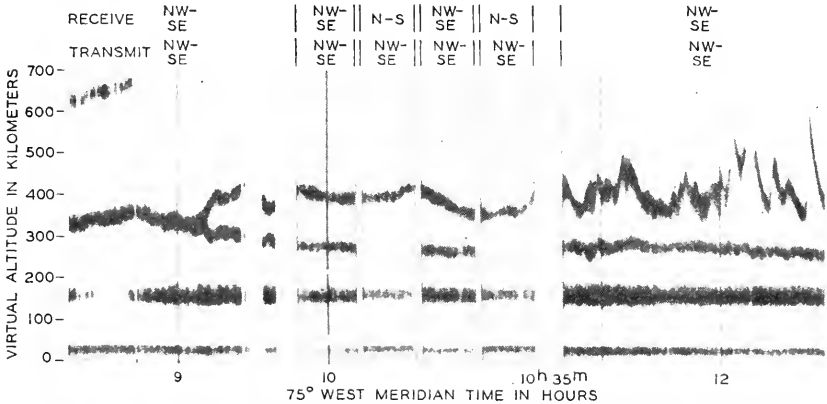


Fig. 13—Echoes of plane-polarized signals near the geomagnetic equator. (Wells and Berkner.)

Now look again at Fig. 4: formerly I asked the reader to ignore one of the curves, but now we will compare the two. The circles and the crosses indicate the o-wave and the x-wave respectively. One is

[7] This is a good place to speak of a question which may already have occurred to many readers, viz. the question why we assume the charged particles in the ionosphere to be free electrons rather than charged atoms or molecules. Were they of atomic or molecular mass, the separation of the o and x echoes would be inappreciable, and the "gyro-frequency" later to be mentioned (page 482) would be quite outside of the radio range. It is not, however, excluded that among the free electrons there may be a great multitude of charged atoms, perhaps even many times more numerous than they, though much less influential.

a fairly close copy of the other, shifted a certain distance along the horizontal axis. From the magnitude of the shift it is possible to compute the strength H of the earth's magnetic field; or let me rather say, it is possible to compute a numerical value which must agree with H, or else the theory will be vitiated. When the computation is made, the value turns out to be just a few per cent less then the field strength at ground-level. The action of the field through the electrons on the waves is exercised only in the ionosphere, which is hundreds of kilometers up in the sky; and it is quite reasonable to believe that these few per cent are actually the falling-off in the field strength from the ground up to that level. Such is the present belief, and many of those who work in terrestrial magnetism are happy over the prospect of measuring thus the field in regions where there seems to be no greater hope of anyone ever actually going, than of going to the moon.

Actually Fig. 4 shows data obtained in England, which is far from the equator; I point this out in order to mention that even when the waves are traveling obliquely to the earth's magnetic field, there is a separation of the signals into pairs of echoes, and these are still amenable to theory. In this general case of oblique transmission, the waves are polarized elliptically—a feature difficult to visualize without a certain amount of specialized knowledge, but lending itself to some very neat and pretty experimental tests.[8] In the special case of transmission parallel to the magnetic field, waves initially plane-polarized should remain of this character but their plane of polarization should rotate as they proceed. There are indeed so many curious and interesting details of the influence of the field through the electrons on the waves, that a writer must be ruthless in ignoring them if he is to observe decent limits of space. I will mention only in closing that the "gyro-frequency" $eH/2\pi mc$—which in our latitudes is around $1.3 \cdot 10^6$ mc.—plays the part of a resonance-frequency. Waves too close to this frequency are liable to great, not to say distressing, anomalies in transmission. Theoretical statements about waves in general are likely to assume two different forms, one appropriate to those of frequency higher and the other to those of frequency lower than the gyro-frequency; it is the former which appears in this paper.

In Fig. 14 there appears something which, if the war had begun before its discovery, would perhaps have been called a "blackout."

[8] It may perhaps be regarded as obvious that the ellipse of polarization should be described in opposite senses in the northern and southern hemispheres, since in one hemisphere the magnetic lines of force are coming out of the ground, in the other they are diving in. This inference was tested by a special experiment in Australia, and the result was taken as establishing the "magneto-ionic theory," as this general theory is often called.

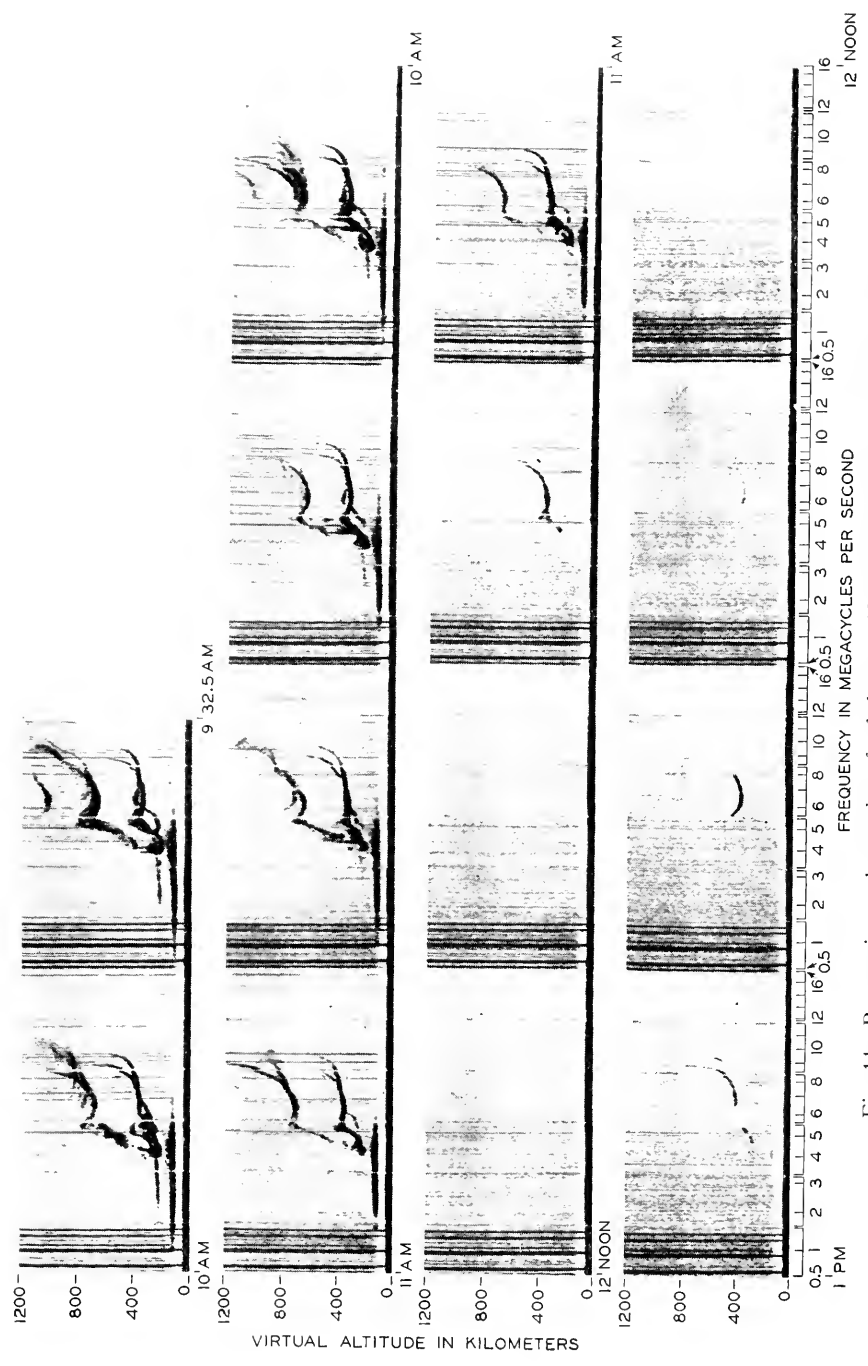


Fig. 14—Progression and recession of a fadeout. (Carnegie Institution of Washington.)

Being discovered however some years before the war, it was and is called a "fadeout." The apparatus was that which I mentioned on page 471; accordingly each of the pictures was traced in fifteen minutes, and as soon as each was finished the next was begun. See how the pattern of (P', f) curves, familiar and sharp in the earlier pictures, dissolves into fragments and then is completely wiped out! Later on it begins to come back piecemeal, and finally is restored as good as ever.

Since attention was focussed on such events in 1935, they have been reported by the scores in every year, varying in duration and in severity. It requires no (h', f) curve to show them, since ordinarily they cut off communication by radio, and with the sharpness of a knife. Many an engineer, to quote from Dellinger, has "dissected his receiving equipment in the vain effort to determine why it suddenly went dead." Over broad areas the extinction is sudden and simultaneous in many fadeouts, more gradual in others; the restoration is as a rule more gradual.

Shall we interpret this strange and striking effect as a sudden vanishing of the ionosphere and all the reflecting layers thereof, or as a swallowing-up of the signals by something which is suddenly created underneath the ionosphere? Against the first suggestion it is to be said, that no one can image anything which might so suddenly frighten all the electrons of the ionosphere back under cover, so to say—drive them into the arms of their parent molecules in a few seconds or minutes—when all day and even at night they manage to hold their freedom. Such a graph as Fig. 15 speaks also against it forcibly. Here in the upper part of the figure we see the critical frequencies [9] of F_2, F_1 and E as located every fifteen minutes on (P', f) curves such as those of Fig. 14. Each flock of data lies along a curve which, intercepted though it is by the fadeout, resumes so nearly at the level where it left off that one can hardly believe that the ionosphere totally vanished in between.[10]

As for the curve marked "f_{MIN}" in Fig. 15, it represents the lowest frequency at which echoes are observed. I have said nothing as yet about there being such a minimum-frequency. How indeed can there be one, and why should signals of any frequency however low fail to be echoed, since the mirror-density for any higher frequency is *a fortiori* more than a mirror-density for any lower?

[9] For the ordinary waves, as indicated by the superscript in symbols such as f^o_E.

[10] In violent magnetic storms the ionosphere is so convulsed that the echoes lose their sharpness entirely, and (h', f) curves like those of Fig. 7 are replaced by broad smudges; or echoes may vanish altogether. These are quite different from fadeouts.

Attempting to supply an answer to this question, I point out that in such part of the theory as I have thus far given, there is nothing corresponding to *absorption*. This is because the electron-current

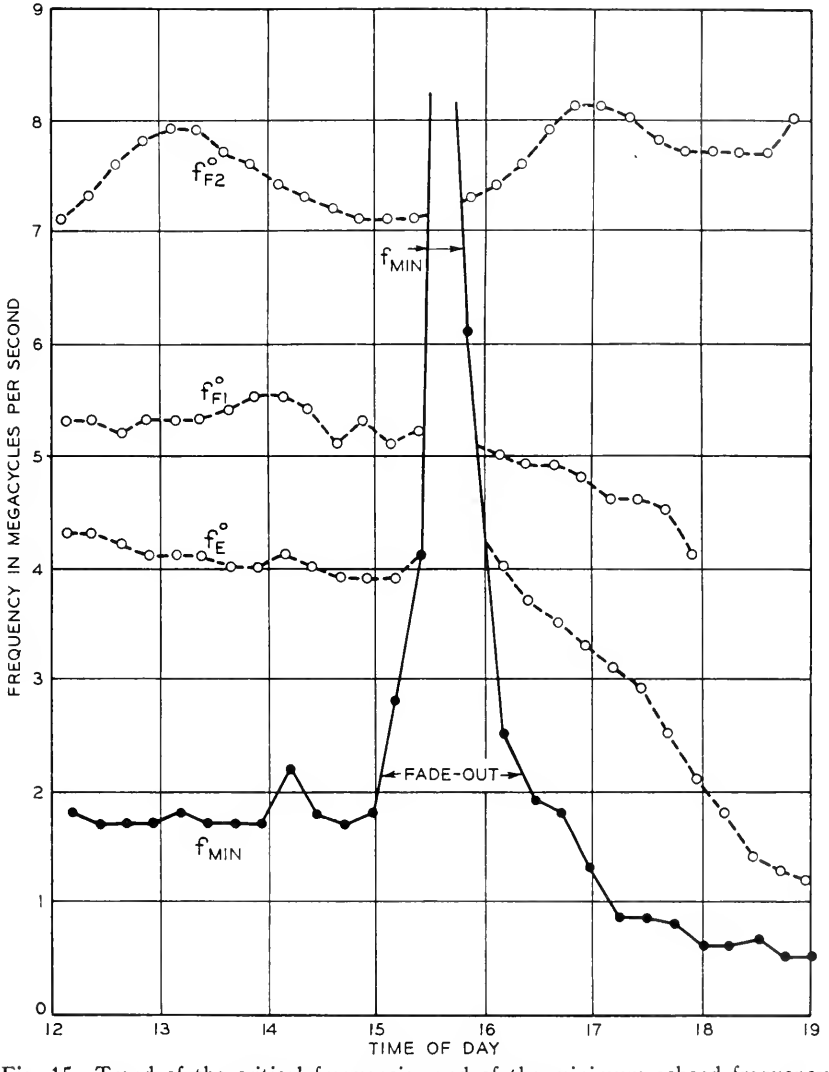


Fig. 15—Trend of the critical frequencies and of the minimum echoed frequency as a fadeout proceeds. (Berkner.)

(I_e in the former notation) is in quadrature with the electric vector in the waves, being thus a "wattless current." This in turn is because we have assumed each electron to be entirely free, oscillating in the

wave-swept aether as though there were nothing else in the world but itself and the waves. If, however, the electron were occasionally to strike and bounce off a molecule it would leave some (even though but a small part) of its kinetic energy behind, and this would have to be replenished by the waves. Indeed as soon as collisions are taken into account, the algebra tells us that the electron-current is no longer in perfect quadrature with the electric vector. Dissipation of energy occurs together with reflection, and if it is sufficiently great —if, that is to say, the electrons collide often enough with molecules— it takes the place of reflection. In the terms of my acoustic simile on an earlier page, the signals are no longer echoed back from the hard slopes of the mountain-range of Fig. 3*A*, but are swallowed up and lost as if in something soft and woolly.

One would expect absorption to occur in the lower reaches of the ionosphere rather than in the upper, since the air is denser there and the electrons suffer many more collisions. The absorbing layer, that is to say, must be situated just where it is able to cut us off from the reflecting layers by dissipating the signals which we send. Why should it do so occasionally with such completeness, and most of the time not do so at all, for any except the lowest frequencies?

For an answer to this query, one looks again to the sun. Ordinarily, we will suppose, the ionizing agent coming from the sun penetrates deep enough into the air to form the reflecting layers high overhead, but is nearly consumed in so doing. Occasionally, though, the sun sends forth a quite abnormal transitory burst of radiation, so strangely constituted that it passes the reflecting layers without contributing to them or weakening itself, and continues so far down that at the level where it at last engenders free electrons they constitute a layer absorptive and not reflective. This would be no more than an *ad hoc* assumption, were it not that brilliant eruptions are frequently seen on the face of the sun at the moments when fadeouts are commencing. To some extent it is still an *ad hoc* assumption, for the light whereby the eruption is seen is certainly not ionizing light, and we must assume that the visible light is attended by rays of other wave-lengths having just the properties desired. Coincidence of fadeout and eruption is, however, so frequent, that it would now take a very sceptical mind to reject the assumption. The trend of the curve "f_{MIN}" in Fig. 15 sustains it.

And so I have now come back to the theory that it is radiation from the sun which makes the upper atmosphere into an ionosphere, by detaching electrons incessantly from the molecules thereof. (The detaching must be incessant, for the electrons are always liable to

recapture by the molecules.) Of this theory it may be said that the major facts confirm it, though at night the E-ionization persists so tenaciously that we are obliged to seek for a separate agent (meteors, perhaps?); while numerous minor discrepancies can be explained away by making special assumptions which can neither be confirmed nor refuted because there is so little independent knowledge of the upper atmosphere. Not a very satisfactory situation for the present, but at any rate one which offers endless promise!

Thus it can readily be seen that as the ionizing light descends from heaven through the upper air, the ionization per unit volume should at first increase (because the air is getting denser) and then decrease (because the light is getting to be used up). This offers an explanation for a layer; and the mathematical working-out of the idea—due in the main to Chapman—shows that not only the existence but the shape of either peak in the curve of Fig. 3A is compatible with the theory. But there are several layers and peaks, not just one; how does this come about? Well, the atmosphere is a mixture of several gases, differently susceptible to the ionizing light; one can attribute a peak to each gas (indeed more than one to a gas, by invoking different states of the molecules). The height of a peak, the N-value at the crown of a layer, should rise and fall as the sun rises and sinks in the sky. This is true of the layers E and F_1, as we saw from Fig. 11, and again there is a quantitative theory by Chapman, which is borne out in some though not in full detail. Of F_2 it is not always true, as Fig. 11 proclaims; there is a minimum at noon in summer, and the highest N-values of all are attained in winter! One tries to cope with the discrepancy by assuming that as the sun climbs higher in the sky, the F_2 region expands so much in the heat that although the total number in the region is properly increasing, the number in unit volume suffers a decline. The layers do not disappear at night, though the N-values shrink. There seems to be plenty of time for recapture of all the electrons between sunset and sunrise, and one is driven to hunt for other causes of ionization which emerge when the sunlight is gone. These remain mysterious. Inrush of meteors into the high atmosphere has been suggested as one of the causes, and also incessant streams of charged particles similar to those which become intense during magnetic storms.

Sunlight is therefore not the only, yet apparently the major factor in maintaining the ionosphere. Not, however, any sunlight that we ever feel! This portion of the sun's outpourings is so thoroughly consumed above that it never reaches down to the levels where we live. Were it not so consumed, we should not be able to communicate by

radio very far over the earth. The reader may think that this is not very important: our ancestors lived without radio, why should we worry about lacking it? Well, it is probably quite true that if the ionosphere were not overhead, we should not be worrying about the lack of radio. We should in fact probably not be worrying about anything at all, for we should not be here to worry. The ultra-violet light of the sun, pouring down upon the surface of the globe unhindered, would work changes so severe on organisms as we know them that life would have to be very different, and perhaps impossible. This lethal light is like an enemy, which in attacking a city spends itself in throwing up a barrier against itself; and the barrier not only keeps the enemy out, but is serviceable otherwise to the dwellers in the city.

Abstracts of Technical Articles by Bell System Authors

Stereophonic Reproduction from Film.[1] HARVEY FLETCHER. On April 9 and 10, 1940, demonstrations of the stereophonic reproduction of music and speech, described in this article, were given at Carnegie Hall, New York, N. Y. These demonstrations represented the latest development in a series of researches by Bell Telephone Laboratories, the first step of which was demonstrated in 1933 when a symphony concert, produced in Philadelphia, was transmitted over telephone wires to Washington, and there reproduced stereophonically and with enhancement before the National Academy of Sciences.

For the present demonstrations, original recordings of orchestra, choir, and drama were made at Philadelphia and Salt Lake City; and at a later audition the artist or director was able to vary the recorded volume and to change the tonal color of the music to suit his taste. At will, he could soften it to the faintest pianissimo or amplify it to a volume ten times that of any orchestra without altering its tone quality, or he might augment or reduce the high or low pitches independently. The music or drama so enhanced is then re-recorded on film, with the result that upon reproduction, a musical interpretation is possible that would be beyond the power of an original orchestra, speaker, or singer to produce.

Wave Shape of 30- and 60-Phase Rectifier Groups.[2] O. K. MARTI and T. A. TAYLOR. The installation of mercury arc rectifiers with a total capacity of 82,500 kilowatts by the Aluminum Company of America at Alcoa, Tennessee and Massena, New York, was accompanied by widespread increases in the inductive influence of the interconnected power supply networks with resultant increases in the noise on exposed telephone circuits. Because of the size of these installations, and the complexity of the supply systems, it appeared impracticable to limit the rectifier harmonics by the use of frequency-selective devices, which have been successfully applied to certain smaller installations. However, the results of a cooperative study indicated that by means of a relatively simple arrangement of phase shifting transformers, the equivalent of 30- or 60-phase operation of the rectifier stations could be secured. In this way, the important harmonic components on the power systems were reduced to relatively small values, and wave shape and noise conditions were restored practically to normal.

[1] *Jour. S. M. P. E.*, June 1940.
[2] *Elec. Engg.*, April 1940.

This paper describes briefly the voltage and current relations, preliminary tests on a small-scale rectifier, the phase shifting transformers and their application to this particular situation, and presents data to indicate their effectiveness.

A New Quartz-Crystal Plate, Designated the GT, which Produces a Very Constant Frequency over a Wide Temperature Range.[3] W. P. MASON. In this paper, a new quartz-crystal plate, designated the GT, is described which produces a very constant frequency over a wide temperature range. This crystal does not change by more than one part in a million over a 100-degree centigrade range of temperature. This crystal obtains its great temperature stability from the fact that both the first and second derivatives of the frequency by the temperature are zero. Both the frequency and the temperature coefficient can be independently adjusted.

This crystal has been applied in frequency standards, in very precise oscillators, and in filters subject to large temperature variations. It has given a constancy of frequency considerably in excess of that obtained by any other crystal. A crystal chronometer, using this type of crystal, was recently lent to the Geophysical Union for measurements on the variation of gravity and the chronometer is reported to have kept time within several parts in 10 million, although no temperature control was used.

Room Noise at Telephone Locations—II.[4] D. F. SEACORD. Room-noise data, based primarily on measurements made at about 900 locations in and around Philadelphia and Chicago under winter conditions, were reported informally to the conference on sound at the 1939 A. I. E. E. winter convention. The present article supplements the earlier material and includes a summary of room-noise conditions expressed in terms of annual averages based on both the winter survey data previously discussed and the summer survey data that had been obtained at about 1,300 locations but had not been completely analyzed at the time of the earlier report. The summer survey included 500 measurements at locations previously measured during the winter in and around Philadelphia and Chicago, and 800 measurements at locations in and around Cleveland, New York City, northern New Jersey, and Philadelphia. Annual average as used in this article is the mean of winter and summer measurements. In addition, the present article includes a brief discussion of outdoor noise and the relative frequency of occurrence of several predominant sources of room noise.

[3] *Proc. I. R. E.*, May 1940.
[4] *Elec. Engg.*, June 1940.

The noise measurements were made with equipment conforming to the specifications described in the A. S. A. "Tentative Standards for Sound Level Meters" (Z24.3–1936), using the 40-decibel loudness-weighting network. The measurements are expressed in terms of sound level in decibels above reference sound level, that is, 10^{-16} watt per square centimeter at 1,000 cycles in a free progressive wave, each measurement being based on the average of 50 individual readings.

Electrical Conductance Measurements of Water Extracts of Textiles.[5] A. C. WALKER. It has been shown that the electrical properties of textiles depend upon chemical composition, water-soluble electrolytic impurities, moisture content and manner of drying the material from the wet state. Selection of a textile for electrical puposes should include consideration of the influence of chemical composition upon the properties of the material, absence of significant amounts of electrolytes, and the method of drying the material from the wet state.

This paper discusses the water extract conductivity method, its correlation with insulation resistance data, and describes a simple, durable electrolytic cell which is convenient to use for the conductivity measurements.

[5] *A. S. T. M. Proc.*, Vol. 39, 1939.

Contributors to this Issue

H. W. Bode, A.B., Ohio State University, 1924; M.A., 1926; Ph.D., Columbia University, 1935. Bell Telephone Laboratories, 1926–. As Consultant in Network Theory, Dr. Bode is engaged in research studies of electrical networks and their applications to various transmission problems.

R. P. Booth, S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1925. American Telephone and Telegraph Company, Department of Development and Research, 1925–34; Bell Telephone Laboratories, 1934–. Mr. Booth has been active in the development of line design methods suitable from the interference standpoint for carrier and broad-band transmission.

Karl K. Darrow, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. As Research Physicist, Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

Frederick J. Given, S.B., Harvard and Massachusetts Institute of Technology, 1919. Western Electric Company, Engineering Department, 1919–1925; Bell Telephone Laboratories, 1925–. Mr. Given has been engaged in design and development of transmission apparatus, including retardation coils, condensers, and transformers as well as loading coils and cases.

K. E. Gould, B.S. in Electrical Engineering, Oklahoma A. and M. College, 1924; M.S. 1925, Sc.D. 1927, Massachusetts Institute of Technology. American Telephone and Telegraph Company, 1927–34; Bell Telephone Laboratories, 1934–. Dr. Gould, formerly engaged in inductive coordination studies, is concerned with transmission measurements at high frequencies.

Victor E. Legg, B.A., 1920, M.S. 1922, University of Michigan. Research Department, Detroit Edison Company, 1920–21; Bell Telephone Laboratories, 1922–. Mr. Legg has been engaged in the development of magnetic materials and in their applications, particularly for the continuous loading of cables, and for compressed dust cores.

TODOS M. ODARENKO, University of Technique in Prague, E.E., 1928. New York Telephone Company, 1928–30; Bell Telephone Laboratories, 1930–. Mr. Odarenko has been engaged in the measurement and study of transmission characteristics of existing and newly developed types of transmission lines.

T. SLONCZEWSKI, B.S. in Electrical Engineering, Cooper Union Institute of Technology, 1926. Bell Telephone Laboratories, 1926–. Mr. Slonczewski has been engaged in the development of electrical measuring apparatus.

VOLUME XIX OCTOBER, 1940 NUMBER 4

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XIX *October, 1940* *No. 4*

The Carrier Nature of Speech

By HOMER DUDLEY

Speech synthesizing is here discussed in the terminology of carrier circuits. The speaker is pictured as a sort of radio broadcast transmitter with the message to be sent out originating in the studio of the talker's brain and manifesting itself in muscular wave motions in the vocal tract. Although these motions contain the message, they are inaudible because they occur at syllabic rates. An audible sound is needed to pass the message into the listener's ear. This is provided by the carrier in the form of a group of higher frequency waves in the audible range set up by oscillatory action at the vocal cords or elsewhere in the vocal tract. These carrier waves either in their generation or their transmission are modulated by the message waves to form the speech waves. As the speech waves contain the message information on an audible carrier they are adapted to broadcast reception by receiving sets in the form of listeners' ears. The message is then recovered by the listeners' minds.

SPEECH is like a radio wave in that information is transmitted over a suitably chosen carrier. In fact the modern radio broadcast system is but an electrical analogue of man's acoustic broadcast system supplied by nature. Communication by speech consists in a sending by one mind and the receiving by another of a succession of phonetic symbols with some emotional content added. Such material of itself changes gradually at syllabic rates and so is inaudible. Accordingly, an audible sound stream is interposed between the talker's brain and the listener. On this sound stream there is molded an imprint of the message. The listener receives the molded sound stream and unravels the imprinted message.

In the past this carrier nature has been obscured by the complexity of speech.[1] However, in developing electrical speech synthesizers

[1] Speech-making processes are here explained in the terms of the carrier engineer to give a clearer insight into the physical nature of speech. The point of view is essentially that of the philologist who associates a message of tongue and lip positions with each sound he hears. This aspect also underlies the gesture theory of speech by Paget and others and the visible speech ideas of Alexander Melville Bell. The author has been assisted in expressing speech fundamentals in carrier engineering terms by numerous associates in the Bell Telephone Laboratories experienced in carrier circuit theory. Acknowledgment is made in particular of the contributions of Mr. Lloyd Espenschied.

copying the human mechanism in principle, it was soon apparent that carrier circuits were being set up. Tracing the carrier idea back to the voice mechanism there was unfolded, a little at a time, the carrier nature of speech. Ultimately the speech mechanism was revealed in its simplest terms as a mechanical sender of acoustic waves analogous to the electrical sender of electromagnetic waves in the form of the radio transmitter. Each of these senders embodies a modulating device for molding message information on a carrier wave suitable for propagation of energy through a transmission medium between the sending and receiving points.

The Carrier Elements of Speech

This carrier basis of speech will be illustrated by simple speech examples selected to show separately the three carrier elements of speech, namely, the carrier wave, the message wave, and their combining by a modulating mechanism. These illustrations serve the purpose of broad definitions of the carrier elements in speech.

The illustration chosen for the carrier wave of speech is a talker's sustained tone such as the sound "ah." In the idealized case there is no variation of intensity, spectrum or frequency. This carrier then is audible but contains no information, for information is dynamic,[2] ever changing. The carrier provides the connecting link to the listener's ear over which information can be carried. Thus the talker may pass information over this link by starting and stopping in a prearranged code the vocal tone as in imitating a telegraph buzzer. For transmitting information it is necessary to modulate this carrier with the message to be transmitted.

For the second illustration, message waves are produced as muscular motions in the vocal tract of a "silent talker" as he goes through all the vocal effort of talking except that he holds his breath. The message is inaudible because the motions are at slow syllabic rates limited by the relatively sluggish muscular actions in the vocal tract. Nevertheless these motions contain the dynamic speech information as is proved by their interpretation by lip readers to the extent visibility permits. Another method of demonstrating the information content of certain of these motions is the artificial injection of a sound stream into the back of the mouth for a "carrier" whereby intelligible speech

[2] The information referred to is that in the communication of intelligence. There is, however, static information in the carrier itself. This serves for "station identification" in radio and may similarly help in telling whether it was Uncle Bill or Aunt Sue who said "ah."

can be produced from almost any sound stream.[3] The need of an audible "carrier" to transmit this inaudible "message" is obvious.

The final example, to illustrate the modulating mechanism in speech production, is from a person talking in a normal fashion. In this example are present the message and carrier waves of the previous

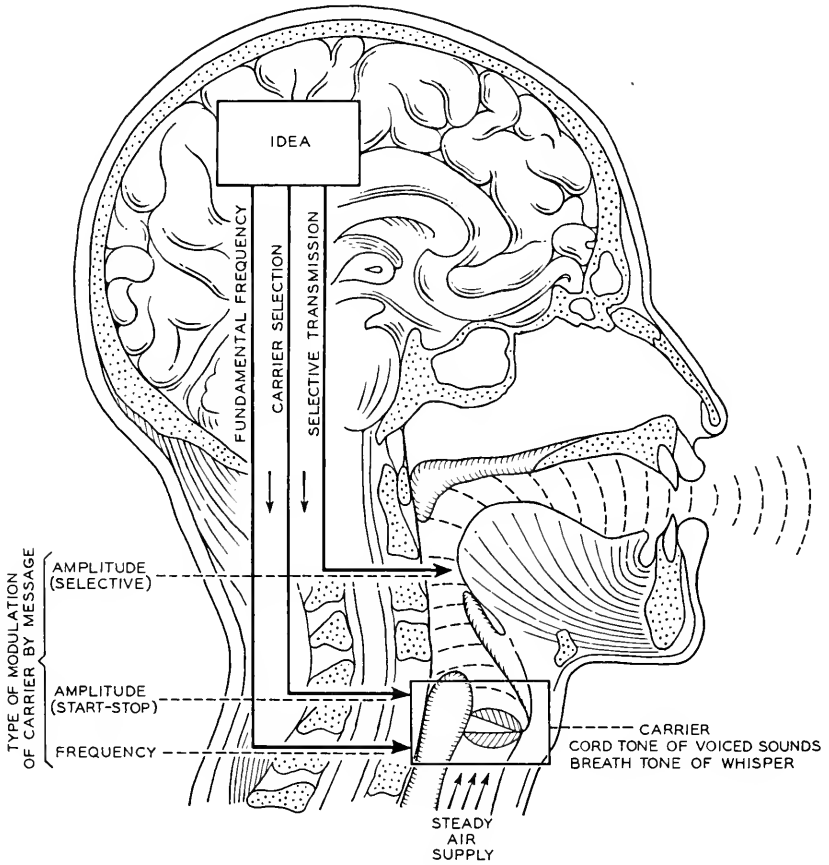


Fig. 1—The vocal system as a carrier circuit.

examples, for both are needed if the former is to modulate the latter. However, the mere presence of the carrier and message waves will not make speech for if they are supplied separately, one by a silent talker and the other by an intoner, no speech is heard but only the audible intoned

[3] R. R. Riesz, "Description and Demonstration of an Artificial Larynx," *Jour. Acous. Soc. Amer.*, Vol. 1, p. 273 (1930); F. A. Firestone, "An Artificial Larynx for Speaking and Choral Singing by One Person," *Jour. Acous. Soc. Amer.*, Vol. 11, p. 357 (1940).

carrier. Ordinary speech results from a single person producing the message waves and the carrier waves simultaneously in his vocal tract, for then the carrier of speech receives an imprint of the message by modulation.

The Speech Mechanism as a Circuit

The foregoing three illustrations by segregating the basic elements in speech production reveal the underlying principles. The present paper treats of these elements as functioning parts of a circuit. In Fig. 1 is shown a cross-section of the vocal system. The idea to be expressed originates in the talker's brain at the left top. Thence, impulses pass through the nerves to the vocal tract with the complete information of the "message," that is to say, what carrier should be used, what fundamental frequency if the carrier is of the voiced type and what transmission through the vocal tract as a function of frequency. The carrier whether voiced or unvoiced is shown for simplicity as arising at the talker's vocal cords. This carrier is modulated to form speech having the complete message imprinted on it preparatory to radiation from the talker's mouth to the ear of the listener, who recognizes the imprinted message.

In discussing the speech mechanism as a circuit, it is clearer to start with a block schematic. Figure 2 has thus been drawn to sketch the

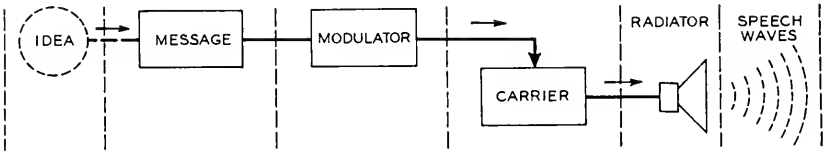


Fig. 2—The basic plan of synthesizing speech.

basic plan of speech synthesizing. As in Fig. 1, the idea gives rise to the message which modulates the voice carrier to produce the speech radiated from the talker's mouth. One can follow the path of the message from its inception in the talker's brain to its radiation from his mouth as an imprint on the issuing sound stream. The progress of the sound stream is also seen from its origin as an oscillatory carrier to its radiation from the talker's mouth carrying the message imprint.[4] The light arrow heads indicate direction of flow while the heavy ones indicate a modulatory control of the carrier by the message. This

[4] Here the carrier path is stressed to show the alteration of the carrier sound stream as it proceeds on its way from the point of origin to the point of radiation. This also accords with the importance of the voice carrier which is received and used by the ear, and thus differs from the treatment of the carrier in simple radio broadcast reception.

modulatory control is exerted on the carrier wave in part as the carrier is generated and in part as it is transmitted after generation.

Relevant Carrier Theory

The heart of the speech-synthesizing circuit of Fig. 2 is the part in which the group of waves making up the message modulate the component waves of the carrier. In any one of these modulations, there is the simple carrier process blocked out in Fig. 3. Here a message [5]

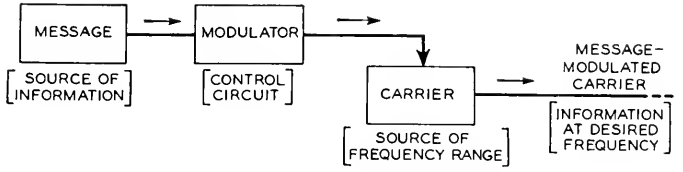


Fig. 3—The elements of a carrier sender.

containing the information modulates a carrier determining the frequency range so that the end product in the form of the message-modulated carrier contains the information of the message translated to frequencies in the neighborhood of the carrier. In this way the carrier sound stream of speech is imprinted with the message.

The prerequisites of the carrier system sender are, as indicated in Fig. 3, first, a carrier wave source; second, a message wave source; and third, a modulating circuit of variable impedance by which the message controls the carrier. The carrier wave is for the simplest case a single sine wave function of time characterized by an amplitude, a frequency and a phase. The message wave as a rule is more complex but may be analyzed as the sum of component sine waves each of which is characterized by its own amplitude, frequency, and phase. In most carrier circuits the frequency range of the message is below that of the carrier. This is true of speech production.

The function of the modulating circuit is supplying a means for the message wave to modify a characteristic of the carrier. If the carrier wave amplitude is modified by the message wave amplitude the process is known as amplitude modulation; if the carrier wave frequency is so modified the process is called frequency modulation while if the carrier wave phase is so modified the process is called phase modulation. No distinction is made as to whether the modification occurs during or

[5] The word "message" has been substituted for the usual carrier term "signal" to avoid confusion since the input signal is commonly speech whereas here the output wave is speech. "Message" seems particularly appropriate with its suggestion of code as in telegraph.

after the generation of the carrier. Modification of the carrier wave characteristics by other than the amplitude of the message need not be considered here. In the voice mechanism significant amplitude and frequency modulations of the carrier occur. Phase modulation takes place also but will not be discussed because the listener's ear is not very sensitive to these phase changes in the carrier.

In attempting to segregate the carrier elements of speech we run into one serious difficulty. In an idealized carrier circuit as shown in Fig. 3 connections can be cut between the two energy sources and the modulator so that each boxed element can be studied independently. With the human flesh of the voice mechanism this is no longer feasible; the use of cadavers would help very little because normal energizing is then impossible. The same difficulty often appears in electrical modulators as, for example, within a modulating vacuum tube where a grid voltage modulates a plate current. In such a case of common parts it is necessary to discuss the action of each of the three elements in the presence of the other two.

With this carrier theory review as a background we are in a position to analyze the three elements making up the carrier transmitting system of the human voice. While the picture presented is over-simplified in details the principles hold and aid in applying carrier methodology to explain the mechanism of speech.

The Voice Carrier

In electrical circuits the carrier is obtained from an oscillatory energy source. The same holds for speech. In the electrical circuit the oscillatory waves (a-c.) are ordinarily generated from a supply of d-c. energy.[6] The same is true in speech with the compressed air in the lungs furnishing the steady supply. Confusion must be avoided, for in speech the conversion of steady to oscillatory energy is often described as *modulation*. Here this conversion of energy form will be considered as an oscillatory action so that the term *modulation* can be reserved for the low-frequency syllabic control of this oscillatory energy to produce the desired speech. *Oscillatory* then will refer to automatic natural responses while *modulatory* will refer to forced responses which are controlled volitionally. This distinction is consistent with carrier terminology.

In the simplest electrical modulating circuits the carrier is a sine

[6] In the usual electrical circuit the carrier is cut off by turning off the output but leaving the carrier oscillator energized as, for example, in voice frequency telegraphy. In the voice mechanism, however, the oscillator is stopped at the source. The difference between the electrical on-off switching and the start-stop switching of speech is not fundamental but results from the use of the most suitable action in each case in view of the conditions prevailing.

wave although this is not true of the damped wave carriers of multi-frequency type once commonly used in spark wave radio telegraphy. The carrier wave in speech is not a simple sine wave. Such a sound would be like a whistle and so too limited for the rich flexibility of speech. Instead the voice carrier is a compound tone having a multiplicity of components of different frequencies which together cover the audible range fairly completely. While these components may be considered as a multiplicity of separate carriers it is simpler to think of the ensemble as a single complex carrier; so this terminology has been used in the earlier carrier illustration and elsewhere in this paper.

Aside from this compound nature of the voice carrier, the voice has two distinct types of carrier, one for voiced and one for unvoiced sounds. Some sounds such as "z" have both types present at the same time but this case may be treated as the superposition of one carrier on the other. For voiced sounds the carrier is the vocal cord tone, an acoustic wave produced by the vibration of the vocal cords consisting of a fundamental frequency component and the upper harmonics thereof. These decrease in amplitude with increasing frequency. For unvoiced sounds the carrier is the breath tone, a complex tone resulting from a constriction formed somewhere in the vocal tract through which the breath is forced turbulently to produce a continuous spectrum of frequency components in the audible range.

These carrier waves must be dissociated from any effects of resonant vocal chambers, for such characterize the speech message rather than the carrier. Furthermore, these carrier waves must be mentally pictured as sustained indefinitely with the starting and stopping of them also characterizing the message wave. Pauses for breath, due to incidental human limitations, do not invalidate the fundamental theory.

The Speech Message

Since a sustained voice carrier has no dynamic flow of information there is need for a source of message waves and a modulating mechanism for imprinting the message on the carrier. Conversely, any variation from the sustained carrier infers the presence of a message wave molding the carrier. The message consists of those articulating, phonating and inflecting motions of the vocal parts which imprint the information on the carrier sound stream. The importance of the message waves cannot be stressed too much. Any impairment of them is an impairment of the message.

The message waves include the motions producing speech changes at infra-syllabic rates, such as the effect of anger when a talker may be high-pitched for many minutes. When the carrier is thus altered over

a long period of time the question arises whether to use a long- or short-term value of the carrier. The answer may well be the same as in the analogous radio problem. If weather causes a carrier frequency to be slightly high all day, this higher value is taken as the normal carrier in studying short-term effects such as the degree of modulation. But in long-term studies of carrier stability the deviations from the mean represent a frequency modulation which is observed as a "message" effect.

Due to the inseparability of the message wave motion and its associated wave of impedance change in the modulating mechanism there may be confusion in distinguishing between the modulating elements and the source of the message waves. The rule followed here is simple. From the standpoint of the human flesh lining the vocal tract, the message source is internal, the modulating elements, external. The message consists of those muscular motions (or pressures or displacements) in the vocal tract which are present in the "silent talker" and are volitional in nature. This definition excludes the oscillatory motions which make up the carrier. The modulating elements are acoustic in nature since the carrier starts as a sound stream and ends as a modulated sound stream.

There are three important variations of the voice carrier and so three types of message and of associated modulation. These variations are: first, selecting the carrier; second, setting the fundamental frequency of the voiced carrier; and third, controlling the selective transmission of the vocal tract.[7] The message waves in the three cases will be discussed with the corresponding modulation reserved for consideration under the next heading.

Selecting the carrier appears as a simple start-stop message, complicated somewhat by the presence of two types of carrier and by locating the constriction for the unvoiced type at several places in the vocal tract. We may think of a start-stop type of message for each point where constrictions are formed, including the vocal cords for the voiced type of carrier. A constriction message may be plotted as the opening between vocal parts at the constriction with critical values for the onset of audible carrier. The constrictions are to a certain extent independent. Thus with the vocal cords vibrating, a constriction from the tongue tip to the upper teeth may also be formed, as in making the "z" sound. Again, in whispering, there may be simul-

[7] A fourth message characteristic prescribes the intensity of the speech. This message may be included in the carrier selection if the carrier is selected for intensity as well as type. The matter of intensity is passed over rather lightly here because a comparison is being developed between the human and electrical speech synthesizers with the final intensity in the latter under control of an amplifier setting.

taneous constrictions, both of the unvoiced type, one at the vocal cords and one in the mouth. As the voice has two distinct types of carrier, the vocal cord tone and the breath tone, the selection sets up one of four carrier conditions at any instant: no carrier, vocal cord tone only, breath tone only, or a combination of vocal cord tone and breath tone. This start-stop message resembles the on-off type of telegraph where switching controlled by other muscular motions sets up speech information in another code, that of telegraph. As mentioned earlier a communication system can be made with the vocal system by starting and stopping a voice carrier in a vocal imitation of a telegraph buzzer. While this would be a clumsy way of communicating information it marks the start-stop control of the voice carrier as a speech message and not part of the voice carrier. Another check is that the "silent talker" does form such constrictions.

The second type of message wave specifies the fundamental frequency with any related voice changes for the voiced type of carrier. This message, in a mechanical form, may be the time variation of the tension of the vocal cords. As the frequency of each upper harmonic is changed in the same ratio as the fundamental frequency, a single parameter suffices for all of the carrier components. The unvoiced carrier has no message of this type impressed since the unvoiced sounds are not characterized by pitch.

The third and final type of message wave controls the selective transmission in the vocal tract. By comparison, the first two types of message are simple, with the selecting of carriers ideally changing all components of the carrier by the same amplitude factor and the fundamental frequency control changing them by a uniform frequency factor. The vocal transmission, however, results from a multi-resonance condition with more than one degree of freedom. There follows a selective amplitude modulation with some carrier components decreasing in amplitude at the same instant that others are increasing. Maximum transmission occurs when a component coincides with an overall resonance, minimum transmission when it coincides with an anti-resonance and intermediate transmission for other cases. The voice message for transmission appears in mechanical form as the displacements of lips, teeth, tongue, etc., with as many such displacements considered as are needed for adequately expressing the speech content. This infers finding the simplest lumped impedance structure equivalent to the distributed impedance structure of the vocal tract to the necessary degree of approximation.

All these mechanical displacements of vocal parts that together constitute the voice message lead to corresponding displacements of

air in the vocal system, resulting in a set of air waves that likewise contain all the information of speech. These airborne message waves, however, are at syllabic rates and so below the frequency range of audibility.

The Voice Modulators

The three voice modulators associated with the three speech messages are the mechanisms of (*a*) selecting the carrier, (*b*) setting the fundamental frequency and (*c*) controlling the selective transmission. The mechanism for starting and stopping a voice carrier is simple. Assume a sustained carrier of either the voiced or unvoiced type. It can be stopped by opening the constriction at which it is formed. This alters the acoustic impedance of the opening which is then the modulating element in this case.

The modulating mechanism for controlling the fundamental frequency appears in the vibrating portions of air at the glottis. The exact mechanism is of no importance here so long as the message wave at the vocal cords finds means for altering the fundamental frequency under the control of the will.[8] This is a case of frequency modulation of multiple carriers harmonically related.

The modulating mechanism for controlling the transmission through the vocal tract as a function of frequency consists of the masses and stiffnesses of air chambers and openings in the vocal tract. These are varied under control of the message in the form of muscular displacements of vocal tract parts. There is a more complicated modulation in the vocal tract than in the usual electrical circuit for amplitude modulation because the varying impedances are reactive in the voice mechanism but resistive in the electrical circuit and also because several independent modulator elements are used in the voice mechanism as against either a single one or a group functioning as a unit in the simple electrical modulator. The reactive nature of the vocal impedances leads to the selective control of the amplitudes of the various harmonics of the voice carrier. The amplitude modulation of each carrier component by the combined message waves produces an output containing the carrier and sideband frequencies.

Comparison of Speech Synthesizing Circuits

The fundamental processes in human speech production are thus analogous to those of electrical carrier circuits. There is a switching of voice carrier energy comparable to that in voice frequency telegraph;

[8] For a simplified theory of the larynx vibration see R. L. Wegel, *Bell Sys. Tech. Jour.*, Vol. 9, p. 207 (1930) and *Jour. Acous. Soc. Amer.*, Vol. 1, Supp. p. 1, April 1930. The analogy of the larynx to a vacuum tube oscillator is described in an abstract, *Jour. Acous. Soc. Amer.*, Vol. 1, p. 33 (1929).

there is an altering of speech frequencies as in frequency modulating circuits; and finally, there is an amplitude modulation to yield a selective transmission of the various carrier components of the voice. However, the voice mechanism differs from the usual carrier circuit markedly as regards complexity. In the voice mechanism there are two types of carrier each with a multiplicity of partial carrier components. The incoming message has a multiple nature. Finally, several modulations take place including both amplitude and frequency types. This multiplicity of carrier relations indicates the wide range of voice phenomena possible.

Any electrical speech synthesizer must be a functional copy of the human speech synthesizer in providing the essential speech characteristics sketched in the preceding paragraph. There have been developed two such electrical synthesizers referred to in the introduction. A brief description of these will be given followed by some circuit comparisons.

These electrical synthesizers are known as the vocoder and the voder. The vocoder was so named because it handles the speech in a coded form; the voder, because it serves as a Voice Operation DEmonstratoR. Considerable interest has been manifested at the public showings of each of these synthesizers, the vocoder in a limited number of lecture demonstrations and the voder at the San Francisco and New York World's Fairs. Circuit details have been published elsewhere.[9]

Of these two speech synthesizers the vocoder was constructed first. It works on the principle of automatically remaking speech under control of spoken speech instantaneously analyzed to derive the code currents for the control. The vocoder as set up for demonstration is shown in Fig. 4.

The voder was derived from the vocoder by substituting manipulative for automatic controls. The resulting voder as displayed at the New York World's Fair is shown in Fig. 5. In the Fair demonstration, repeated continuously at intervals of about five minutes, the male announcer gives a simple running discussion of the circuit with the girl operator replying to his questions by forming sounds on the voder and connecting them into words and sentences. She does this by manipulating fourteen keys with her fingers, a bar with her left wrist and a pedal with her right foot. This requires considerable skill by the operators. The vocoder, automatic in nature, presents no problem of operating technique.

[9] The vocoder in the *Jour. Acous. Soc. Amer.*, Vol. 11, pp. 169–177, October 1939, "Remaking Speech," Dudley; the voder in the *Journal of the Franklin Institute*, Vol. 227, pp. 739–764, June 1939, "A Synthetic Speaker," Dudley, Riesz and Watkins.

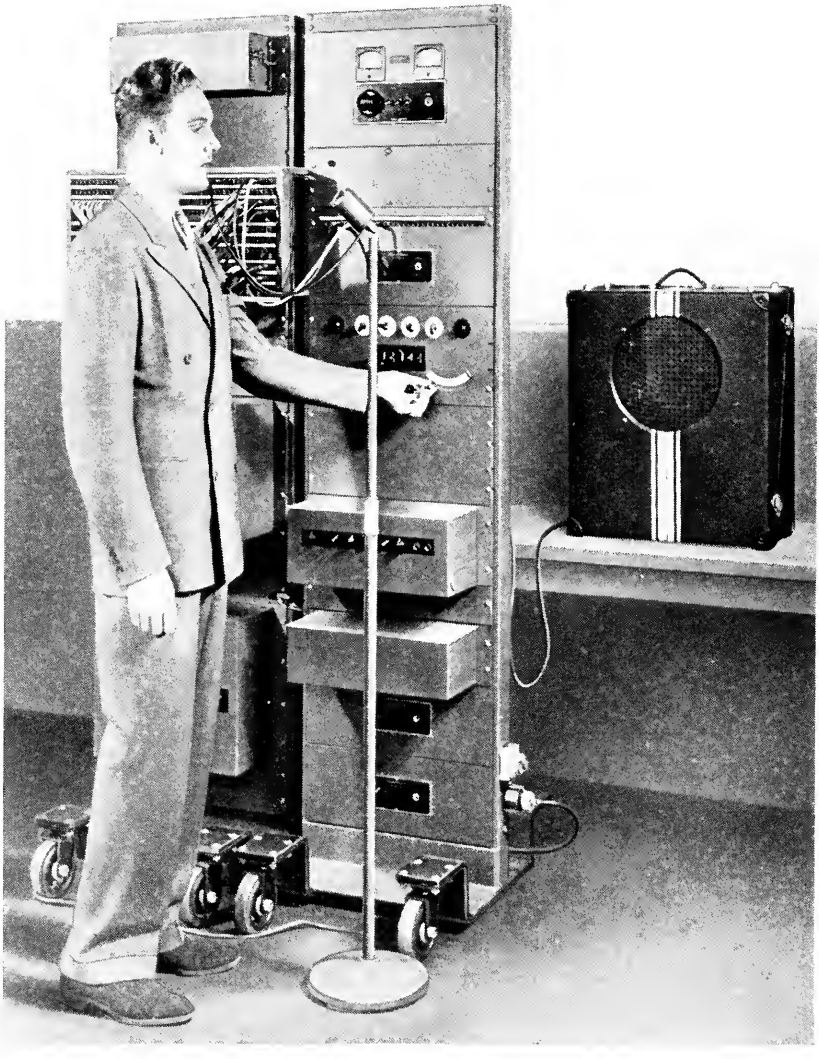


Fig. 4—The vocoder as demonstrated.

Circuit diagrams supply a shorthand for expressing the salient features of electrical circuits. In the next three figures comparative block circuits will be shown for the human and the two electrical speech synthesizers, tracing the communication from the origin of an idea in the communicator's brain to final expression as speech. In each cir-

Fig. 5—The voder being demonstrated at the New York World's Fair.

cuit, the arrangement in Fig. 2 will be followed with sufficient detail to show the functional relations of the parts discussed in this paper.

Figure 6 gives a block diagram of the voice mechanism of Fig. 1 with approximating electrical circuit symbols. The same communication paths can be traced. Thus from the talker's brain are sent nerve impulses that set up the message as a set of muscular displacements containing information as to the voice carrier to use, the fundamental frequency for the voiced carrier, and the selective transmission of the vocal tract. The air expelled from the lungs sets up as carriers the

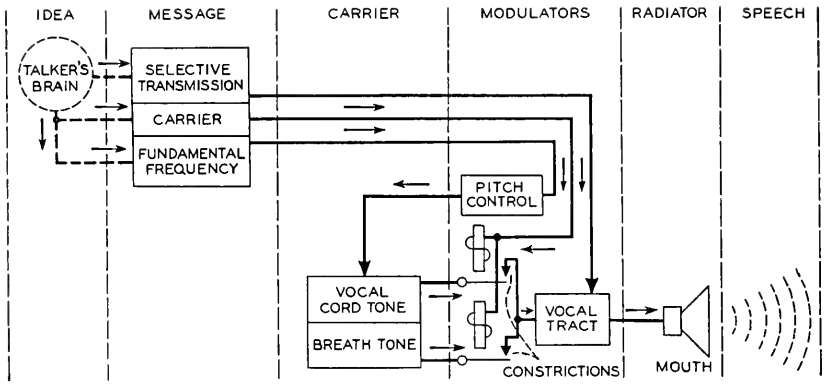


Fig. 6—Block diagram of the voice mechanism.

breath tone for unvoiced and the vocal cord tone for voiced sounds. For simplicity the carrier selection is shown after instead of before the carrier generation. These carriers are modulated by the message wave to produce the output of speech in the form of the message-modulated carrier in the audible range of frequencies.

Figures 7 and 8 show similar block schematics for the vocoder and the voder. The voder circuit has been simplified by the omission of a few controls for easier operation. In these electrical synthesizers, the carrier is provided by a buzzer-like tone from a relaxation oscillator for the voiced sounds and from a hiss-like sound from a gas-filled tube for

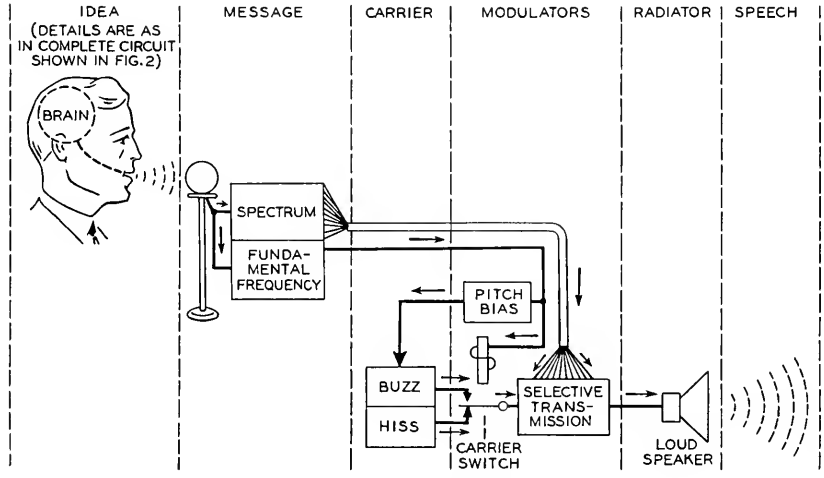


Fig. 7—Schematic circuit of the vocoder.

the unvoiced sounds. In the vocoder, for simplicity's sake, one or the other of these energy sources is used according to whether the sound is voiced or unvoiced, with no provision for the mixed types of sounds found in the human voice. The analyzer of the vocoder derives the

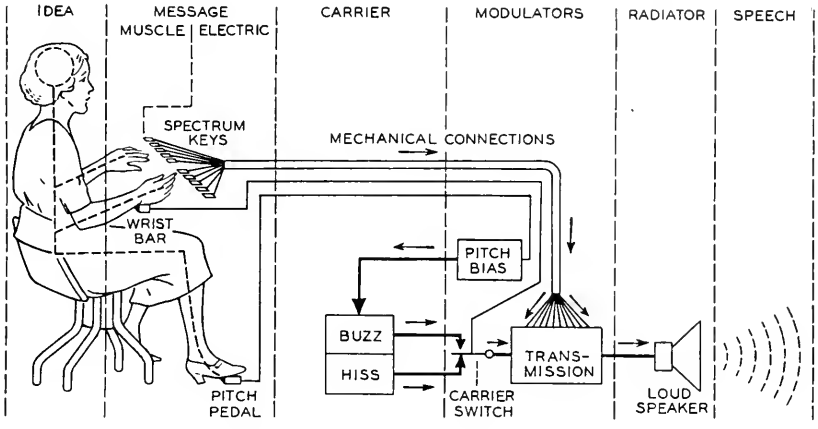


Fig. 8—Schematic circuit of the voder.

original speech message in terms of a modified set of parameters. This analyzer suppresses the original carrier of the talker and so resembles the demodulator in radio reception. The analyzer acts as an electrical ear to tell the artificial vocal system of the vocoder what to say, the whole vocoder acting as a synthetic mimicker.

The basic similarity of the electrical and human speech synthesizers is seen in these figures. In all three cases the message is originated by the brain of the sender of the speech information. There is in each case a transmission of control impulses by the talker's nervous system to the appropriate muscles. The muscles produce displacements of body parts formulating the speech information as a set of mechanical waves. These waves appear in the vocal tract in the case of normal speech; in the fingers, wrist and foot in the case of the voder, but in the case of the vocoder use is made of electrical currents derived from and equivalent to the vocal tract displacements in ordinary speech. In each case the message contains the speech information in syllabic waves. In all cases the message waves control the choice of carrier, the fundamental frequency of the voiced type carrier and the spectrum of power distribution in the speech output. Differences arise in the details rather than in the principles.

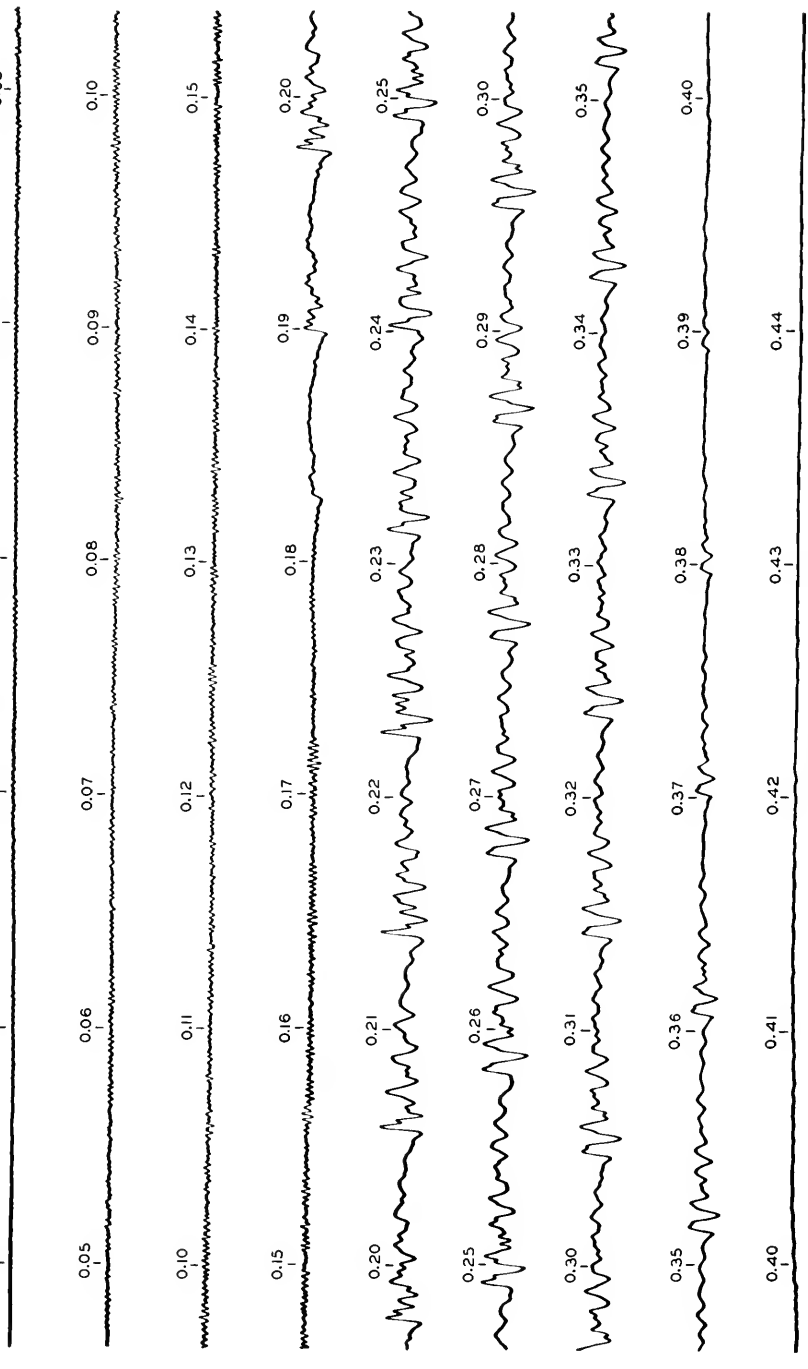


Fig. 9—Oscillogram of the sound "sa."

Speech Characteristics from the Carrier Point of View

Now that the mechanism of speech has been described in carrier terms it is of interest to observe carrier features as they manifest themselves in the characteristics of speech. Some of these can be seen by the eye in speech oscillograms. Some can be demonstrated to the ear with a speech synthesizer such as the vocoder.

For a visual illustration there is shown in Fig. 9 a high quality oscillogram taken from Crandall [10] of the sound "sa" (Plate No. 160. Spoken by M. B.) for a medium-pitched male talker. The carrier shown by the oscillogram is of the unvoiced type for the earlier and of the voiced type for the later part. As one looks at the oscillogram he sees a great mass of the high-frequency components of the carrier. Scrutiny, however, reveals modulated on the carrier the message information in terms of switched energy sources, controlled fundamental frequency and varied transmission characteristic. Shortly after .17 second the switching off of the unvoiced carrier begins. Remnants of the unvoiced carrier can be seen in the voice period just before .19 second and the one starting at about .19 second. The switching on of the voiced carrier appears just after .18 second and seems to be reasonably well completed at the end of the second voice period just before .20 second. This switching was not instantaneous. However, the ear probably does not observe the duration time of the switching. The fundamental frequency falls rapidly at the beginning followed by a leveling out and then a final slight fall in the last few periods. It starts at 140 cycles per second, dropping to around 110 in the level portion, and then to 101 at the end. The resonance conditions cannot be followed too well by eye. However, around .20 second there is a major lower-frequency resonance of about 800 cycles. At .33 second this resonance appears to have increased to 1100 cycles or so. A similar alteration of resonance conditions may be observed if the little shoulder on the rear side of the peak just in front of the .25 second mark is traced in adjacent periods. It can readily be followed back to the third period just before .20 second and can still be seen in the last distinct voicing period starting before .39 second. The dynamic variation of the speech at syllabic rates in accordance with the message content is thus revealed.

For another visual illustration of the speech message Fig. 10 shows a set of oscillograms [11] from the vocoder analyzer for the words "She saw Mary." The oscillogram of the input speech is the trace next to

[10] *Bell Sys. Tech. Jour.*, Vol. 4, p. 586, 1925.

[11] This figure is a copy of Fig. 3 in the paper "The Automatic Synthesis of Speech," Dudley, *Proc. Nat. Acad. Sci.*, Vol. 25, pp. 377–383, July 1939.

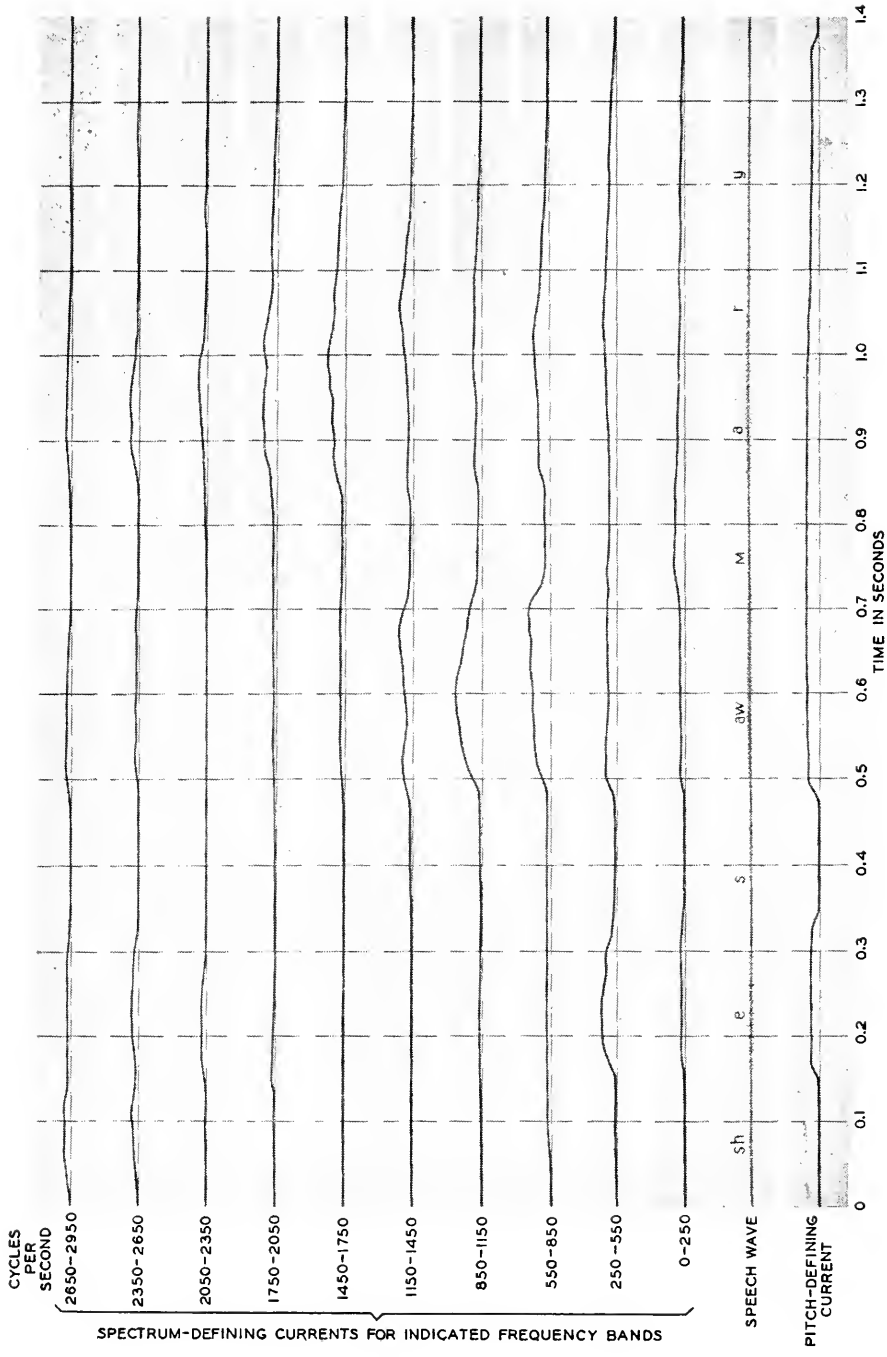


Fig. 10—Coded speech (derived message currents) for the words "She saw Mary."

the bottom. The trace below shows the defining current for the fundamental frequency, while the ten traces above show currents indicating the rectified power in ten frequency bands of 300 cycles width except that the lowest one extends from 0 to 250 cycles. The slow rates of change are noted in the message currents when compared to the original speech wave.

Demonstrations of the vocoder indicate to the ear the carrier nature of speech. Thus the carrier used for remaking speech, whether a monotone or a hiss sound, is observed to have no intelligibility when heard alone. The message currents derived from spoken speech are not audible. However, intelligible "speech" is produced by the modulation of either type of carrier by the message currents of selective transmission. Similarly, there can be used for the carrier a wide variety of sound from the puffs of a locomotive to instrumental music. Upon imprint of the transmission message currents from spoken speech, new forms of odd sounding but nevertheless intelligible "speech" are produced.

The carrier conception of speech reveals what is important and not important in evaluating speech characteristics. An example of interest is the matter of phase. It has long been known that phase was unimportant to the ear at reasonably low listening levels. From the carrier point of view this is natural, for the phase changes referred to are those in the carrier and so, unimportant. When the phases of the message components are altered, there is a very noticeable effect on the ear, for phonetic units are now being shifted.

The great advance in recent years in the application of carrier circuits has been guided by mathematical theory. Since in electrical speech synthesizers the carrier and message currents are separated physically, it is possible to use carrier equations expressing the modulation phenomenon. Similar equations may be written for the voice mechanism as represented by Fig. 6. This has been done in the attached appendix, thus separating speech into syllabic and carrier factors.

APPENDIX

Mathematical Relations

The speech concepts developed in the body of the paper may be expressed in mathematical terms which not only give the fundamental relations in simplest form but also aid in the application of the well-established carrier technique to speech. For voiced sounds, periodic

by nature, the carrier C_v may be written as a function of the time t thus:

$$C_v = \sum_{k=1}^{n} A_k \cos (kPt + \theta_k). \tag{1}$$

Here C_v is composed of n audible harmonics of relatively high frequencies with the kth of amplitude A_k, frequency kP radians per second, and phase θ_k. The choice of fundamental frequency P is somewhat arbitrary but may well represent the average of the talker over the period of interest.

By modulation processes, there is molded on to this carrier the total message information at the relatively low syllabic frequencies. The message is divided into three parts: (*a*) the starting and stopping of the carrier; (*b*) the instantaneous fundamental frequency; and (*c*) the selective transmission through the resonant vocal tract.[12] These three message functions as they manifest themselves in varying the carrier will be represented by s, p, and r, respectively. Equation (1) will be modified to indicate the effect on the carrier of each of these modulations separately, after which the equation will be rewritten to show the effect of all three acting simultaneously.

The effect of starting and stopping the carrier is described mathematically as a function of time by multiplying C_v by the switching function $s(t)$, giving:

$$\text{Switched } C_v = s(t) \sum_{k=1}^{n} A_k \cos (kPt + \theta_k). \tag{2}$$

For simple on-off switching, $s(t)$ alternately equals zero and unity, although it may in general represent more gradual changes or even any variations of intensity over the frequency range.

The instantaneous fundamental frequency is obtained by multiplying P by the inflecting factor $p(t)$. The effect of the frequency modulation [13] is represented by substituting for Pt the integrated quantity

$$\int_0^t Pp(t)dt = P \int_0^t p(t)dt.$$

Writing this value for Pt in equation (1) gives the inflected carrier wave:

$$\text{Inflected } C_v = \sum_{k=1}^{n} A_k \cos \left[kP \int_0^t p(t)dt + \theta_k \right]. \tag{3}$$

[12] As in the body of the paper, the effect of phase modulation is neglected here.

[13] "Variable Frequency Electric Circuit Theory with Application to the Theory of Frequency Modulation," J. R. Carson and T. C. Fry, *Bell Sys. Tech. Jour.*, Vol. 16, p. 513 (1937).

The effect of the selective transmission is allowed for by multiplying C_v by the transmitting factor $r(\omega, t)$, ω indicating that the transmitting factor is a function of frequency at any instant. Applying this factor in equation (1) gives:

$$\text{Transmitted } C_v = \sum_{k=1}^{n} r(\omega, t) A_k \cos (kPt + \theta_k). \tag{4}$$

The r factor is placed inside the summation to indicate that as k changes the different frequencies have different values of the multiplying factor r. If a multiplicity of carrier waves is assumed, the transmitting factor would be $r_k(t)$, individual to the kth component.

In normal voiced speech, S_v, these three modulations are all present simultaneously, so that:

$$S_v = s(t) \sum_{k=1}^{n} r(\omega, t) A_k \cos \left[kP \int_0^t p(t)dt + \theta_k \right]. \tag{5}$$

Equation (5) shows how the message in the form of the s, r, and p functions has imprinted its characteristics on the original carrier C_v of equation (1).

The derivation of (5) was for voiced speech. Unvoiced speech, however, is also covered by (5) as a degenerate case. Nevertheless, further information is presented by writing out the unvoiced carrier separately. For unvoiced speech, the frequency P approaches zero and the number of terms, n, approaches infinity, giving an integral instead of a finite sum of components in equations (1) and (5). The unvoiced carrier C_u is then:

$$C_u = \int_{\omega_1}^{\omega_2} A(\omega) \cos [\omega t + \theta(\omega)] d\omega \tag{1'}$$

and the unvoiced speech:

$$S_u = s(t) \int_{\omega_1}^{\omega} r(\omega, t) A(\omega) \cos [\omega t + \theta(\omega)] d\omega \tag{5'}$$

with the continuously variable frequency ω (radians per second) varying over the audible range of energy contribution from ω_1 to ω_2 and the unvoiced carrier spectrum defined by amplitude $A(\omega)$ and phase $\theta(\omega)$. The unvoiced speech has no inflecting factor but does have switching and transmitting factors to make up the message impressed on the carrier.

Manufacture of Quartz Crystal Filters

By G. K. BURNS

Quartz crystal filters used in modern carrier systems present new problems in manufacturing technique. In the assembly and testing of the filters and in the production of component crystals, coils and condensers, special factory facilities are required for accurate measurement of frequency and control of atmospheric conditions. The manufacture of quartz crystal plates in particular combines several fields of applied science, including crystallography, precision grinding, vacuum technique and high frequency electrical measurement. Inductance coils and fixed and variable condensers for use in crystal filters must consistently meet advanced requirements, especially in regard to stability. The assembly of these components into filters resembles the manufacture of radio receivers, differing mainly because of smaller quantity requirements. Testing equipment must permit rapid shop adjustment and test of the completed filters with laboratory precision.

Introduction

ELECTRICAL wave filters employing quartz crystals [1] are used extensively in broad band carrier systems [2, 3] recently introduced into commercial service. Such crystals exhibit the property of piezo-electricity; that is, an electrical voltage applied to the terminals of a crystal causes a mechanical distortion of the quartz, and vice versa. Because of this interrelation a plate of quartz, at frequencies near its mechanical resonance, behaves electrically like the coil and condenser combination shown in Fig. 1. The series inductance and capacitance

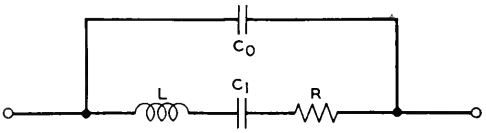


Fig. 1—Equivalent circuit of a quartz crystal plate. Elements L, C_1 and R are associated with the piezo-electric property and mechanical resonance of the crystal, while C_0 represents capacitance between the electrodes.

represent the mass and elasticity of the plate, respectively, while the shunt condenser represents the capacitance between faces of the crystal. The damping of such a plate may be made very low, giving a ratio of reactance to resistance (commonly termed Q) of 15,000 or

[1] Numbered references are listed at end of paper.

higher, as compared with a practical limit of 300 for coils. Stability of resonance frequency and compactness of dimensions are two further respects in which quartz crystals surpass the best coils and condensers available.

Filters designed to utilize these properties generally consist of one or more crystal plates, plus such condensers, inductance coils and resistances as may be required to give the desired overall performance. The principal types used in the Bell System operate at frequencies ranging from 40 to 600 kilocycles and transmit bands varying from 5 cycles to 6 kilocycles in width. Physical dimensions range up to $3 \times 5 \times 16$ inches.

Unusual manufacturing requirements are imposed by the nature of these filters and of the systems in which they are used. Adjusting tolerances and stability requirements, for example, range from $\pm$ 20 to $\pm$ 200 parts per million on crystals and on coil-and-condenser circuits used in crystal filters. Transmission losses must be measured to accuracies of the order of $\pm$ .03 db at 100 KC. To insure stability of adjustment during service life, component apparatus must be protected against dust and excessive humidity. Methods of assembly and testing must be adaptable to a variety of types of filters, one of which, the channel filter,[4] is manufactured by the Western Electric Company in quantities of 1500 to 5000 per year, while the others range from 10 to 1000 per year. Long service life must be assured by proper choice of materials and technique.

In order to satisfy such requirements special manufacturing procedures are necessary. In reviewing these features it will be convenient to consider first those methods or facilities which are used in several or all stages of the manufacture of crystal filters, second the methods employed in producing component apparatus for such filters—particularly crystals, coils and condensers—and finally the technique of assembling and testing the complete filters.

General Facilities

A primary requisite in the adjusting and testing of both crystals and crystal filters is the precise measurement of frequency. The equipment used for this purpose includes a standard frequency generator containing a 100 KC crystal oscillator. This generator normally maintains a frequency accuracy of about 1 part in 2,500,000 operating under the control of the Bell System master frequency standard in New York, but will remain accurate within 1 part in 1,000,000 even though the master signal is interrupted for as much as 24 hours. Three sub-harmonics of 100 KC, namely, 100 c.p.s., 1000 c.p.s. and

10,000 c.p.s., are distributed to all test positions. Oscillators supplying the individual test sets are provided with cathode ray oscilloscopes, by means of which they can be synchronized with any multiple of the three standard frequencies. To set up an odd frequency not coinciding with any multiple it is necessary to interpolate dial readings between two synchronized points.

Control of atmospheric conditions also plays an important part in the manufacture of crystal filters. The temperature coefficient of frequency of the crystals most commonly used is about 15 parts per million per degree Fahrenheit. For some filters, in order to secure uniform performance throughout the temperature and frequency ranges encountered in service, these crystals must be adjusted within tolerances as small as 40 parts per million. Fluctuations of as little as 2° F., in such cases, must be taken into account during the adjustment of the crystals. In addition, crystals, coils and condensers are all sensitive to the effects of excessive humidity. To minimize such difficulties, the assembly and testing of these components and of the filters in which they are used are carried out in air conditioned rooms controlled at 75° ± 2° F. and approximately 40 per cent relative humidity.

Crystals

Of the several component parts used in crystal filters, the first to be considered in detail are logically the quartz crystals themselves. Their properties of low loss and high stability are primarily responsible for the unusual performance of filters in which they are employed.

Natural deposits in the earth constitute the sole source of supply of quartz crystals, since no practical method of producing them synthetically has been developed. "Raw" crystals suitable for use in filter manufacture must be unusually large and free from flaws. The principal source is Brazil, the bulk of the quartz being brought in by native prospectors to trading posts and shipped to this country via Rio de Janerio and other coastal cities. The crystals usually range between 3 and 10 pounds in weight, with occasional pieces reaching 100 pounds.

The raw quartz passes through successive stages of inspection and selection, commencing at the trading post and culminating in careful examinations before and during the cutting operations. A concentrated beam of light from an arc lamp (see Fig. 2) is used in locating internal flaws, which generally appear as small bubbles and inclusions of foreign matter. Quartz takes two distinct forms, left-hand and right-hand, having opposite piezo-electric polarities. Portions of raw crystals containing both forms are not usable. This condition, called

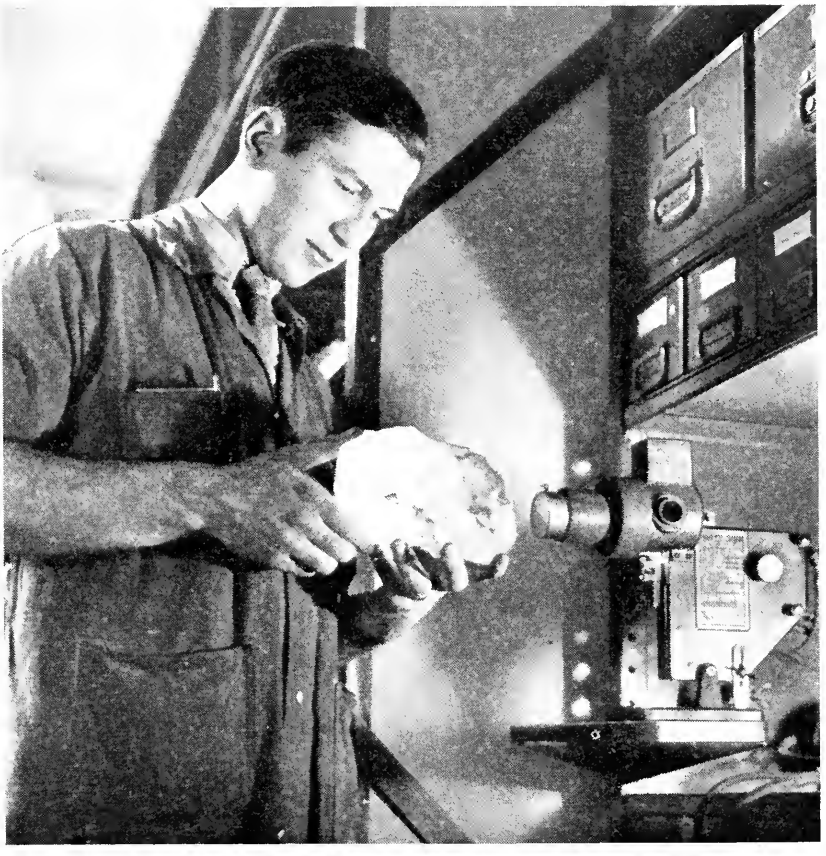


Fig. 2—Inspection of quartz crystals. An arc light beam aids in the detection of internal flaws.

"twinning," appears as shown in Fig. 3 when observed with polarized light.

For use in filters, quartz must be cut into rectangular plates properly oriented with respect to the electrical, mechanical and optical axes of the crystal, as shown in Fig. 4. A polariscope and an X-ray spectroscope are used in locating these axes to an accuracy of $\pm$ 0.25 degree. For the majority of applications the plate is cut in the plane of the mechanical and optical axes, with the long dimension set at an angle of 18.5° from the mechanical axis. This orientation eliminates secondary resonances in the completed crystal and makes the primary resonance frequency relatively independent of slight errors in orientation. For applications requiring a low coefficient of resonance frequency versus

Fig. 3—Right and left-hand twinning in quartz as seen by polarized light.

temperature, plates are cut with their long dimension 5° from the mechanical axis. Tolerances in cutting and grinding to thickness, length and width prior to calibrating are of the order of .001 mm., requiring the use of technique similar to that employed in the manufacture of gage blocks. A few standard thicknesses, ranging from .020 to .060 inch, are used for most crystal plates. Lengths vary from 0.5 to 2.0 inches while widths range from 0.15 to 1.5 inches. Because of unavoidable waste in the cutting and grinding operations and the rejection of quartz containing flaws, only a small portion of the material entering the cutting room finds its way into finished plates.

Up to this point the cutting and grinding are purely mechanical operations, directed toward securing prescribed physical dimensions. During final adjustment and in service, however, the crystal plate must be connected as an electrical element. Electrodes are provided by coating the major surfaces of the plate with aluminum, using a process of evaporation and condensation in a vacuum, similar to that employed in the silvering of telescope mirrors. If the plate is to be used in a balanced filter section which requires a pair of crystal elements of the same frequency, as is frequently the case, the plating on each face is then divided in half along the longitudinal axis. This division, one-hundredth of an inch wide, must have a d.c. insulation resistance of at least 100 megohms to insure proper operation in some types of crystal filters.

Preliminary tuning is accomplished with a fixture, simulating the final holder, which grips the plate at the center by four contact points, one on either side of the division in the plating of each face. These

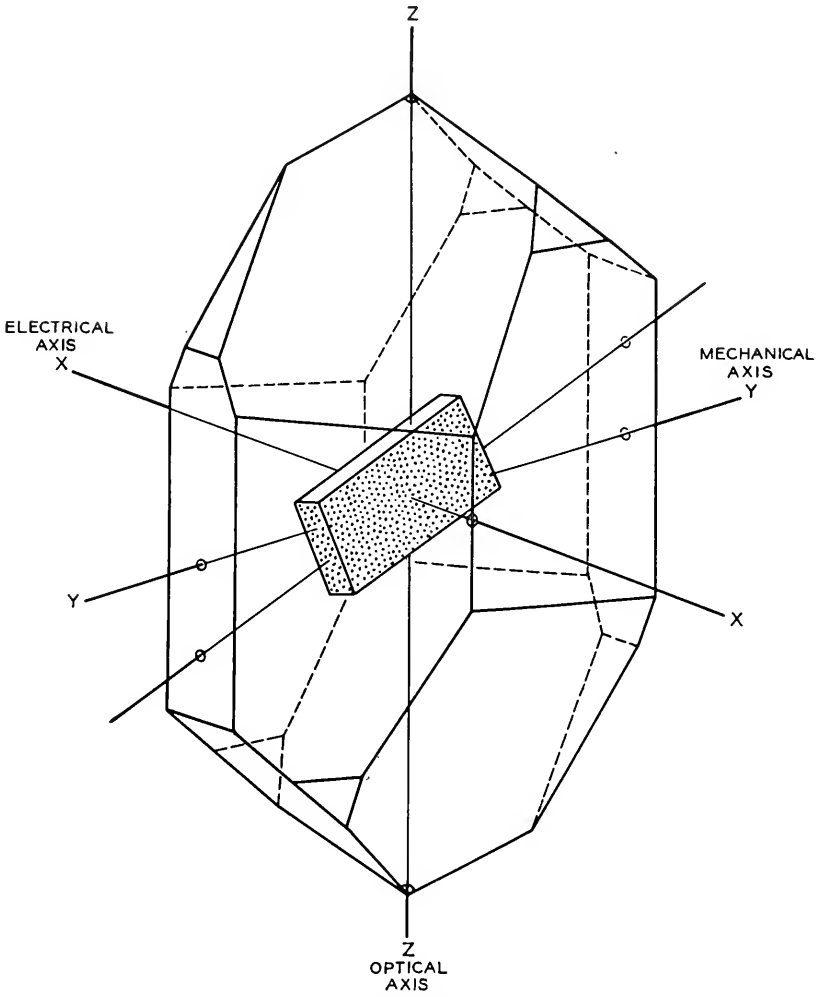


Fig. 4—Orientation of a typical quartz plate with respect to its electrical, mechanical and optical axes.

contacts introduce very little damping, since the mode of vibration normally employed is longitudinal, with maximum amplitude at the ends of the plate and a node at the center. The test set-up normally used consists of two oscillators with a meter arranged to read the

difference between their frequencies. One oscillator is controlled by the crystal plate being tuned and operates at its resonance frequency; the other is controlled by a standard crystal of the desired frequency. Starting 100 to 200 cycles low, the plate is ground on the ends until its frequency approaches that of the standard.

The plate is then transferred from the fixture to its final holder, shown in Fig. 5. This mounting normally accommodates two plates

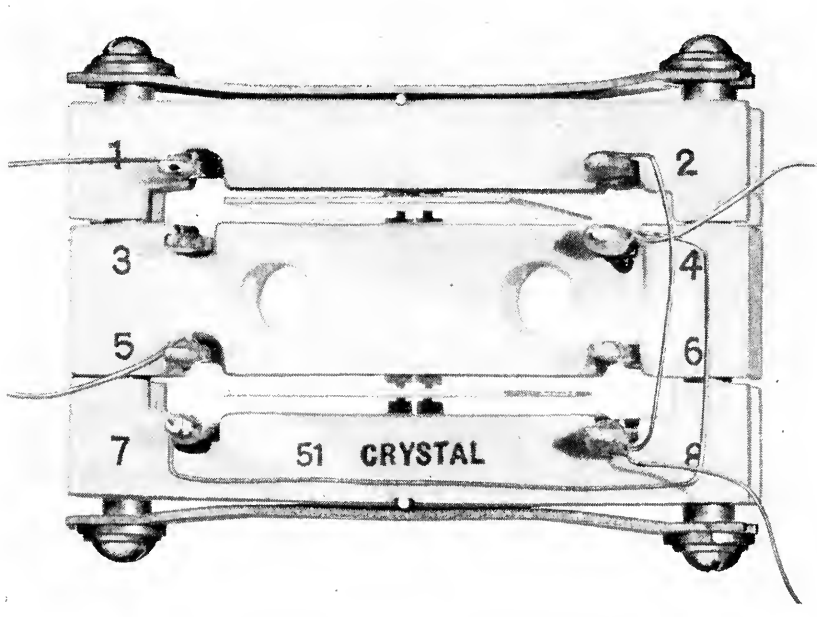


Fig. 5—Crystal plates mounted in holder. The four points at the center of each plate provide electrical contact and mechanical support.

of different frequencies, each supported at its nodal point by contacts projecting from ceramic blocks. The entire assembly is held together by a spring suspension in order to apply uniform pressure at all contacts. To minimize damping, the contacts must be accurately aligned and the quartz plates must be carefully centered upon them.

A final adjustment of frequency is now performed, as shown in Fig. 6. Permissible tolerances vary from $\pm$ 20 to $\pm$ 150 parts per million for different types of crystals. Crystals having the broader tolerances and substantial quantity requirements are adjusted by comparison with a standard crystal, as in the case of preliminary tuning. The test set shown at the left in Fig. 6 is being used for this purpose. The upper and lower panels are the oscillators controlled by

the standard and the test crystals, respectively, while the center panel indicates the frequency difference between them. For very accurate work and for periodic checks of the standard crystals it is necessary to use a precision oscillator, shown at the right.

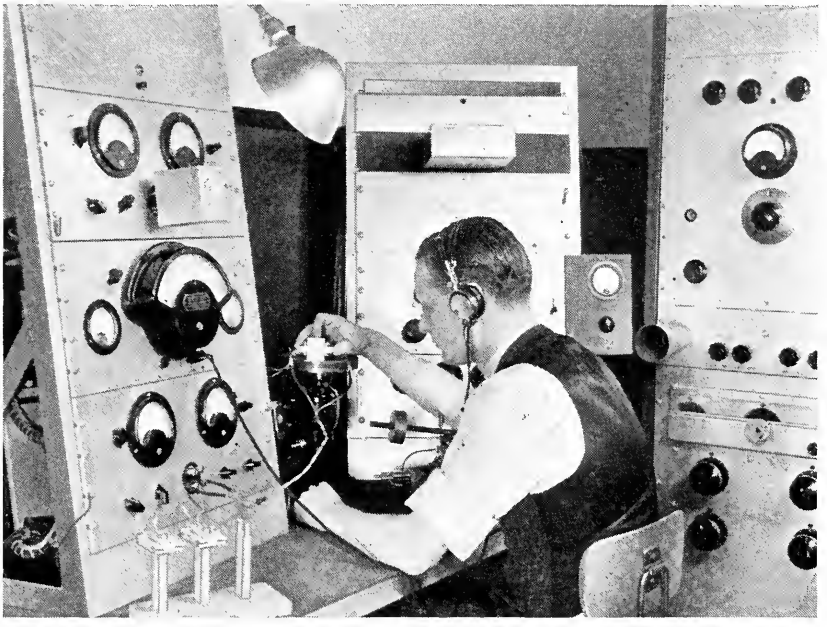


Fig. 6—Final tuning of a crystal plate using a standard crystal (in small box at upper left) for comparison.

Occasionally, in the course of adjustment, plates are carried too high in frequency. In such instances, as a result of the standardization of thicknesses mentioned previously, the plate normally can be salvaged by grinding it to the dimensions of the next higher frequency plate of the same thickness.

Aging occurs in both resonance frequency and effective resistance, as slight strains created in the quartz and in the contacts during adjustment relieve themselves. The greater part of the aging takes place during the first few hours after calibration and nearly all of it during the first week. In general, the frequency rises a few cycles and the resistance drops slightly. Crystals on which the frequency tolerance is approximately equal to the shift due to aging are stabilized by one or more temperature cycles, prior to final measurement of frequency and resistance.

Coils

Inductance coils are used in some crystal filters, particularly the channel filter for the newer types of carrier telephone systems. Since it may be necessary to connect as many as ten such filters in tandem in a long-distance circuit without appreciable impairment of the quality of transmission, the filters must meet exacting requirements not only

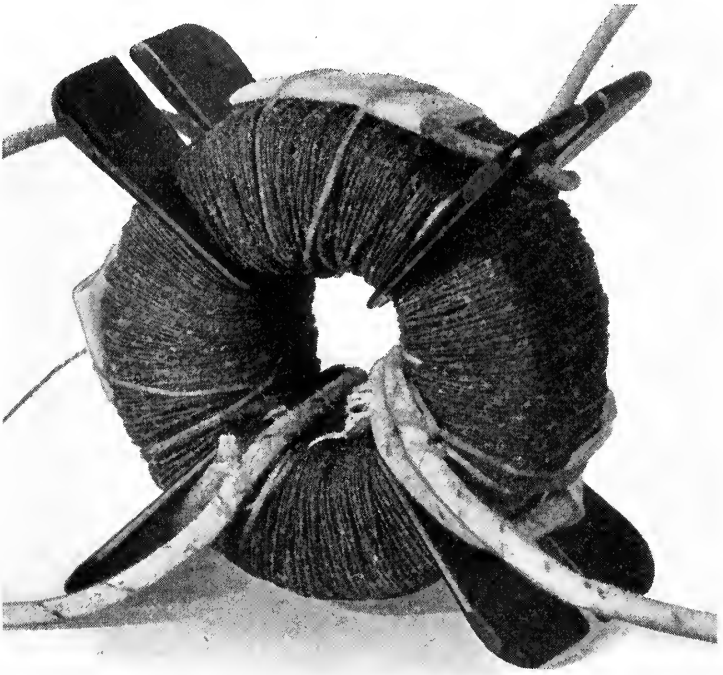


Fig. 7—Toroidal inductance coil used in crystal channel filter, shown before potting.

at the time of their manufacture but throughout service life. Consequently inductance coils used in the filters must exhibit little aging or shift with temperature, either in inductance or in effective resistance. Losses must be kept low in order to meet a Q requirement of approximately 200. The types employed in channel filters range from 25 to 50 millihenries in inductance and from 60 to 120 kilocycles in operating frequency.

Unusual features of design and manufacture are employed to meet these requirements. The coil is essentially a toroidal winding with low distributed capacitance, applied to a permalloy dust core, im-

pregnated and potted in wax. A molded jacket with protruding fins, placed around the core, reduces the capacitance from windings to core and improves the uniformity of the windings. The coils are adjusted to within ± 1 per cent for inductance and 2 per cent for inductance unbalance by removal of excess turns, all adjustments being made at low frequency. Figure 7 shows a coil at this stage of manufacture.

The coil is then potted in a copper can and a cover soldered in place. Final test simulates actual service conditions. The coil is resonated with an external variable condenser at the operating frequency for which it is designed.

Condensers

Nearly all crystal filters contain condensers shunted across the crystal elements. These condensers must meet stability requirements similar to those already mentioned in connection with coils.

One form of fixed condenser, used where small values of capacitance and high stability are required, is illustrated in Fig. 8. Silver is fused

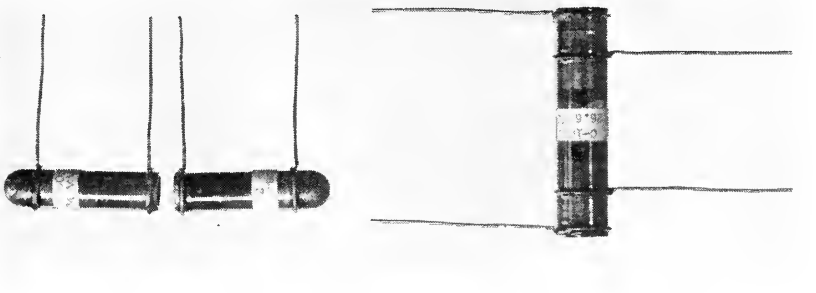


Fig. 8—Silvered glass condensers used in crystal filter applications where high stability is required.

to the inside and outside of a glass tube by applying a coating of silver paste and firing the tube in an oven. A gap is left uncoated on the outer surface near the open end and leads are soldered to the silver on both sides of the gap. The capacitance is then adjusted to the required value, within approximately ± 1.5 mmf., by scraping off a portion of the silver coating. Capacitances up to 80 mmf. are realized by this means. Two condensers may be combined in a single unit, as shown at the right in Fig. 7. The completed condenser is dipped in varnish to protect the silver from corrosion.

Pairs of such condensers, matched to each other within 0.4 mmf., are required in some types of crystal filters. This precision is achieved by manufacturing a quantity of condensers of the correct nominal capacitance and sorting them into close-limit groups after final measurement.

Variable air condensers are used in adjusting the assembled filter. For the channel filter four such condensers are manufactured on a single ceramic base, as shown in Fig. 9, to eliminate unnecessary parts

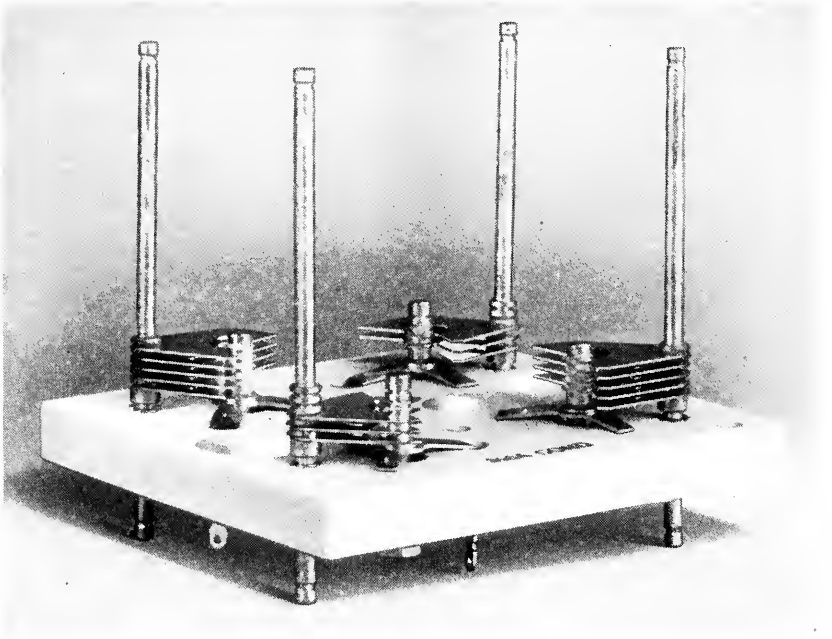


Fig. 9—Four-section variable air condenser.

and reduce assembly cost. The posts supporting the stator plates are extended both upward and downward to serve as convenient terminals for leads from adjacent pieces of apparatus. Freedom from binding is important, since condensers in crystal filters must be adjusted through angles as small as 2 minutes. To insure smooth adjustment the rotor shafts and their bearings are held to close dimensional tolerances and lubricated with petrolatum. Stability is secured by the use of thrust springs providing a substantial holding torque.

Assembly

The foregoing components—crystals, coils and condensers—are assembled into complete filters by methods somewhat similar to those employed in the manufacture of radio receivers, the principal differences arising from smaller volume requirements. The channel filter alone is produced in sufficient quantities (1500 to 5000 per year) to warrant a substantial degree of tooling. Figure 10 illustrates the

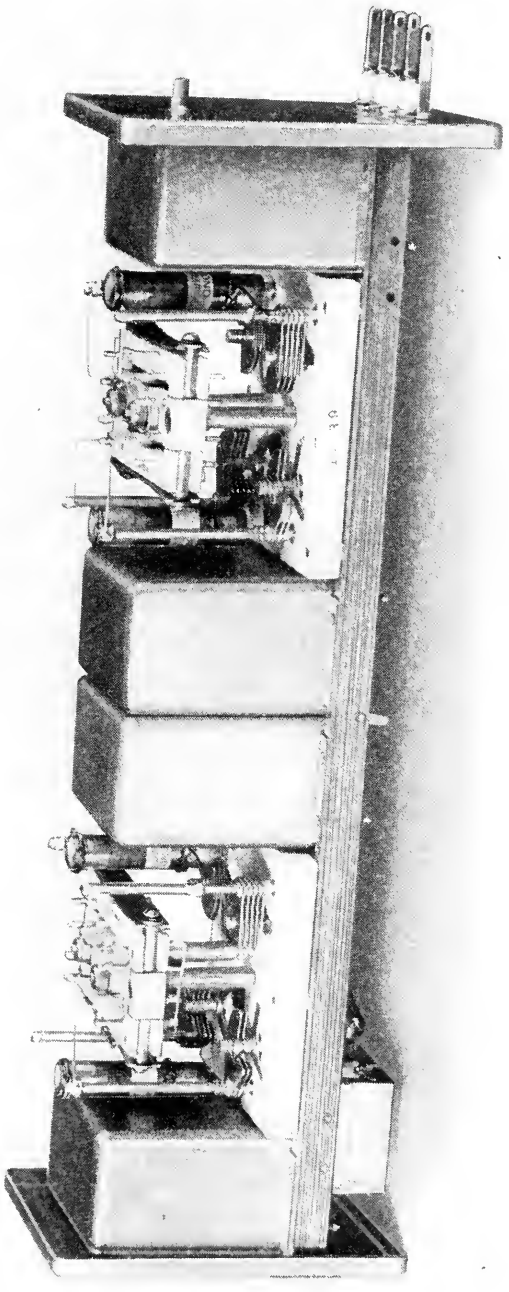


Fig. 10—Internal assembly of crystal channel filter, used in Type *J*, *K* and *L* carrier systems.

internal assembly of this type of filter. The chassis consists of a pair of perforated brass angles running the length of the assembly and spot-welded to a cover at each end. Coils and condensers are riveted to the angles, while the crystal holders are mounted with rubber shock absorbers on studs extending upward from the ceramic bases of the variable condensers. External leads are brought out through copper-to-glass seals to terminals, as shown in Fig. 11, since in final assembly

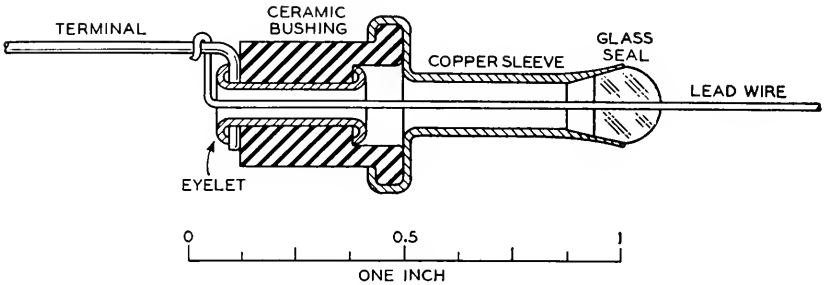


Fig. 11—Terminal used in hermetically sealed filters. The copper sleeve, bonded to the lead wire by means of an insulating glass bead, is soldered into the container of the filter in final assembly.

the filter must be hermetically sealed to protect components from moisture and dust.

In wiring the filter special precautions are taken to prevent foreign materials from being deposited on crystal plates, thereby introducing mechanical damping, and from lodging in variable condensers, where electrical leakage must be avoided. Internal connections are made with bare tinned wire. Rosin flux remaining on soldered connections is washed off with a solvent. Dust and other particles in variable condensers are blown out with air. The filter then undergoes a careful visual inspection and a 500 volt d.c. insulation test.

At this stage it is generally necessary to adjust certain of the filter elements, usually variable condensers, in order to compensate for manufacturing variations in other elements and for parasitic effects such as capacitance of the wiring to ground. A general view of the testing equipment used for this purpose is shown in Fig. 12. One or more of three methods of adjustment are employed, namely, (*a*) transmission loss, (*b*) resonance and (*c*) capacitance. The first and second of these are utilized on the channel filter, the schematic of which is shown in Fig. 13. The two sections are adjusted independently before the resistance pad, seen at the center of the figure, is inserted to connect them.

In transmission loss adjustment of the channel filter, the attenuation

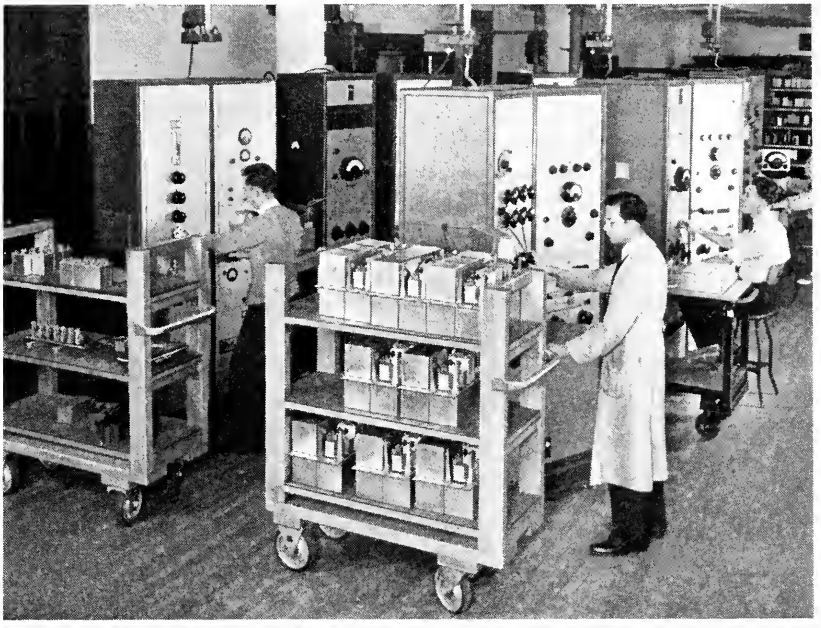


Fig. 12—Filter testing area. The standard-frequency outlets are seen above the test sets.

of each filter section is brought to a peak at a specified frequency. The filter is placed in a test shield simulating its final container. Voltage from a precision oscillator is applied to the input terminals of the section and the voltage at the output terminals is measured with a sensitive detector preceded by a variable attenuator. The condensers designated C_{TL} in Fig. 13 are adjusted with a non-metallic tool until the loss reaches a peak of 50 to 70 db. Each section contributes two such peaks.

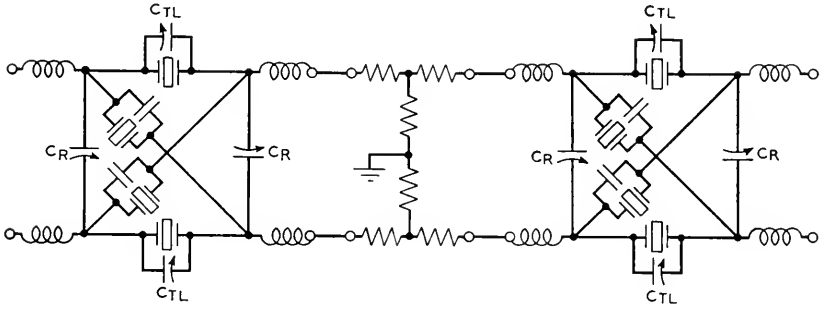


Fig. 13—Schematic of crystal channel filter.

The filter is then transferred to a resonance bridge and the impedance looking into either end of each section, with the opposite end open-circuited, is adjusted to series resonance at a specified frequency by means of the condenser C_R. This adjustment primarily controls the shape of the loss characteristic of the filter in the transmission range. The resistance at resonance is recorded for later reference.

In some types of filters, adjustment must be made to secure the correct absolute capacitance between certain points in the filter rather than to obtain desired attenuation peaks or resonances. A capacitance bridge is employed for this purpose.

In all of these adjustments, test leads connecting the filter to the test set play an important part. Shielding, balance, capacitance to ground, dielectric loss, stability and other characteristics of the leads must be carefully controlled or compensated by adjustments within the test sets in order to meet precision requirements of the order of $\pm$ 0.01 per cent.

After adjustment, in the case of the channel filter, the individual sections are connected through a resistance pad selected to complement the values of resistance measured during resonance adjustment. Uniformity of overall transmission loss, regardless of manufacturing variations in components, is secured by this means.

The completely wired filter is now placed in a copper shell and hermetically sealed with solder, except for an inlet and an outlet vent. In order to remove vestiges of moisture which might affect the crystals or other components during service life, a current of air of less than 3 per cent relative humidity is then passed through the filter for 12 hours and the vents are sealed off.

Final test consists of measurements of transmission loss at a series of frequencies in the transmission and attenuation bands of the filter, using equipment similar to that on which the peaks were adjusted. The variety of product which must be tested with these facilities demands maximum flexibility and minimum set-up time. This requirement is met with plug-in terminating impedances, pads, leads, etc., and with oscillators and detectors tuning continuously over a wide range of frequencies. Several filters of the same type are normally tested simultaneously, all being measured at one frequency before the next frequency is set up. Contact fixtures for particular types are provided when justified by quantity requirements, in order to facilitate the transfer of test leads from one filter to the next.

Transmission loss characteristics of channel filters under various conditions are shown in Fig. 14. The solid curve illustrates a normal filter. The loss in the passband is approximately 5.6 db, with distor-

tion of about 0.25 db over a band 3 KC wide. The peak losses are from 75 to 90 db, with the intervening "valleys" approximately 65 db. The other curves illustrate three types of defects occasionally observed, namely: (*a*) displacement of one attenuation peak caused by

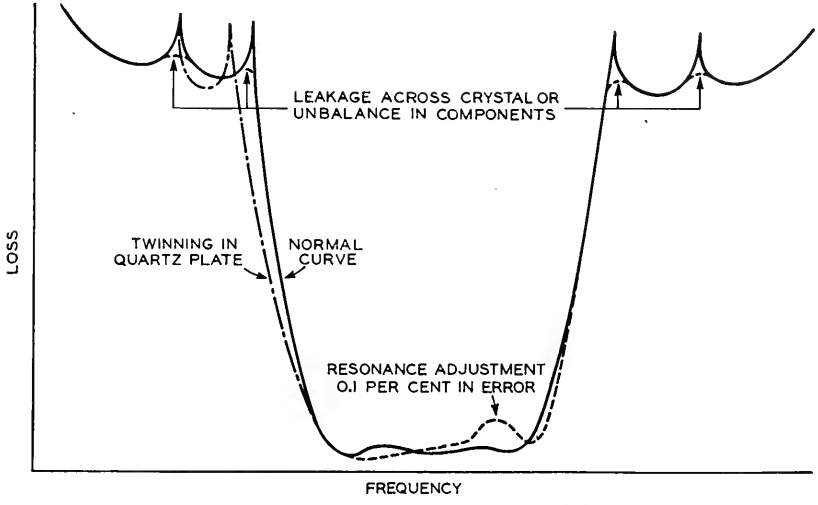


Fig. 14—Insertion loss characteristics of crystal channel filters, showing the effects of deviations from normal conditions.

twinning in one of the crystal plates, (*b*) abnormal distortion caused by a 0.1 per cent error in resonance adjustment, and (*c*) low loss at peaks caused by leakage across a crystal or by components which are inadequately balanced to ground. As an aid to locating the particular components or adjustments which are responsible for such defects, a catalogue of "trouble-shooting" instructions, arranged by classes of filters and types of symptoms, has been compiled.

Cleaning, finishing and labelling constitute the remaining operations on crystal filters. High temperature processes such as vapor degreasing and baking of the finish are inapplicable here because of the nature of the component apparatus in the filter. The surface is scratch-brushed, washed with a solvent and sprayed with aluminum lacquer. Rubber stamps and printers' ink are then used to apply the terminal and type designations.

Conclusion

Crystal filters exemplify the trend toward higher frequencies and higher precision in modern carrier systems. These advances in design have required the development of new manufacturing processes and refined methods of adjusting and testing, and demand increased care

and skill in every step. Ten years ago this technique had not reached even the laboratory stage. Today, in the commercial production of crystal filters, it has become commonplace to deal with capacitances expressed in tenths of a micromicrofarad, transmission losses in hundredths of a decibel, crystal dimensions in thousandths of a millimeter and frequency measurements in thousandths of a per cent. The attainment of such precision at moderate shop cost is the primary engineering problem in the manufacture of the higher frequency types of carrier telephone facilities.

Bibliography

1. "Electrical Wave Filters Employing Quartz Crystals as Elements," W. P. Mason, *Bell System Technical Journal*, 1934, pages 405–452.
2. "A Carrier Telephone System for Toll Cables," C. W. Green, E. I. Green, *Electrical Engineering*, 1938, pages 227–236.
3. "A Twelve-Channel Carrier Telephone System for Open-Wire Lines," B. W. Kendall, H. A. Affel, *Bell System Technical Journal*, 1939, pages 119–142.
4. "Crystal Channel Filters for the Cable Carrier System," C. E. Lane, *Electrical Engineering*, 1938, pages 245–249.

Results of the World's Fair Hearing Tests

By J. C. STEINBERG, H. C. MONTGOMERY, and M. B. GARDNER

A hearing test for musical tones formed part of the Bell System exhibit at the New York and San Francisco Fairs, and the test records obtained have made possible a study of the hearing of a large group of the United States population. The variation of hearing acuity with age and sex is described in considerable detail. Hearing is also related in lesser degree to other factors, such as place of residence, economic status, and race, and these relations are discussed.

The data are applied to the United States population by indicating certain allowances which should be made for differences between the Fair groups and the population, particularly with respect to distribution of ages and economic status.

Accuracy of the test is discussed in relation to ability of visitors to understand the test procedure, disturbing effect of background noise, and calibration of the test equipment.

Certain results of the survey are expressed in terms of ear canal pressure and equivalent free field intensity, and on this basis a comparison is made with the results of other surveys of hearing.

A criterion is given for deciding how much hearing should vary from average before being considered abnormal. Application of this criterion indicates, in the case of children, a suggestive similarity between incidence of adenoid growth as reported in medical surveys and abnormal hearing for high frequency tones.

WITH the opening of the New York and San Francisco World's Fairs in 1939, an opportunity became available for a survey of hearing in a large group of the United States population. One of the Bell System Exhibits consisted of a hearing test whereby visitors could test their hearing for tones of musical pitch. At the end of the test, each visitor was asked to permit an attendant to make a photographic copy of his hearing test card so that a study of the records might be made. Before making the copy, the attendant indicated by a check mark whether the visitor was male or female, colored or white, and to which of the five age groups, 10–19, 20–29, 30–39, 40–49, or 50–59, she judged him to belong. In all, some 550,000 photographic records were obtained, and it is estimated that about 80 per cent of the visitors who tested their hearing for musical tones cooperated in the survey. A somewhat similar test for spoken words was also provided, but the survey was concerned principally with the results of the musical tones test.

The value and usefulness of such a large collection of records is dependent very directly upon the accuracy of the test. Therefore considerable attention was given to the calibration of the hearing test equipment and to the evaluation of factors which might affect the results of the test. There seems little doubt that the records accurately portray the hearing characteristics of that section of the population taking the tests.

One of the principal objectives of the study was to determine the hearing acuity and the prevalence of defective hearing in the United States population. The visitors who tested their hearing were not a representative sample of the population with respect to factors affecting hearing. Consequently a second objective was to determine the relation of hearing to such factors as age, sex, place of residence, economic status, etc. This information is necessary in order to apply the Fair data to the whole population or to specialized groups within the population.

It is believed that the two important factors, age and sex, have been satisfactorily evaluated. Information on other factors although less complete, is sufficient to justify many applications of the data. In other applications, it is necessary to make reservations and these are described in the text.

Description of the Test

The tests were made in sound-insulated rooms arranged to seat seven visitors, each partially screened from the others, as shown in Fig. 1. The test and suitable instructions were recorded on phonograph records and given through a telephone receiver which the visitor held to his ear. In the musical tone test a pure tone was sounded one, two, or three times, and the listener was instructed to write in a space on a form that was given him the number of times he heard the tone. For a given pitch, nine such sets of tones were sounded, each set fainter than the preceding one. When the tones became too faint to be heard, the listener could not write the number correctly, and thus a measure of his hearing acuity was obtained. This test was made with tones of five different frequencies in the following order: 440, 880, 1760, 3520, and 7040 cycles. A typical hearing test record is shown in Fig. 2.

The correct numbers, which appear in the spaces between the columns of Fig. 2, were printed on the back of the blanks in such a way that they would show through in these spaces when the blank was placed on a brightly illuminated glass shelf. The designations "normal or good," "slightly impaired" and "impaired," which show through

Fig. 1—Interior view of one of the hearing test booths.

opposite the eighth, fifth, and second steps respectively, were intended to give a qualitative indication of hearing acuity. Thus the visitor could correct his own test and obtain an indication of his hearing acuity for each frequency.

A scale of hearing acuity for expressing the results of the tests was set up, using as a zero or reference level the mean test score of men and women at the Fairs in the age group 20 to 29. Hearing acuity is expressed as a hearing loss in the usual way, i.e., the departure in db of a given test result from the reference level. As shown in a later section, this reference level gives an ear canal pressure which corre-

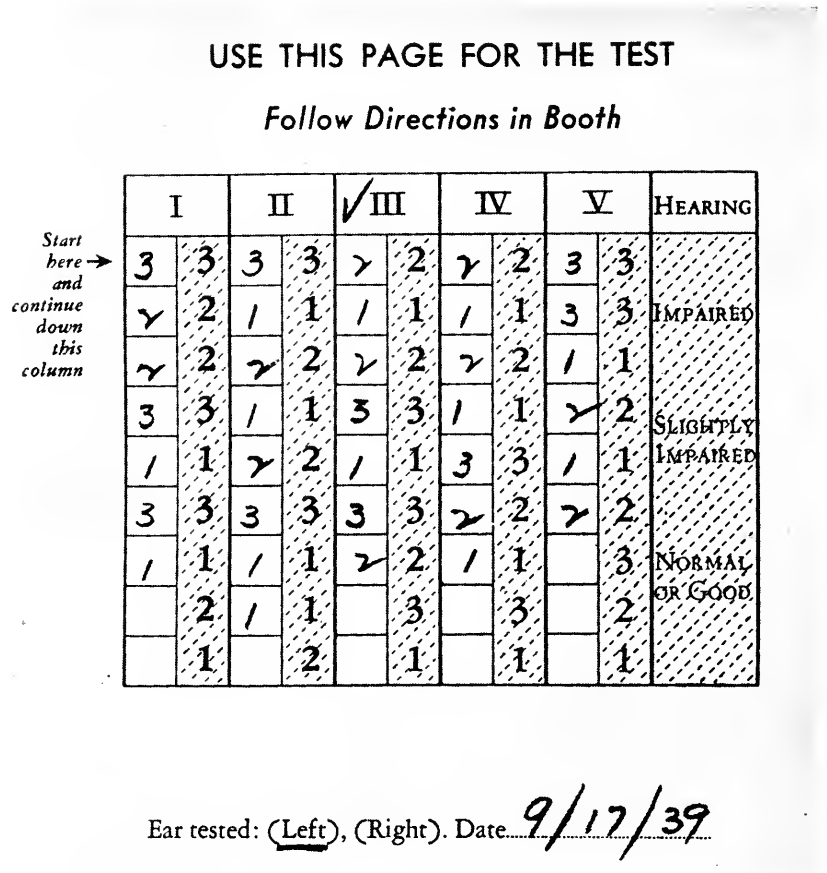


Fig. 2—Typical hearing test record as it would appear when illuminated from underneath.

sponds closely to that for zero hearing loss on the 2A Audiometer.[1] Hence hearing losses given here are comparable in magnitude with audiometric measurements.

The range of the test is shown in Table 1, which gives the hearing loss corresponding to each test step.[2] The range covered is 62 db in the first four columns and 48 db in the last.

[1] Some confusion has been occasioned by referring to the audiometer zero as normal hearing. Actually it is supposed to represent average normal hearing, where normal hearing refers to a range about the average value. If an individual has a hearing loss, his hearing is not necessarily abnormal. The amount of hearing loss which should be considered abnormal is discussed in a later section.

[2] The hearing loss values given in Table 1 correspond to the actual tone levels in the test. Throughout this paper it is assumed that the threshold of an individual lies between the last level at which he correctly records the number of tones and the first level at which he does not, and in computing mean values the loss is reckoned mid-way between these two levels.

TABLE 1

HEARING LOSS FOR EACH STEP IN THE MUSICAL TONE TEST

Column Frequency	I 440	II 880	III 1760	IV 3520	V 7040
Step 1	52*	52	52	46	33
2	42*	42	42	36	27
3	32	32	32	26	21
4	22	22	22	16	15
5	14	14	14	8	9
6	8	8	8	2	3
7	2	2	2	−4	−3
8	−4	−4	−4	−10	−9
9	−10	−10	−10	−16	−15

* These tones were used in the instructions for the test.

The voltage levels used at each frequency in the hearing test were selected so that the average young person would be able to hear the tones on the first six or seven steps, but would miss the last two or three.

RELATION OF HEARING TO AGE AND SEX

The relation of hearing loss to age and sex may be summarized by giving the average hearing loss. More detailed information may be obtained from tables giving the frequencies of occurrence of different amounts of hearing loss. In this section, both types of data are given for men and women separately, in five age ranges.

Trends in Average Hearing

Average hearing as indicated by the mean hearing loss for men and women in five different age groups is shown in Table 2. The number

TABLE 2

MEAN HEARING LOSS IN DB

Age Group		Frequency					Number of Tests
		440	880	1760	3520	7040	
Men	10–19	1.0	.3	−.3	−1.2	−.4	4132
	20–29	.0	−.2	−.1	2.0	1.5	3287
	30–39	1.4	1.3	2.3	8.2	7.7	3197
	40–49	3.7	4.5	7.0	17.7	16.8	4528
	50–59	6.8	7.7	12.1	25.6	24.0	1935
Women	10–19	.5	.2	−1.1	−4.4	−3.6	3417
	20–29	.0	.2	.1	−2.0	−1.5	4208
	30–39	2.6	2.6	2.9	2.4	4.8	3978
	40–49	6.0	5.8	6.7	7.8	11.9	4369
	50–59	10.3	9.8	11.0	13.8	19.7	2538

of tests used in obtaining the mean values is given in the right-hand column. For each age group, they were selected in a random manner from the New York and San Francisco tests, about two-thirds from New York and one-third from San Francisco.[3]

Certain trends in average hearing are evident in Table 2. On the average, both men and women show increasing hearing impairment with increasing age. For high-frequency tones, and especially at 3520 cycles, the effect is more pronounced in men than in women, but for

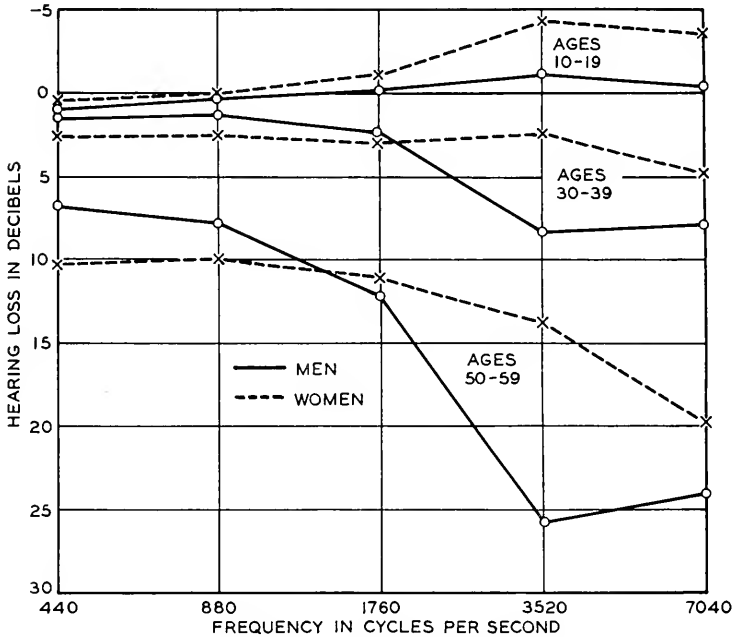


Fig. 3—Mean hearing loss in db for men and women in the youngest, middle, and oldest age groups.

low tones the opposite is true, although to a smaller degree. For the 1760-cycle tone there appears to be little difference between the hearing of men and women. These trends are shown in graphical form in Fig. 3, for the youngest, middle and oldest age groups.

At the lower frequencies, the hearing of the youngest group in Table 2 is slightly poorer than that of the next older group. It is

[3] Several sampling procedures were used, all based on selection of tests by some arbitrary rule, such as taking four tests in order then skipping twelve. In general, the same rule was used throughout a whole day's tests. The days selected were well scattered throughout the season, and week days and week-end days were used in the proper proportion. A larger sampling proportion was used for the older age groups, to make the groups in the sample more nearly equal in size.

believed that this is due principally to the greater difficulty of the younger children in understanding the test and writing their responses on the test blank.

TABLE 3

STANDARD DEVIATION OF HEARING LOSS IN DB

Age Group		Frequency				
		440	880	1760	3520	7040
Men	10–19	10.9	8.7	9.4	12.7	14.2
	20–29	8.5	7.1	8.7	13.8	14.5
	30–39	10.1	7.9	9.9	16.6	16.2
	40–49	11.6	11.0	13.4	19.3	17.0
	50–59	12.1	12.6	15.8	19.2	15.1
Women	10–19	10.6	8.6	8.6	9.4	11.6
	20–29	9.3	7.9	8.5	10.1	11.8
	30–39	11.1	9.9	10.2	12.1	13.8
	40–49	12.4	11.8	11.9	14.1	15.8
	50–59	14.0	14.0	13.9	16.2	15.4

Table 3 shows the standard deviations of the hearing losses for single tests. They range in magnitude from 7 to 20 db, and tend to increase with increasing age and tone frequency. An exception occurs for the 440-cycle tone where the values are mostly larger than for the 880-cycle tone. Since previous surveys have not shown a tendency for the standard deviation to increase below 880 cycles, it seems likely that the present increase occurred because the 440-cycle tone was the first one in the test, and initial unfamiliarity produced a greater scattering of the results. Carelessness in holding the receiver snugly against the ear would produce a similar scattering of results at this frequency.

The mean hearing losses and standard deviations for the older groups for the two high-frequency tones would be somewhat larger were it not for the restricted scale of the test. At 3520 cycles the range from zero is only 46 db, and at 7040 cycles, only 33 db. In computing means and standard deviations all results lying beyond the range of the test were grouped one test step beyond the extreme value included in the test.

Distribution of Hearing Loss

In addition to giving the trends in average hearing, the hearing test scores afford a means of determining the frequency of occurrence of different amounts of hearing loss. A convenient method of presenting the occurrence rates of different degrees of deafness is by means of curves or tables showing cumulative distributions. Table 4 shows

TABLE 4

DISTRIBUTION OF HEARING LOSS

Percentage of tests failing to show a correct response at the test step indicated, based on 23,320 tests from the N. Y. Fair and 12,269 tests from the S. F. Fair.

Frequency	Test Step	Hearing Loss	Men, Age Group					Women, Age Group				
			10–19	20–29	30–39	40–49	50–59	10–19	20–29	30–39	40–49	50–59
440	1	52*	—	—	—	—	—	—	—	—	—	—
	2	42*	—	—	—	—	—	—	—	—	—	—
	3	32	2.7	1.2	2.5	4.1	5.4	2.5	1.5	3.5	5.4	9.9
	4	22	4.6	2.1	4.2	7.1	9.8	4.1	3.0	6.1	10.7	17.9
	5	14	9.2	5.3	8.3	13.9	21.2	8.4	6.6	11.8	20.9	33.5
	6	8	16.0	11.0	14.7	22.3	32.1	15.2	12.1	19.4	30.9	44.3
	7	2	40.0	32.3	38.7	46.7	60.9	37.7	33.5	43.3	56.2	68.9
	8	−4	67.1	67.9	71.3	77.6	85.7	64.7	65.1	72.9	81.8	88.1
	9	−10	86.2	92.2	93.6	95.1	97.8	84.9	90.7	94.0	96.2	98.0
880	1	52	.4	.1	.1	.7	1.4	.4	.2	.8	1.2	1.9
	2	42	.7	.2	.4	1.8	3.2	.6	.5	1.4	2.6	5.3
	3	32	.9	.6	.9	3.2	5.4	1.0	1.1	2.1	4.1	7.6
	4	22	2.1	1.3	2.3	6.8	12.0	2.2	2.1	4.4	8.7	16.1
	5	14	4.9	2.9	5.5	13.5	21.6	4.6	4.3	8.5	16.0	27.1
	6	8	10.3	7.4	11.7	23.0	33.6	9.3	8.2	15.3	26.4	40.9
	7	2	34.3	30.2	37.5	49.3	62.3	32.5	29.9	40.8	55.1	68.0
	8	−4	77.5	76.1	81.8	87.3	92.3	77.4	78.1	84.0	89.6	93.3
	9	−10	88.9	93.8	96.2	97.6	98.9	90.6	95.0	97.4	98.2	98.8
1760	1	52	.5	.2	.4	1.7	4.3	.2	.2	.6	.8	1.6
	2	42	.6	.4	.7	2.9	6.3	.5	.4	.9	1.8	3.9
	3	32	1.0	.7	1.5	5.6	11.0	.8	.8	2.1	3.8	8.3
	4	22	1.9	1.4	3.5	11.0	19.5	1.4	1.9	4.4	8.8	16.6
	5	14	5.0	5.7	9.5	20.3	33.4	3.8	4.9	9.4	19.5	31.9
	6	8	11.3	11.5	18.3	32.9	47.8	7.9	10.7	17.8	32.9	47.5
	7	2	31.5	31.8	40.3	55.6	69.1	26.7	30.7	43.0	58.7	70.4
	8	−4	63.5	63.6	74.6	82.8	90.2	60.1	66.4	77.2	85.1	91.2
	9	−10	85.1	89.4	93.3	95.6	97.9	85.9	91.0	94.5	97.3	97.8
3520	1	46	1.8	2.5	5.8	15.3	25.6	.2	.6	1.5	3.0	6.7
	2	36	2.3	3.9	8.8	20.4	33.8	.4	1.1	2.3	5.0	10.7
	3	26	3.8	6.3	13.6	29.9	45.8	1.0	1.9	4.7	9.6	19.8
	4	16	7.8	11.7	23.5	44.8	61.7	2.9	4.5	10.4	21.9	36.2
	5	8	13.3	20.2	36.1	60.1	75.5	6.6	10.2	21.2	37.7	54.7
	6	2	26.7	35.7	54.0	73.9	86.8	16.4	21.9	38.0	56.2	72.3
	7	−4	54.5	64.1	76.9	88.8	95.3	45.0	52.5	68.6	81.7	89.7
	8	−10	78.4	85.8	93.2	97.0	99.2	71.8	81.2	91.3	95.9	98.0
	9	−16	90.6	95.6	98.1	98.9	99.7	89.0	95.5	98.2	98.9	99.3
7040	1	33	6.3	7.8	16.1	35.0	53.1	1.6	2.7	8.4	20.1	38.3
	2	27	7.1	8.9	18.1	37.5	56.6	2.0	3.3	9.6	22.7	41.1
	3	21	8.7	10.7	20.8	41.8	60.8	3.0	4.4	11.9	27.2	45.9
	4	15	13.2	15.3	28.6	50.9	70.5	6.2	7.9	19.1	38.0	57.7
	5	9	20.6	23.8	38.6	61.7	78.4	13.8	15.7	31.1	50.7	70.2
	6	3	32.5	35.7	52.4	72.6	86.4	24.1	28.2	47.2	64.9	82.1
	7	−3	49.4	54.7	70.4	85.0	94.0	41.3	49.3	67.7	81.3	92.6
	8	−9	70.2	75.6	85.0	93.8	97.5	64.0	72.3	86.0	92.7	97.6
	9	−15	82.7	88.1	94.5	97.6	98.9	80.0	87.9	95.1	97.5	99.1
No. of Tests			4132	3287	3197	4528	1935	3417	4208	3978	4369	2538

* These tones were used for the instructions for the test.

such distributions for the 35,589 test scores that were used in calculating the mean hearing loss values of Table 2. It is arranged to show, for each of the five tones, the cumulative distributions separately for men and women in the five age ranges. It gives the percentage of tests failing to show a correct response at the test step indicated, or the percentage of individuals having a greater hearing loss than that corresponding to the indicated step. For example, the table shows that only 0.7 per cent of the men in the 20–29 age group have hearing losses greater than 32 db for a 1760-cycle tone, while 11 per cent of those in the 50–59 group have this much loss.

Zero hearing loss falls between steps 7 and 8 for the first three tones, and between steps 6 and 7 for the last two tones. The last step corresponds to very good hearing, and individuals able to hear this step have hearing acuities at least 10 db better than average. Some 10 or 15 per cent of the youngest age group, but only 1 or 2 per cent of the oldest group, were able to hear the tones on the last step. For the tone of lowest frequency, there were seven young persons for every older person who could hear the last step, but at 7040 cycles, there were 18 young persons for every such older person.

The tabular data for the age groups 20–29 and 50–59 are shown graphically in Fig. 4 for four of the tones, beginning at 880 cycles.[4] The curves are cumulative distributions and the ordinate gives the percentage of individuals having hearing losses greater than the value indicated by the abscissa.[5] At 880 cycles hearing losses in excess of a given amount tend to be more prevalent among women than among men. At 1760 cycles the distribution curves for men and women are much the same. At the two higher frequencies, the prevalence of deafness in excess of a given amount is greater among men than among women.

A hearing loss of 25 db at frequencies up to 1760 cycles begins to be a handicap. The individual will usually be aware of such an impairment, and will experience difficulty in understanding speech under conditions of public address, such as in the church or theater or around the conference or dinner table. The distribution curves show that only about 1.5 per cent of the young people taking the test, or three out of 200, have a hearing loss of 25 db or more for tones of these frequencies. In the oldest age range almost ten times as many, or every seventh person, shows this much impairment.

[4] The distributions for 880 cycles may be used for 440 cycles as well, it being assumed that the small difference shown in Table 4 is due to practice.

[5] The ordinates are shown on an arithmetic probability scale, which has the property that a normal distribution plots as a straight line whose slope is proportional to the standard deviation of the distribution. It is convenient because it shows the small values more accurately and because on this scale the standard errors of the ordinates are approximately equal in all parts of the range.

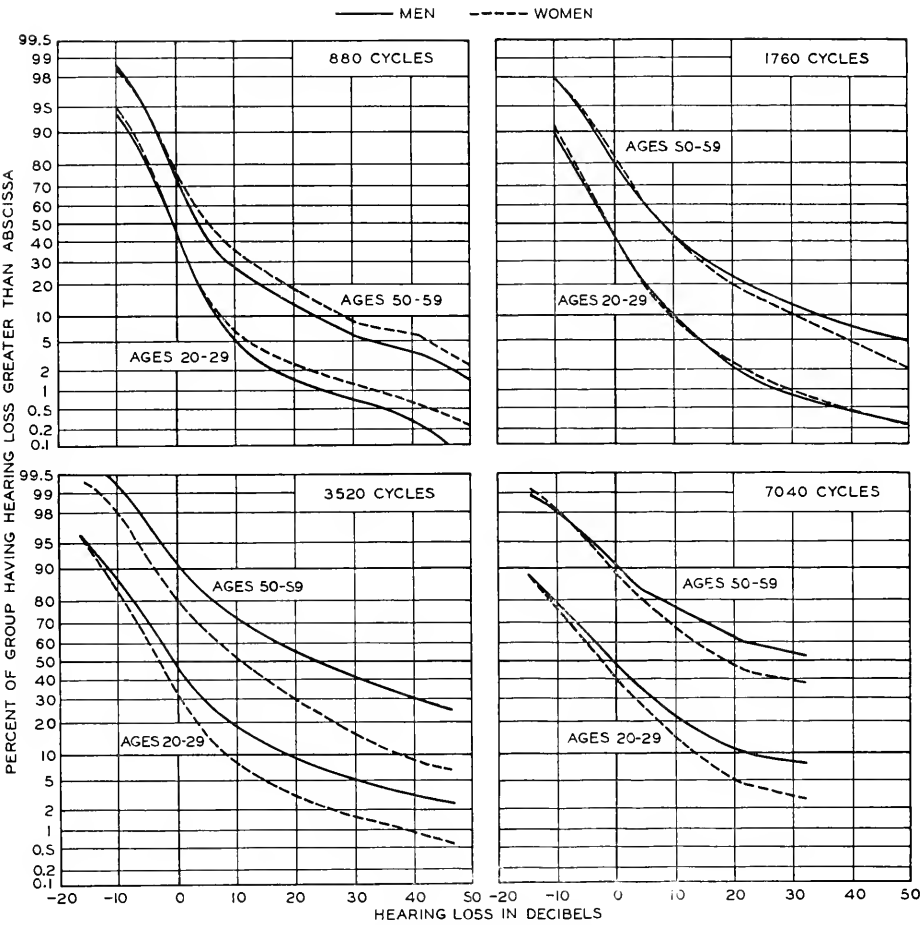


Fig. 4—Percentage of people in a given age and sex group having a hearing loss greater than any given value.

A hearing loss of 45 db for frequencies up to 1760 cycles will usually make it difficult to hear direct conversation even when the speaker is about two or three feet away. Individuals with this much loss usually need some sort of hearing aid. Table 5 gives the percentages of various groups that have losses in excess of 25 db and 45 db at the various frequencies.[4]

Acuity for the two high-frequency tones is less important than for the low tones for understanding speech, so that a loss for the high tones is not such a serious handicap. High-tone deafness is of particular interest, however, to the extent that it is indicative of a pro-

TABLE 5

PERCENTAGE OF TESTS WITH HEARING LOSS GREATER THAN 25 AND 45 DB.

Age Group		25 db Loss, Frequency				45 db Loss, Frequency		
		440, 880	1760	3520	7040	440, 880	1760	3520
10–19	Men	1.7	1.6	4.5	8.0	.6	.6	1.8
	Women	1.8	1.2	1.2	2.4	.6	.4	.3
20–29	Men	1.1	1.2	7.0	9.5	.1	.3	2.7
	Women	1.8	1.6	2.2	3.5	.4	.3	.7
30–39	Men	1.8	3.5	15.0	19.0	.3	.6	6.0
	Women	3.5	3.5	5.5	10.0	1.2	.8	1.6
40–49	Men	5.5	9.5	32.0	39.0	1.4	2.6	16.0
	Women	7.0	7.0	11.0	24.0	2.1	1.5	3.0
50–59	Men	9.5	17.0	48.0	58.0	2.6	6.0	27.0
	Women	13.0	14.0	22.0	43.0	4.0	3.0	7.0

gressive condition which may later involve tones of lower frequency. It is striking to note that, of the people taking the hearing test at the World's Fairs, some 6 per cent in the 20–29 year age range showed a hearing loss in excess of 25 db for the 7040 cycle tone. In the oldest age range half of the people showed such a loss. It is likely that an even larger proportion would be found in a random sample of the population, for, as will be discussed in a subsequent section, it is believed that the people taking the test at the Fair represent an economic status that is average or better, and there are indications of a greater prevalence of hearing defects in the lower economic groups.

The Estimation of Age

As previously indicated, estimates were relied upon to furnish information on the ages of the visitors taking the hearing test. The attendant making these estimates was changed every hour or so, about fifty being used in the course of a week at the New York Fair. In order to determine the accuracy of the estimates, 267 test blanks of members of Bell Telephone Laboratories families were examined and the real age compared with the estimated age indicated on the blank. In most cases the attendant who made the estimate was not aware that the individual was a member of the Bell System. It was found that 62 per cent of the real ages were within the estimated 10 year age group, while 83 per cent were not more than 3 years outside of the group and 96 per cent were not more than 8 years outside of the group. There was a general tendency to estimate high for young people and low for older people.

The average real age in any estimated age group depends on the distribution of ages within the group, the amount of random error involved in making the estimates, and any consistent tendency to estimate high or low. The distribution of estimated ages at the two Fairs is shown in Table 6, together with the age distribution of people

TABLE 6

PERCENTAGE OF INDIVIDUALS OVER 10 YEARS OF AGE FALLING IN VARIOUS AGE GROUPS ACCORDING TO ESTIMATES AT THE TWO FAIRS, AND FROM THE 1930 U. S. CENSUS

Age Group	New York	San Francisco	United States Population
10–19	34	27	24
20–29	25	23	21
30–39	23	23	19
40–49	14	16	15
Over 50	4	11	21
Total	100	100	100

over 10 years of age in the U. S. population (1930 census). At New York, the youngest group was considerably larger than the corresponding group in the population while the oldest group was very much smaller. The same tendency, but to a lesser degree, is seen in the San Francisco distribution. The distribution of ages fluctuated greatly on different days of the week and at different seasons. The table is based on counts on about 35 days at each Fair, well scattered through the season.

From the results obtained with the control group described above and the age distributions of Table 6, the average real age in each age group was judged to be as follows:

Age Group	10–19	20–29	30–39	40–49	50–59
Average real age	15	23	33	44	56

These values should be reliable to the extent that it is reasonable to assume that the accuracy of the age estimates of the control group was representative of the whole group at each Fair. The oldest group, which is designated 50–59 throughout this paper, included that range at New York, but at San Francisco included all persons over 50. Because of this the average real age in this group is judged to be 54 at New York and 60 at San Francisco. The value in the table is the weighted average of these two figures.

Relation of Hearing to Other Factors

In the preceding section it was shown that hearing acuity varies to quite an important extent with the age and sex of the group. The next problem is to identify any other factors to which hearing acuity is related to a significant degree. A sensitive method of determining whether such factors exist is provided by the control chart technique developed in connection with the statistical control of manufactured product.[6] When the hearing tests results, corrected for age and sex differences, were plotted on a control chart there was very definite evidence of lack of statistical control. This indicates that one or more factors exist to which hearing is significantly related, and experience in other fields in which control chart technique has been applied suggests an excellent chance of being able to identify these factors. The most straightforward procedure would have been to make a careful study of those individuals whose tests fell outside of the control chart limit and discover the factors responsible for the abnormal scores. This was not feasible at the Fairs, and other less direct methods were used as described below.

In the discussion which follows, a judgment must often be made as to whether an apparent relation between hearing and some factor under discussion is significant. The customary formula for the significance of a mean

$$\sigma_m = \frac{\sigma}{\sqrt{n}},$$

where σ_m is the standard deviation of the mean of a group of n observations and σ is the standard deviation of a single observation, has not been used, because modern statistical theory shows that this relation is valid only for data which are in a state of statistical control. The data of this section do not meet this requirement, and attempts to use the above relation as the sole test of significance are often misleading. The judgments which are expressed as to the significance of a relation are based upon the consistency with which the relation appears when the data are broken up into small groups. Space does not permit showing all the evidence on which these judgments are based, but the summaries which are presented indicate the magnitude of such relations as are judged to exist.

Place of Residence

Data from the New York and San Francisco Fairs were compared to discover any differences which might be attributed to sectional

[6] W. A. Shewhart, "Statistical Method from the Viewpoint of Quality Control" (Grad. School, U. S. Dept. of Agriculture, 1939).

TABLE 7

DIFFERENCE IN MEAN HEARING LOSS AT NEW YORK AND SAN FRANCISCO

		Frequency					No. of Tests	
		440	880	1760	3520	7040	N. Y.	S. F.
Men	10–19	.8	.0	.0	2.4	3.6	2839	1293
	20–29	.6	−.3	−.1	1.8	3.6	2219	1068
	30–39	.8	−.4	−.4	2.6	3.1	2193	1004
	40–49	−.5	−.8	.0	4.0	2.7	3171	1357
	Aver.	.4	−.4	−.1	2.7	3.2		
Women	10–19	.0	−.8	−1.0	−.2	.5	2172	1245
	20–29	.3	−.6	−.9	−.2	1.8	2848	1360
	30–39	.7	−.5	−.9	.2	2.3	2733	1245
	40–49	.6	−.2	−.1	.4	.8	3119	1250
	Aver.	.4	−.5	−.7	.0	1.4		

differences in hearing acuity. Table 7 gives the difference in mean hearing loss between corresponding age groups at New York and San Francisco for the age groups below 50. A positive difference indicates greater hearing loss at San Francisco.

At the three lower frequencies, the differences are insignificant. The differences at the higher frequencies are not large enough to be conclusive evidence of a sectional difference in hearing, but it seems quite probable that the men taking the test at San Francisco were, on the average, some three decibels less acute than those at New York for the frequencies 3520 and 7040. No important differences in standard deviation or general form of the distribution of hearing loss were noted in comparing the two Fairs.

For approximately a week at each Fair a question was printed on each hearing test blank asking whether the visitor lived (*a*) in the city where the Fair was held, (*b*) within commuting distance, or (*c*) beyond commuting distance. From the replies to these questions it was found that roughly one-quarter of the visitors at each Fair lived in the city, and another quarter within commuting distance. The remaining one-half were probably well scattered. The test results were analyzed in relation to place of residence. Most of the differences in mean hearing loss were so small that they could easily be attributed to sampling variations. Only one of the comparisons among the various groups showed differences sufficiently large and consistent to be significant. Women from New York City had greater hearing loss at all frequencies than women from the commuting area or beyond, as shown in Table 8. The groups compared number 600 and 1400 respectively. The differences did not show any trend with

age. The table gives the average difference between corresponding age groups in the age range from 10 to 49. A similar difference was not found among the men nor among the corresponding groups at San Francisco.

TABLE 8

Frequency	440	880	1760	3520	7040
Hearing Loss Difference	1.4	2.7	2.2	1.6	2.6

The differences just discussed are small enough so that the average hearing values computed from all the data will not be critically dependent on the weight assigned to the various geographical groups. On the other hand, some of the differences are large enough to suggest that a more efficient segregation into geographical groups, taking account of past as well as present place of residence, might uncover some substantial differences in hearing.

Personal Characteristics

An attempt was made to determine the relation to hearing acuity of several such factors as economic status, intelligence, and general appearance. This was done by observing individuals at the New York Fair as they submitted their test blanks for the photographic record. Some ten or fifteen seconds of observation were usually available, and the individual was classed as below, average, or above in respect to the characteristic being studied. Although separate estimates were attempted for each of the three characteristics named above, it was concluded that in each case the same thing was being estimated, namely general personal appearance. Accordingly all the data were combined. Table 9 summarizes the findings. Each figure is the mean of the mean hearing losses for the age groups below 50.

TABLE 9

VARIATION OF MEAN HEARING LOSS WITH PERSONAL APPEARANCE

	Frequency					No. of Tests
	440	880	1760	3520	7040	
Men—Below	4.3	2.4	1.8	4.4	5.7	95
Average	1.5	1.5	1.9	8.2	7.6	560
Above	−1.2	0.5	0.8	6.0	2.0	184
Women—Below	4.4	3.2	3.8	4.3	6.7	52
Average	2.2	2.7	2.5	2.4	3.6	658
Above	0.6	1.4	1.7	0.8	2.3	259

The differences shown suggest a relation between hearing acuity and general personal appearance, although the evidence is not conclusive. In each of the ten comparisons given in the table, there is an increase in hearing acuity in going from average to above average and in nine out of ten there is an increase in going from below average to above average. Personal appearance is somewhat related to economic status and intelligence. A more accurate index of these might show a more striking relation with hearing acuity.

Race

The number of tests of negroes tabulated thus far is too small to give a satisfactory picture of the hearing trends among them. However, there is no indication of substantial departure from the results reported by Bunch and Raiford,[7] who determined that the hearing of negro men and women is similar to that of white women.

Awareness of Hearing Impairment

People whose hearing is impaired are often quite sensitive, and it seems possible that some may have avoided the hearing test for this reason. On the other hand, a person with impaired hearing might be especially attracted by the opportunity to measure it. Whether the data show too high or too low an incidence of hearing impairment depends on which of these factors predominates. This is a possibility for bias that is present in any survey where participation is voluntary. No satisfactory method of evaluating it has been discovered. However, the following discussion is intended to give some idea of the magnitude of the error which may be involved.

Since a person is scarcely aware of a hearing loss of less than 25 db, it may be assumed that neither of these factors would affect the distributions below that value. For greater losses some effect may be expected, gradually increasing so that above 40 db the possibility of a substantial bias in the distributions must be considered. The shift of the mean values of hearing loss is probably not very pronounced. For example, Table 10 shows the shift in mean hearing loss at 1760

TABLE 10

Age	20–29	30–39	40–49	50–59
Men	0.4 db	0.7 db	2.6 db	4.8 db
Women	0.4	0.9	1.6	2.9

[7] C. C. Bunch and T. S. Raiford, "Race and Sex Variations in Auditory Acuity," *Arch. of Otolaryng.*, **13**: 423–434 (1931).

cycles which results when the number of cases of hearing loss over 42 db is tripled. Eliminating all cases over 42 db would produce about half as much shift in the opposite direction. However, the form of the distribution curves for large values of hearing loss may be substantially in error.

Hearing losses at 3520 and 7040 cycles are much more common, but are not likely to be noticed except when accompanied by a loss at a lower frequency. Consequently, the biasing effect of a selective process based on awareness of hearing loss is less pronounced at these frequencies.

Right and Left Ears

The physical arrangements at the Fairs made it awkward for a right-handed person to test his right ear. As a result, about 80 per cent of the recorded tests were for the left ear. No appreciable difference was found between the test results for right and left ears. However, this may be taken as only a rough indication of equality between the ears, in view of the differences in test conditions for the two ears.

Time of Day

Data from each Fair were studied to determine whether there was any significant variation in hearing with time of day. Table 11, which gives values of mean hearing loss covering a period of about two weeks at the San Francisco Fair is typical.

TABLE 11

VARIATION OF MEAN HEARING LOSS WITH TIME OF DAY

	Frequency					No. of Tests
	440	880	1760	3520	7040	
Morning (9–1:30)	1.0	1.2	1.9	3.5	3.5	1285
Afternoon (1:30–5:30)	2.1	1.4	2.3	3.3	4.6	1637
Evening (5:30–10)	1.4	1.6	2.3	3.4	4.4	1511

The figures given are averages of several age groups. An apparent slight trend to poorer hearing in the afternoon is probably not significant, because detailed study of this and other data showed that this trend was not consistent. It was concluded that there were no trends of hearing acuity with time of day in any age group of sufficient magnitude to be revealed by the survey.

Variation Over Longer Periods

A comparison of data obtained at the New York Fair during a period early in the summer and another period early in the fall is given in Table 12. This table gives the difference in mean hearing loss for the two periods for the various age and sex groups. A positive difference indicates greater hearing loss in the later period.

TABLE 12

DIFFERENCE BETWEEN MEAN HEARING LOSSES DETERMINED DURING TWO DIFFERENT PERIODS

Ages		Frequency					No. of Tests
		440	880	1760	3520	7040	
Men	10–19	.4	.2	.0	.7	.6	1509, 1330
	20–29	−.7	−.8	−.6	.9	−.2	1110, 1109
	30–39	−1.0	−.6	.5	1.1	1.4	1053, 1140
	40–49	.2	.2	.2	1.4	1.1	1802, 1369
	50–59	1.8	1.9	2.6	5.0	3.1	432, 511
Women	10–19	−1.4	.0	−.4	−.7	−.4	1028, 1144
	20–29	−1.0	.1	.5	.9	.6	1645, 1203
	30–39	−.7	.2	.3	.1	1.0	1363, 1370
	40–49	−.5	−.3	−.3	.6	1.0	1722, 1397
	50–59	.3	1.0	1.3	1.8	1.1	440, 643
Average	10–49	−.6	−.1	.0	.6	.6	

The average difference is rather small, and may be taken to indicate good stability on the part of the test equipment and lack of any pronounced seasonal trends in hearing over this interval.

DISTRIBUTION OF HEARING ACUITY IN THE UNITED STATES POPULATION

The tests at the two Fairs constitute a large cross section of the United States population. It is not a representative cross section in certain respects, the most important of which are described below. It is believed that by taking into consideration the limitations mentioned below an estimate of the distribution of hearing acuity in the United States population can be obtained which is sufficiently accurate for most practical purposes.

The two Fairs taken together probably represent a good geographical cross section of the country, except that areas near the Fairs were too heavily represented. However, since no pronounced differences in hearing were found for those living near the Fairs, the geographical sampling may be regarded as fairly satisfactory. Although it is quite

possible that sections might be found in which hearing differed markedly from the Fair values, it is unlikely that such areas would be extensive enough to affect the overall result.

With regard to economic status, intelligence, and amount of education, the Fair groups were judged to be somewhat above average, and probably representative of the upper two-thirds or three-quarters of the population. If hearing is related to such factors, as seems probable, the hearing of the population is not quite so good as indicated by the Fair tests.

The portion of the distribution curves relating to large hearing losses at the three lower frequencies must be accepted with reservations on account of the possible biasing effect of awareness of hearing impairment, as discussed in the preceding section. However, the curves should be reliable below about 35 db loss, and the curves for the two highest frequencies should be reliable throughout the test range.

The distribution of ages at the Fairs was quite different from that in the population. The first step in allowing for that difference was made by recombining the distributions of hearing loss for various age and sex groups shown in Table 4, weighting each according to the size of the group in the population. The resulting distributions are shown in Fig. 5, and apply to the age range 10–59 years.[4]

In a similar manner, the figures in Table 5 for the incidence of hearing loss of 25 db or more were weighted according to the size of the groups and combined, leading to the values given in the first line of Table 13, for the age range 10–59. This process can be extended to include the whole age range as follows. It is assumed that the incidence for ages under 10 is the same as in the 10–19 group. This may not be strictly true, but is a sufficiently good approximation for this purpose. For ages above 60 a minimum estimate was obtained by assuming that the incidence of hearing loss is the same as in the 50–59 group, and a maximum estimate by assuming 100 per cent incidence above age 60. The actual value may be expected to be somewhere between these limits, which are shown in Table 13.

Except for the reservations stated in the first four paragraphs of this section, it is believed that the data of Figs. 3, 4, and 5 and Tables 2, 3, 4, 5, and 13 should apply fairly closely to the U. S. population as a whole. They may also be applied to groups in the population who are not specialized in regard to any factor related to hearing. It would be unsafe to apply them to a group of very low or unusually high economic status, college graduates, unskilled laborers, foreign groups,

[4] Loc. cit.

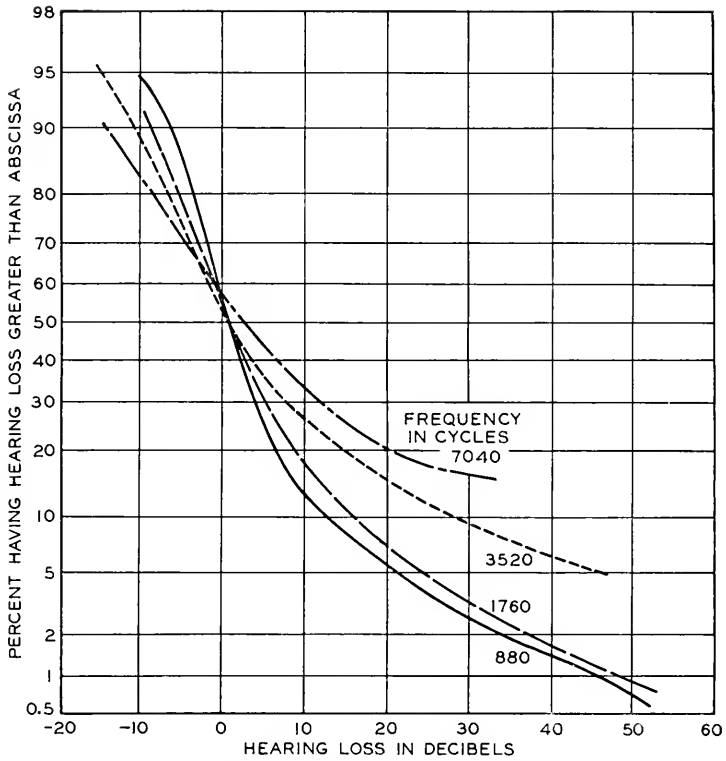


Fig. 5—Percentage of people, both men and women, in the age range 10–59 having a hearing loss greater than any given value.

TABLE 13

PERCENTAGE OF PEOPLE HAVING HEARING LOSS OF 25 DB OR MORE

		Frequency			
		440, 880	1760	3520	7040
Ages 10–59		3.8	4.5	12	18
All ages	Minimum	4	5	12	18
	Maximum	11	12	18	22

etc. without further knowledge of the relation between hearing and the factor in which the group was unusual.

ACCURACY OF THE HEARING TEST

The participation of the visitors in the test was entirely voluntary, and nothing is known from the test records concerning their reasons

for taking the test. Observation indicated that the great majority of people took the test seriously, and made a conscientious attempt to test their hearing.

It was noticed that a very small percentage of the people, mostly in the youngest age group, altered their scores by filling in all of the missing numbers before having them photographed, thus giving a false appearance of a perfect test. Some of these were detected from the differences in writing, and were eliminated from the tabulations, but others were probably included. It is believed that the number of false scores included was too small to affect the hearing loss distributions appreciably.

Some people undoubtedly secured poor scores in the test on account of failure to understand the test, interruptions, or other causes not connected with hearing. A study of this factor was made by observing about 1200 tests, picked at random, and interviewing all those who failed to fill in more than three squares of their test blanks. About 1.5 per cent failed the test in this sense, and were subsequently interviewed and watched to see if they permitted their test scores to be photographed. The interview revealed that about two-thirds of this group failed the test because they were definitely hard of hearing. Also, as it happened, about two-thirds of them submitted their test scores to be photographed, so that the number of recorded failures tended to be the same as the number that were actually hard of hearing.

The noise conditions under which the tests were given were quite favorable. The exhibit building was quiet due to the generous use of sound absorbent walls and carpeted floors. Parts of the building containing air-conditioning and other machinery were constructed on a separate foundation from the part containing the hearing test booths. The booths were carefully insulated, the attenuation of the walls to air borne sounds being 30 db or more over a wide frequency range. Additional isolation was provided by a glass partition between the booths and the lobby. Noise in the booths from external sources was nearly inaudible, and it is probable that most of the distributing noise was caused by the people participating in the test.

Sound level meter measurements with flat weighting were made in a booth where noise from external sources was judged to be most objectionable, and while regular tests were in progress. The average and maximum readings are given in Table 14. After making allowance for the attenuation of the telephone receiver covering the ear, the masking computed for the average noise level was less than 5 db at 440 cycles and zero at the higher frequencies. The masking of a

TABLE 14

SOUND LEVEL METER READINGS, WITH FLAT WEIGHTING, IN DB ABOVE .0002 DYNE PER SQUARE CENTIMETER

Frequency Band	100–300	300–500	700–900
Average reading	43	<25	≪25
Maximum reading	50	35	26

steady noise equal in magnitude to the maximum noise levels was computed to be 11 db at 440 cycles, 3 db at 880 cycles, and zero at higher frequencies.[8] The interpretation of these results is in some doubt because the people in the booths tended to be more quiet when the test level approached threshold, and also because the disturbing effect of sounds of irregular character may not be properly indicated by masking computations based on experiments with steady sounds. Accordingly a more direct method of evaluating the disturbing effect of noise was tried.

Members of Bell Telephone Laboratories who had taken the test at the Fair under routine conditions at various times during the season were retested at the Laboratories after the Fair closed. The same equipment and procedure were used, except that only one person was tested at a time under conditions free from any disturbing noise except that created by the observer himself. On the average, the tests indicated more acute hearing during the retest at the Laboratories, particularly at the low frequencies. The average shift was 2.9 db at 440 cycles, 1.4 db at 880 cycles, 1.1 db at 1760 cycles, and negligible at the higher frequencies. Since the test at the Laboratories was given last in every case these shifts may have been partly due to improvement with practice. However, they serve to set an upper limit to the average disturbing effect of noise. This comparison is based on tests of 150 ears of 106 people, whose average age was 39 and whose average hearing acuity was somewhat better than an equivalent age group at the Fair.

The equipment for the tones hearing tests consisted of eight machines at New York and two at San Francisco. These machines were maintained in an equipment room some distance from the test booths, one machine being connected to each booth. Each machine consisted of a phonograph reproducer, amplifier, attenuation network, and seven

[8] These values of masking apply to an ideal observer having a threshold approximating the minimum audible pressure curve of Fig. 6. Since a great majority of observers at the Fairs had higher thresholds, the masking would be correspondingly less for them.

telephone receivers in parallel on the output.[9] Vertical cut phonograph records contained instructions for the test and the tones used in the test. To insure a favorable ratio of signal to record and amplifier noise throughout the test, the test tones were recorded at constant level, and the desired level changes were obtained by changes in the attenuation network made in synchronism with the turntable.

Output of each phonograph reproducer and amplifier was checked daily, and held within limits which varied from ± 0.5 db at the lowest frequency to ± 2 db at the highest frequency. Performance of the attenuation networks was determined twice during the season by careful measurements of voltage at each test level and each frequency. They were found to give the expected values of attenuation within 1 db at all levels. The efficiency of each receiver was measured on a rigid closed coupler. The standard deviation of all the receivers used was about 1 db at the lower frequencies and 3 db at 7040 cycles. Check measurements were made at intervals of about one month on each receiver. The mean response of all the receivers varied by less than 1 db during the season. The ten machines were alike in output (at the test level nearest the reference level) within ± 1.0 db at the three lower frequencies and within ± 1.5 db at the two higher frequencies.

In addition to the above measurements, listening tests were made daily by one of the engineers in charge of the equipment, and the girls who conducted the test listened frequently throughout the day by means of monitoring receivers.

Hearing Test Results in Terms of Pressure and Intensity Level

In order to compare the results of the Fair tests with other data on hearing, calibrations have been made of the receivers that were used in the tests. This was done by measuring the pressure levels developed by the receiver at the opening of the ear canal for a small group of people, using a special search tube transmitter so designed that the tube could be inserted under the receiver cap into the opening of the ear canal. Such a calibration gives an ear canal pressure level in terms of receiver voltage levels. The authors are indebted to Mr. W. A. Munson of these Laboratories for the calibrations. They are preliminary in character and may need modification in the light of subsequent studies.

With the aid of these calibrations, ear canal pressure levels may be

[9] F. A. Coles, "Hearing-Test Machines at the World's Fairs," *Bell Laboratories Record*, 18: 290, June 1940.

calculated from the receiver voltage levels measured in the tests. Such calculations for the reference level or condition of zero hearing loss as used in this paper are shown in Table 15. The resulting reference ear canal pressure levels are plotted in Fig. 6. For comparison,

TABLE 15

CALIBRATION OF HEARING TEST EQUIPMENT AT THE REFERENCE LEVEL

	Frequency				
	440	880	1760	3520	7040
Reference voltage level across receivers—db above one volt	−104	−112	−115	−114	−76
705A receiver calibration—db above 0.0002 dyne per sq. cm. per volt	133	134	133	134	98
Reference ear canal pressure level—db above 0.0002 dyne per sq. cm.	29	22	18	20	22

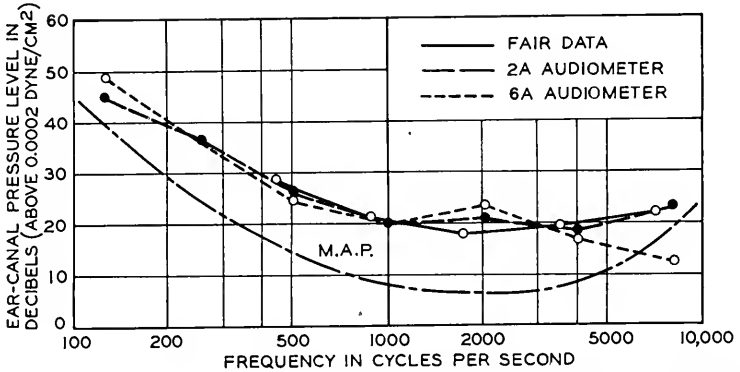


Fig. 6—Ear canal pressure level for certain reference conditions. The measurements for the M. A. P. curve were made nearer the ear drum than those for the other curves. See text.

the ear canal pressure levels corresponding to zero hearing loss on the 2A and 6A audiometers [10] and the minimum audible pressure curve derived by Sivian and White [11] are shown. The audiometer

[10] J. C. Steinberg and M. B. Gardner, "Auditory Significance of Hearing Loss," *Jour. Acous. Soc. Amer.*, **11**: 270 (1940).

In using the Audiometer it is customary to record as the hearing loss the lowest dial setting at which the tone is heard. Threshold would, on the average, be half a dial step lower than the recorded setting. Hence the curves given here for zero hearing loss are 2.5 db lower than those given for zero dial setting in the reference. See also footnote 2.

[11] L. J. Sivian and S. D. White, "Minimum Audible Sound Fields," *Jour. Acous. Soc. Amer.*, **4**: 288–321 (1933).

curves and the Fair curve are based on ear canal pressures measured at the ear opening in the manner just described. The minimum audible pressure curve is based on pressures measured about 1 cm. from the ear drum, which correspond more nearly to ear drum pressure levels. The two types of measurements are undoubtedly quite comparable below 1000 cycles. For frequencies above 5000 cycles and possibly around 2000 cycles it is believed that the pressure at the ear opening is somewhat smaller than the corresponding ear drum pressure.

A comparison of the Fair data with data from two other surveys of hearing is shown in Fig. 7. One curve shows the mean threshold

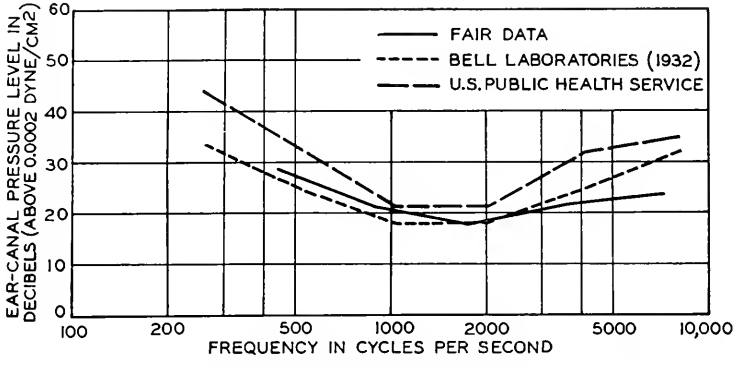


Fig. 7—Comparison of several surveys of hearing, giving mean ear canal pressure level for men aged 20–29.

pressure for men in the 20–29 age group from the Fair data. Another gives values for the same age and sex group in a survey conducted in 1936 by the United States Public Health Service using the 2A Audiometer.[12] This curve is for a somewhat selected group, including only individuals who stated when the test was made that they believed their hearing was normal. The third curve is for members of Bell Telephone Laboratories in the same age and sex group, who were tested in 1931 with a 2A Audiometer.[13] In comparing these results, it should be remembered that differences may be due to three general causes. The groups of people tested may have differed in hearing acuity. The calibrations by which the ear canal pressures were established are subject to error, especially at high frequencies. The conditions of the test, including technique, concentration of subjects, receiver fit, and background noise, were not alike in all cases. Con-

[12] W. C. Beasley, *National Health Survey, Hearing Study Series, Bulletin 5, Table 3*, The United States Public Health Service, Wash. D. C. (1938).

[13] H. C. Montgomery, "Do Our Ears Grow Old," *Bell Laboratories Record*, 10: 311 (1932). Note that median values were given in this reference, differing slightly from the mean values used here.

sidering all the possible causes of variation the curves seem to be in fairly good agreement.

In order to give a preliminary picture of the prevalence of deafness in terms of free field intensity and frequency, the distribution curves of Fig. 5 were converted into the contour lines shown in Fig. 8. For

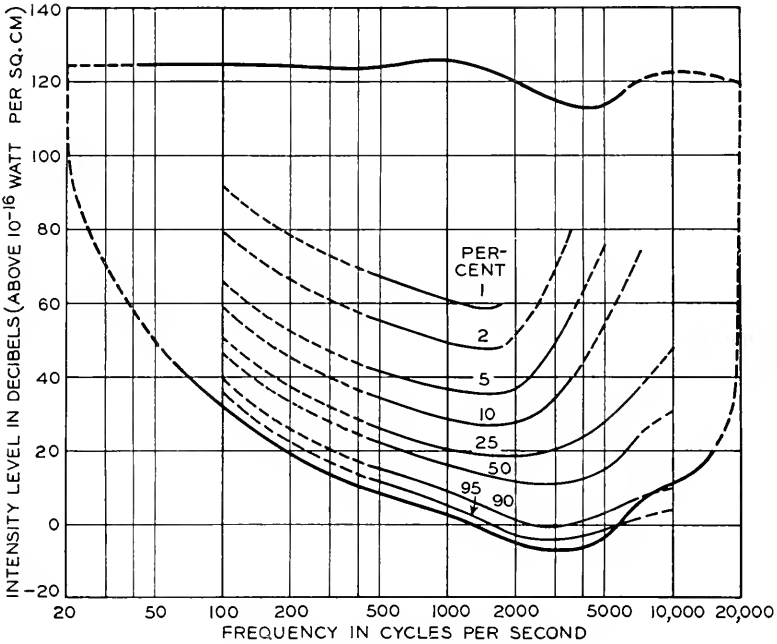


Fig. 8—Contour lines above which lie a given percentage of thresholds for the age group 10–59.

the low frequencies, this conversion was made by applying the differences between the minimum audible pressure and minimum audible field curves of Sivian and White [11] to the ear canal pressure levels of Table 15. For the high frequencies, the conversion was based on a free field calibration of the receivers. The free field intensity levels apply for the condition in which the observer faces the source and listens with both ears. The resulting contours purport to show the percentage of people in the population within the age range 10 to 59 years who cannot hear tones below the given level. The boundary lines forming the auditory sensation area represent the picture of the limits of useful hearing, based upon earlier studies.[14]

The solid portions of the contour lines represent the distributions

[14] H. Fletcher, "Auditory Patterns," *Reviews of Modern Physics*, 12: 47–65 (1940).

obtained from the results of the Fair tests. The dotted portions represent extrapolations of the distributions beyond the intensity and frequency ranges used in the tests, and are of course speculative in character. The extrapolations to lower frequencies are based on the shape of the contours below 1760 cycles and the high correlation that has been found to exist in individual audiograms for frequencies from 64 to 512 cycles.[15,16] The extrapolations to large hearing losses for the 3520- and 7040-cycle tones were made by extending the curves of Fig. 5 as suggested by comparison with the results of other surveys.

The contours show several interesting things. The range over which hearing acuity varies is quite uniform up to about 2000 cycles, and 90 per cent of the group lie within a range of 30 db. Above 2000 cycles the range increases rapidly. Since most of the sounds met with in daily life have intensity levels greater than the 25 per cent contour, fully three-fourths of the people can hear ordinary sounds throughout the frequency range from 100 to 10,000 cycles.

The Onset of Deafness

With the increasing attention being given to the prevention of deafness by early detection, it is of considerable practical importance to define the beginning, or onset, of deafness. Perfectly normal ears are not exactly alike; some are more acute and others are less acute than the average. How much less acute than average may an ear become before deafness begins? In a preceding section, hearing loss was evaluated on the basis of the handicap that it would impose. In this section we take a different viewpoint and use the term "beginning of deafness" to mean a departure from average hearing acuity sufficient to justify the expectation of associating the departure with a specific cause.

In Fig. 9 there is shown a typical distribution curve for hearing loss. It shows the relative frequency of occurrence of various degrees of hearing acuity among a given class of people. Such curves can be well described for many practical purposes by two quantities, the average hearing loss and the standard deviation. The latter quantity, designated as σ, is a measure of the spread of the individual values from the average.

The experience of statisticians with distributions of observations of widely different character indicates that there is little chance of assigning specific causes for the deviations of observations which lie closer

[15] E. G. Witting and W. Hughson, "Inherent Accuracy of Repeated Clinical Audiograms," *Laryngoscope*, **50**: 259 (1940).

[16] W. C. Beasley, "Correlation Between Hearing Loss Measurements," *Jour. Acous. Soc. Amer.*, **12**: 104–113 (1940).

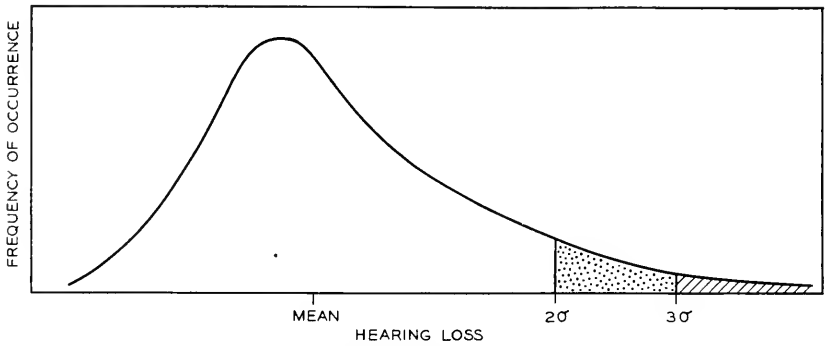


Fig. 9—Typical distribution curve of hearing loss.

than 2σ to the average, because, in this range, many causes are operating and none is predominant. In the range from 2σ to 3σ, the dotted area of Fig. 9, there is a good chance of isolating causes, although the attempt might not be justified if it involved great discomfort, danger or expense. Beyond 3σ, the cross-hatched area of Fig. 9, there is an excellent chance that the cause, or causes, can be isolated.[17] This does not imply that findable causes of hearing impairment do not exist in individuals lying closer to the average than these limits, but merely that the hearing test is not useful in selecting them. The identification of such causes and their treatment are, of course, medical problems.

TABLE 16

HEARING LOSS AT WHICH DEAFNESS BEGINS, IN THE SENSE GIVEN IN THE TEXT

Limit		Frequency			
		440, 880	1760	3520	7040
2σ	Boys	14	15	19	22
	Girls	14	13	11	15
3σ	Boys	21	22	29	34
	Girls	21	20	18	24

This criterion leads to the limits shown in Table 16, beyond which hearing loss is significant in the sense described. Limits are given

[17] It is important to note that the standard deviation to be used in fixing the limits is to be determined after eliminating the effects of impaired hearing. Methods by which this can be done have been developed in connection with statistical methods of manufacturing control. See, for example, W. A. Shewhart, "Economic Control of Quality of Manufactured Product" (Van Nostrand, New York, 1931). The values of σ used in fixing limits in this section are 20 per cent smaller than those given in Table 3.

only for the youngest group because we are primarily concerned with early detection of deafness. Limits for older groups can be obtained in a similar manner.

The table indicates that a smaller hearing loss is significant at low than at high frequencies. At the higher frequencies a smaller loss is significant for girls than for boys. The percentage of the youngest group at the Fairs falling outside of the 2σ limits is given in Table 17.

TABLE 17

PERCENTAGE OF CHILDREN WITH A SIGNIFICANT AMOUNT OF HEARING LOSS

	Frequency			
	440, 880	1760	3520	7040
Boys	4	5	7	8
Girls	5	5	5	6

Use of the 3σ limits would lead to percentages about half as great. It is of interest to note that the 2σ limits are smaller than the amounts of hearing loss which ordinarily constitute a handicap. In other words, a hearing acuity test may have diagnostic significance before the hearing loss is great enough to produce appreciable functional impairment.

Incidence of High-Tone Deafness and of Adenoid Growth

The characteristic difference between men and women in the hearing of high tones seems now to be well established. It has been previously observed by Bunch and Raiford,[7] by Ciocco [18] and more recently by Beasley in a survey of hearing during the health census of 1936.[19] The reason for the difference is not known. It may be occupational in part, although the difference is in evidence in the youngest age group. Recent work reported from Johns Hopkins University,[20] indicates that children with high-tone deafness frequently show pronounced lymphoid tissue growth at the openings of the Eustachian tubes in the throat. In a number of cases, the deafness was cured or held in check by removing the tissue and inhibiting its growth by irradiation with radium during the adolescent period.

If there is a connection between adenoid growth and high-tone deafness, one would expect a greater incidence of adenoid growth in

[18] A. Ciocco, "Observations on the Hearing of 1980 Individuals," *Laryngoscope*, **42**: 837–856, Nov. 1932.

[19] W. C. Beasley, *National Health Survey, Hearing Study Series, Bulletins 5 and 6*, United States Public Health Service, Wash. D. C. (1938).

[20] S. J. Crowe and J. W. Baylor, "Prevention of Deafness," *Jour. Amer. Med. Assn.*, **112**: 585–590, Feb. 1939.

boys than in girls. The results of physical examinations of school children conducted by public health officers and doctors in different parts of the country indicate that such is the case. In all but one of seven surveys,[21] involving more than 18,000 children, the occurrence of adenoid growth was more frequent in boys than in girls. On the average, 6 per cent of boys and girls from 4 to 18 years of age showed pronounced adenoid growth. The ratio of the percentage of girls to the percentage of boys having the defect had an average value of 0.68; for every 100 boys affected there were only 68 girls similarly affected. Analysis of the World's Fair hearing test records shows that 7 per cent of boys and girls from 10 to 19 years of age are deafened for a 7040-cycle tone to the extent of the 2σ limit described in the previous section. The ratio of the percentage of deafened girls to the percentage of deafened boys is 0.75, or 75 girls for every 100 boys. Although it should not be concluded that the similarity of these ratios establishes a correlation between adenoid growth and high-tone deafness, it is believed that they are sufficiently suggestive to justify further study of these defects.

Acknowledgment

This survey was made possible by the cooperation of a large number of people in many parts of the Bell System. The planning, design, and construction of the exhibit were shared by the American Telephone and Telegraph Company, The Western Electric Company, Electrical Research Products, Inc., and Bell Telephone Laboratories, Inc., and to them the authors are indebted. We wish to express our gratitude to the Pacific Telephone and Telegraph Company and the New York Telephone Company for their efficient operation of the exhibits and for the large share which they had in obtaining the data used in the survey; to the tabulating and mathematical groups at the Laboratories for many hours of painstaking labor in treating the data; and to many of our associates whose suggestions and criticisms were a valuable aid in the analysis of the information. We also wish to express our appreciation to the large group of interested visitors to the Fairs whose participation in the hearing tests constituted the basic material of this survey.

[21] "The Health of the School Child," *Public Health Bulletin 200, Table 46, page 141*, United States Public Health Service, Wash., D. C. W. Franklin Chappel, "Exan ination of the Throat and Nose of 2000 Children," *Jour. of Med. Sciences*, **97**: 148–154 (1889). Wm. R. P. Emerson, "Physical Defects in 1000 Children," *Amer. Jour. of Diseases of Children*, **33**: 771–778 (1927).

The Subjective Sharpness of Simulated Television Images

By MILLARD W. BALDWIN, JR.

1. Introduction and Summary

OF the many factors which influence the quality of a television image, the one which is generally indicative of the value of the image and the cost of its transmission is the resolution, or sharpness. This resolution factor has always been reckoned in purely objective terms, such as the number of scanning lines, or the number of elemental areas in the image, or the width of the frequency band required for electrical transmission at a given rate. The subjective value of sharpness has not previously been considered. Some recent tests with a small group of observers, using out-of-focus motion pictures in a basic study of the visual requirements on images of limited resolution, have thrown new light on the evaluation of resolution and sharpness. The results appear of sufficient interest, particularly when interpreted in terms of television images, to warrant this presentation. We shall use the word *sharpness* in the sense of a subjective or psychological variable, with a strict technical significance in keeping with our experimental method, and we shall use the word *resolution* in the sense of an objective or physical variable.

We find that as images become sharper, their sharpness increases more and more slowly with respect to the objective factors. We find also that as images become sharper the need for equal resolution in all directions becomes less and less, and that with images of present television grade the tolerance for unequal horizontal and vertical resolutions is already remarkably wide. These conclusions are supported by our experiments with small-sized motion pictures viewed at a distance of 30 inches, about 4 times the picture height. It would not be safe to extrapolate the results of these experiments to the large-screen conditions of motion-picture theaters, because the visual acuity of the eye may be expected to increase with distance in the range in question,[1] and for other reasons.

2. Exposition of Method

Image sharpness is to be measured by subjective test, employing psychometric methods [2] which have been widely used in the measurement of other subjective values. Test images are to be projected onto a screen from 35 mm. motion picture film in such a way that the reso-

lution of the image can readily be varied over a substantial range, and with provision for making the horizontal resolution different from the vertical. The use of motion pictures instead of actual television images permits sharpness to be studied independently of other factors, and facilitates the experimental procedure.

The relationship between the television image and the motion picture which simulates it will be determined on the basis of their subjective equality in sharpness. For that purpose, a television image reproduced by an apparatus * of known characteristics is to be compared with a projected out-of-focus motion picture of the same scene, under the same conditions of size, viewing distance, brightness and color. (The motion picture will in general be superior in the rendition of tone values and in respect to flicker, and will of course not show the scanning line structure of the television image or any of the degradations commonly encountered in electrical transmission.) When the two images are judged to be equally sharp by the median one of a group of observers, the size of the figure of confusion of the motion picture is to be taken as the measure of the resolution of the compared television image.

The figure of confusion of the motion picture is that small area of the projected image over which the light from any point in the film is spread. Every point produces its own figure of confusion, of proportionate brightness, and the overlapping of the figures in every direction accounts for the loss of sharpness. When the projection lens is "in focus," the figure of confusion is a minimum one set by the aberrations of the optical system and by diffraction effects. As the lens is moved away from the "in focus" position, the figure of confusion becomes larger and assumes the shape of the aperture stop of the projection lens. If the illumination of the aperture stop is uniform, this larger figure of confusion is a well-defined area of uniform brightness. We used a rectangular aperture stop, at the projection lens, whose height and width could be varied reciprocally so as to maintain constant area of opening, and we used a calibrated microscope to measure the departure of the lens from the "in focus" position. Thus we could produce images of various degrees of sharpness and of unequal horizontal and vertical resolutions.

This method of specifying the resolution of an image in terms of the size of the figure of confusion affords an important advantage. It avoids the necessity for postulating any particular relation between the resolution and the spatial distribution of brightness values about

* The television apparatus comprised a mechanical film scanner and an electronic reproducing tube designed specifically for television. A description of it is given in reference 9.

originally abrupt edges in the image. The variety of such relations assumed by others[3, 4, 5, 6, 7] has led to a variety of conclusions with respect to resolution in television. We find subjective comparison of images to yield results of fairly small dispersion.

Let us consider now the measurement of sharpness in subjective terms. Here we find no familiar units of measurement, no scales or meters. We find no agreement as to the meaning of a statement that one image looks twice as sharp as another. We can say of two images only that (*a*) one image looks sharper than the other, or (*b*) the two images look equally sharp. When the images are quite different, there will be agreement by a number of observers that the one image is the sharper. When the images are not different in sharpness, there may be some judgments that one of them is the sharper, but these will be counterbalanced in the long run by an equal number of judgments that the other is the sharper. When the images are only slightly different in sharpness, an observer may reverse his judgment from time to time on repeated trials, and he may sometimes disagree with the judgment of another observer. It is within this region of small sharpness differences, in the interval of uncertain judgments where the observer is sometimes right and sometimes wrong with respect to the known objective difference, that it becomes possible to set up, on a statistical basis, a significant quantitative measure of sharpness difference.

Suppose that in judging two images of almost equal sharpness the observers have been instructed to designate either one or the other of the images as the sharper; that is, a judgment of "equally sharp" is not to be permitted for the present. An observer who discerns no difference in sharpness is thus compelled to guess which image is sharper, and his guess is as likely to be right as it is to be wrong, with respect to the known objective difference. Suppose, further, that the sharpness difference has been made so small that only 75 per cent of the judgments turn out to be right, the remaining 25 per cent being wrong. On the basis that these wrong judgments are guesses, we must pair them off with an equal number of the right judgments, so that 50 per cent of the total are classed as guesses. The other 50 per cent are classed as real discriminations. (The pairing of an equal number of the right judgments with the wrong judgments goes back to the equal likelihood of right and wrong guesses; it affords the best estimate we can make of the number of guesses.) When real discrimination is thus evidenced in one half of the observations, that is, when 75 per cent of the judgments are right and 25 per cent of them are wrong, we shall designate the difference in resolution as the *difference limen*.*

* The term *limen* is frequently used in psychometry in lieu of older terms such as *just-noticeable-difference*, *threshold value*, *perceptible difference*, etc. It has the virtue

It is seen that the value of the limen is arrived at statistically, taking into account the variability of individual judgments. Smaller differences than the limen are not always imperceptible, nor are larger differences always perceptible.

The difference in sharpness, or in sensory response, which corresponds to a difference of one limen in resolution may be said to be one unit on the subjective scale of measurement. We shall designate this as a *liminal unit.** It will be understood that the word *liminal* has here a particular and precise significance, by reason of the one-to-one correspondence between the liminal unit and the statistically-derived value of the difference limen. A liminal unit of sharpness difference may be considered as the median of a number of values of sensory response to a difference of one limen in resolution.

3. Sharpness and Resolution

Figure 1 shows how we find the sharpness of an image to vary as the number of elemental areas in the image is changed. Sharpness is expressed in liminal units, based on measurements of the limen at four different values of resolution, indicated by the four pairs of points on the curve. Resolution is expressed as the number of figures of confusion in a rectangular field of view whose width is 4/3 of its height and which is viewed at a distance of 4 times its height.† This conventional field of view was chosen as typical of viewing conditions for motion pictures and television images. (The conventional field is 19° wide by 14° high.) The range of the curve in Fig. 1 may be stated very roughly as from 150-line to 600-line television images.

The significant feature of this curve is its rapidly decreasing slope with increasing sharpness. It shows that sharpness is by no means proportional to the number of elemental areas in the image, and demonstrates that the use of objective factors as indices of sharpness should be regarded with more than the usual amount of caution. It shows

that its meaning may be precisely defined in terms of the particular experimental method under consideration, without the extraneous significance which might attach to the more commonplace words.

* There seems to be no accepted name for such a unit. Guilford[2] calls it simply "a unit of measurement on the psychological scale." In discussing the measurement of sensory differences which are equal to each other but not necessarily of liminal size, the terms "sensory value" and "scale value" have been used.

† We have used relative values here in order that our results might be applied to other images not too different in size from the small ones we actually used. Other values of aspect ratio in the neighborhood of 4 to 3, and other values of viewing distance in the neighborhood of 4 times the picture height, may be brought within the scope of our data on the assumption that the sharpness is the same if the solid angle subtended by the area of the figure of confusion is the same. For example, a square field of view containing 60 thousand figures of confusion, and viewed at 5 times its height, would be equal in sharpness to our conventional field containing 125 thousand figures of confusion [$125 = 60 \times 4/3 \times (5/4)^2$].

that images of present television grade are well within a region of diminishing return with respect to resolution, a region, however, whose ultimate boundary is still well removed. (We estimate that the sharpest image our motion picture machine could project would be repre-

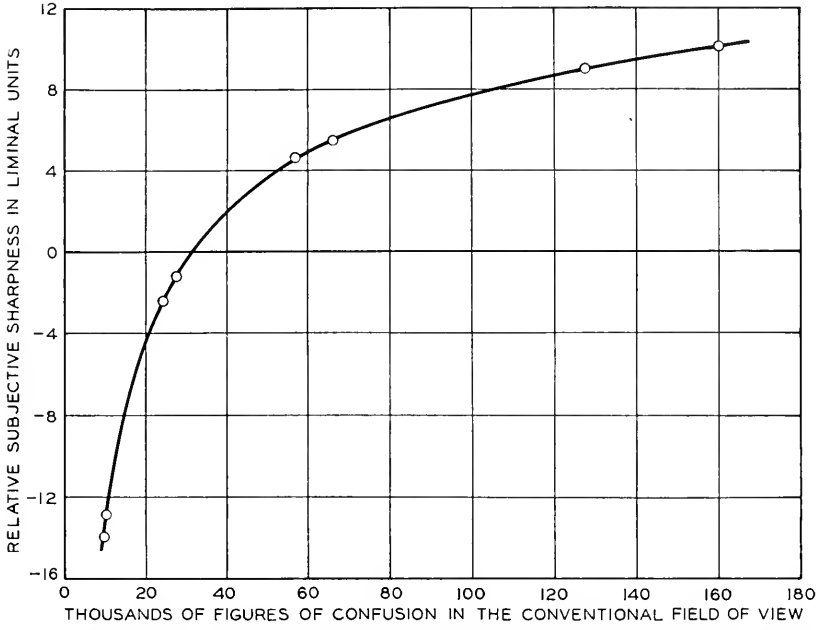


Fig. 1—Sharpness of small-sized motion pictures as a function of resolution. The conventional field of view is a rectangle whose height is 1/4 the viewing distance and whose width is 4/3 the height. Reference sharpness is approximately that of a 240-line, 24 frame per second, 806 kilocycle television image. Curve based on 1,080 observations at a viewing distance of 30 inches.

sented in Fig. 1 by a point in the neighborhood of + 18 units.) It must be remembered that the curve represents judgments made by trained observers under optimum conditions for distinguishing small differences, and that a change as small as one liminal unit, under the conditions of ordinary television viewing, would probably be largely unnoticed.

A better understanding of the meaning of this curve relating sharpness to resolution may be had by examining the experiment in detail. An individual observation was made when one of the observers, watching the projected image, caused the projection lens to be moved from a reference position to a neighboring one and reported which position he judged to yield the sharper image. The motion picture scene was a close-up of a fashion model turning slowly against a plain neutral

background, and was repeated every quarter minute. The observer could have the lens moved whenever and as often as he wanted to before reporting, so that he soon acquired the habit of observing only the most critical portions of the scene. As soon as his report was recorded, completing that observation, he was shown a new pair of lens positions, the same reference one with a different neighboring one, and asked again to report which he judged to yield the sharper image.

We believe that there were no contaminating influences and that only the size of the figure of confusion was varied. No change in brightness or in magnification could be detected. A minute lateral shifting of the image, because of play in the focusing mount of the lens, was completely masked by the continual weave of the film in the gate and the natural motion of the model. Any significance of the position of the observer's control key was destroyed by reversing its connections from time to time, between observations, without the observer's knowledge. No tell-tale sound accompanied the small motion of the lens, and none of the operator's movements could be seen by the observer.

Each one of 15 observers made 84 separate observations of sharpness difference. Expressing the resolution in terms of the angle at the observer's eye subtended by the side of the square figure of confusion, there were four main reference values, namely 0.71, 1.1, 1.7 and 2.8 milliradians (1 milliradian is equal to 3.44 minutes of arc). At each of these reference values there were seven neighboring values, namely 0, 0.045, 0.090, 0.13, 0.18, 0.22 and 0.27 milliradians greater than the reference value. (The 0 in that set means that the reference value was shown against itself, or that the observer was asked to judge a null change; this was intended to keep him on his guard and alert, not to furnish primary data.) Each pair of values was presented to each observer three times, so that there were 45 observations on every pair. The pairs were presented in irregular order according to a schedule, the variation about one reference value being completed before going on to the next. The differences were set up on the basis of preliminary trials to include some which almost none of the observers could detect and some which almost all could. It was explained that some of the differences to be judged would probably be too small for discernment, and that a "no choice" response would be permitted whenever reasonable effort failed to establish a definite choice.

The primary data are shown in Fig. 2. Each point shows the proportion of the observations in which the variable image, which had the poorer resolution by reason of its larger figure of confusion, was nevertheless judged to be sharper than the reference image. Such a judg-

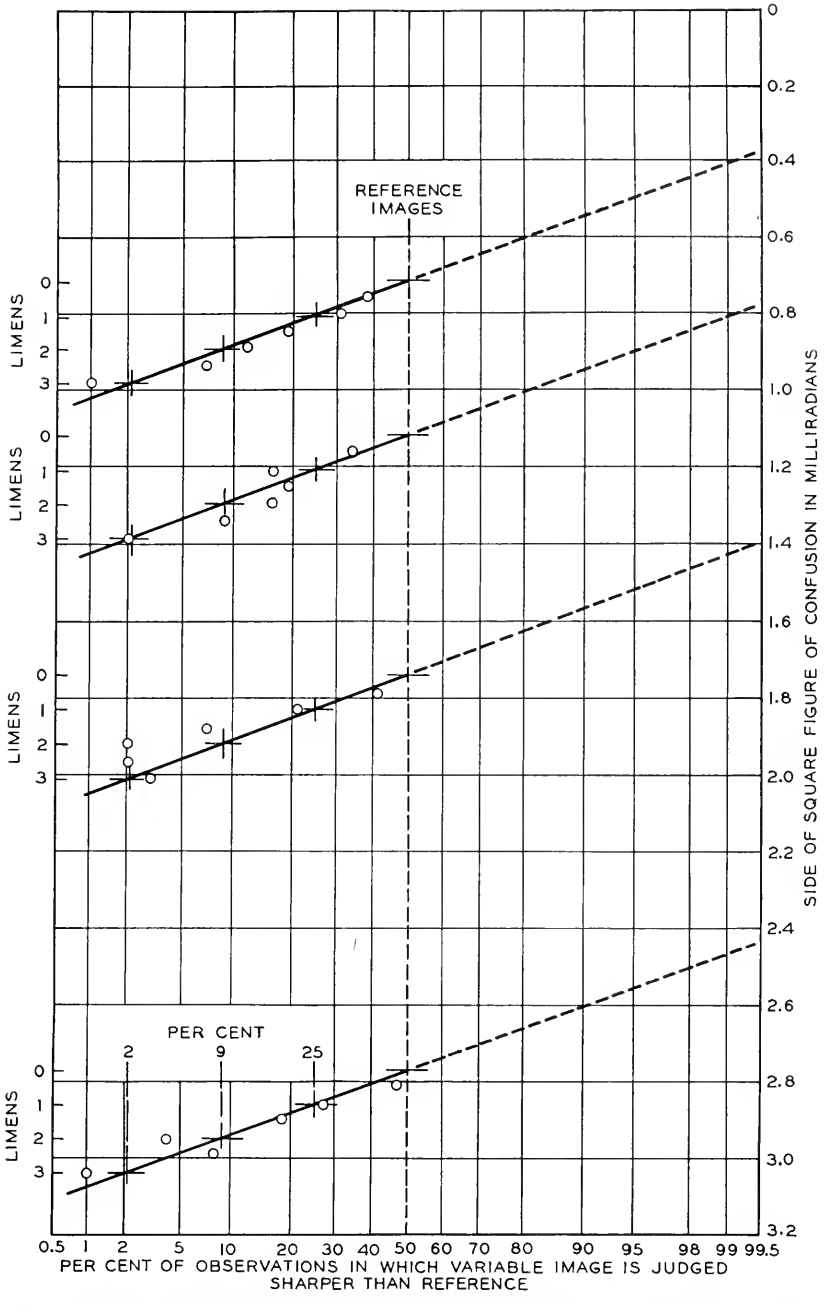


Fig. 2—Distributions of judgments of sharpness differences. The scales of limens denote subjectively-determined units, as explained in the text. Each point represents 45 observations of a small-sized motion picture at a viewing distance of 30 inches.

ment would, of course, be classed as wrong. All reported "no choice" judgments have been distributed equally between the "right" and "wrong" classes. It will be noticed that there was some discrimination at even the smallest change made, that is 0.045 milliradian, and that there was lack of complete discrimination at the largest change, that is 0.27 milliradian. The "no choice" judgments comprised 15 per cent of the total at the smallest change and only 2 per cent at the largest.

It is interesting to note that for the null changes the "no choice" judgments comprised only 17 per cent of the total, indicating either that the observers were reluctant to admit that they were guessing or that they were judging coincidental small changes in the film due to its bending in the gate or to its photographic processing. (We did observe, in establishing the lens position for sharpest focus, that film at the start of a reel required a slightly different lens setting from that at the end of the reel, and we ascribed it to the varying tension during projection, or to the varying degrees of curvature in storage on the reel.)

The four sets of points in Fig. 2 exhibit rather striking similarities. Each set may be fairly represented by a normal error curve (straight line on this arithmetic probability paper). We have drawn in four such normal curves, passing each one through the 50 per cent point at the null change and giving each a common slope. The appropriate value of slope was determined by inspection of an auxiliary plot in which the four reference values were superimposed and the four sets of points were plotted to a common ordinate scale of differences. These normal curves are considered to represent the data as well as any more elaborate relations that might have been used.

We varied the resolution only in the direction of decreasing it with respect to the reference values. We presume that had the variation been in the opposite direction the data would have been represented equally well by the same normal curves, which are accordingly extended in dotted lines.

In Fig. 2 we have indicated the magnitude of a difference of one limen by means of supplementary scales of ordinates. Since the four normal curves have a common slope, the difference limen turns out to have a constant value, 0.090 milliradian (0.3 minute of arc), independent of the size of the figure of confusion in the range from 0.71 to 2.8 milliradians. Why this should be so is a problem of physiological optics which is rather beyond the scope of this paper. The supplementary scales also serve to illustrate the meaning of differences two and three times as large as the difference limen. That is, a change in the side of

the figure of confusion of 0.18 milliradian would be twice as large as the limen of 0.090 milliradian, and would result in wrong judgments in 9 per cent of the observations, corresponding to real discrimination in 82 per cent of them. Likewise a change of 0.27 milliradian would result in real discrimination in 96 per cent of the observations. Any change larger than about three times the limen would be discriminated in practically every instance, under the conditions of our experiment.

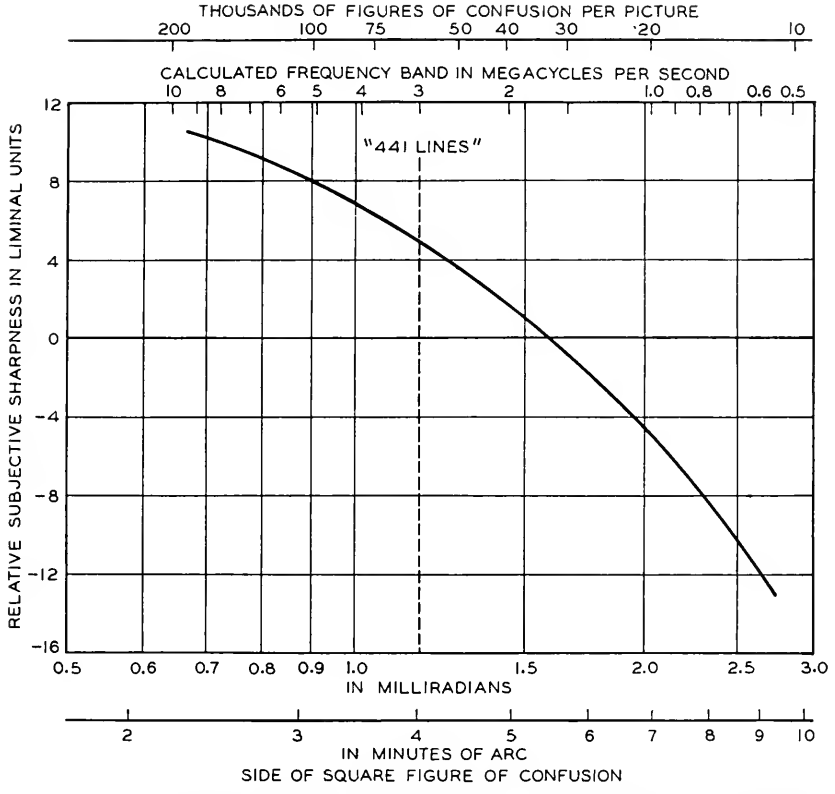


Fig. 3—Sharpness of small-sized motion pictures at a viewing distance of 30 inches. The frequency band is calculated on the basis of a 10-inch by 7½-inch television picture, 30 frames per second, with 15 per cent horizontal and 7 per cent vertical blanking, under the condition of equal horizontal and vertical resolutions.

Figures 3 and 4 show the curve of Fig. 1 replotted in terms of some additional objective variables. A scale of nominal frequency band width required for transmission of the image signal over a video circuit has been worked out on the basis of our comparison of the out-of-focus motion picture with a television image of known characteristics, to be described in section 5. We see that in order to effect an increase in

sharpness which would be practically always discriminated under our experimental conditions, that is, a change of three or four liminal units, the frequency band would have to be increased from say 2.5 megacycles to about 4.5 megacycles. To effect an additional increase of the same

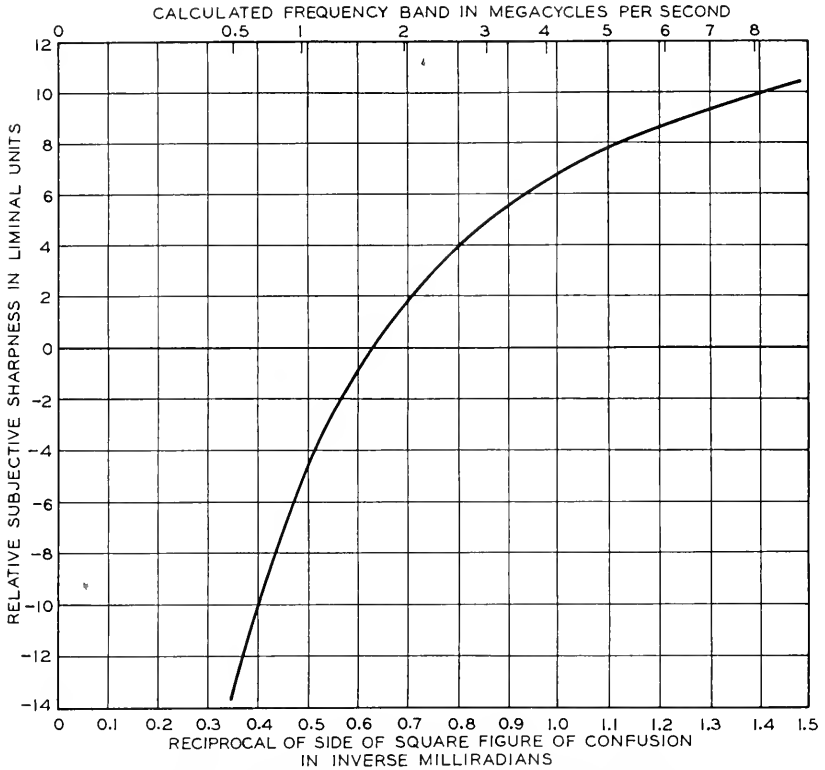


Fig. 4—Sharpness of small-sized motion pictures at a viewing distance of 30 inches. The frequency band is calculated on the same basis as in Fig. 3.

subjective amount would require that the frequency band be increased from 4.5 megacycles to about 10 megacycles. The diminishing return in sharpness is possibly better illustrated by the continually decreasing slope of the curve in Fig. 4, in which the abscissa is proportional to the square root of the frequency band, a factor which may perhaps be interpreted to represent roughly the cost of electrical transmission over a long system. We might infer from this curve that transmission costs are likely to increase faster than image sharpness, other things being equal.

4. Horizontal and Vertical Resolutions

The effect of unequal horizontal and vertical resolutions upon sharpness is shown in Fig. 5. The various rectangular figures of confusion, which were intercompared in a manner which will be described pres-

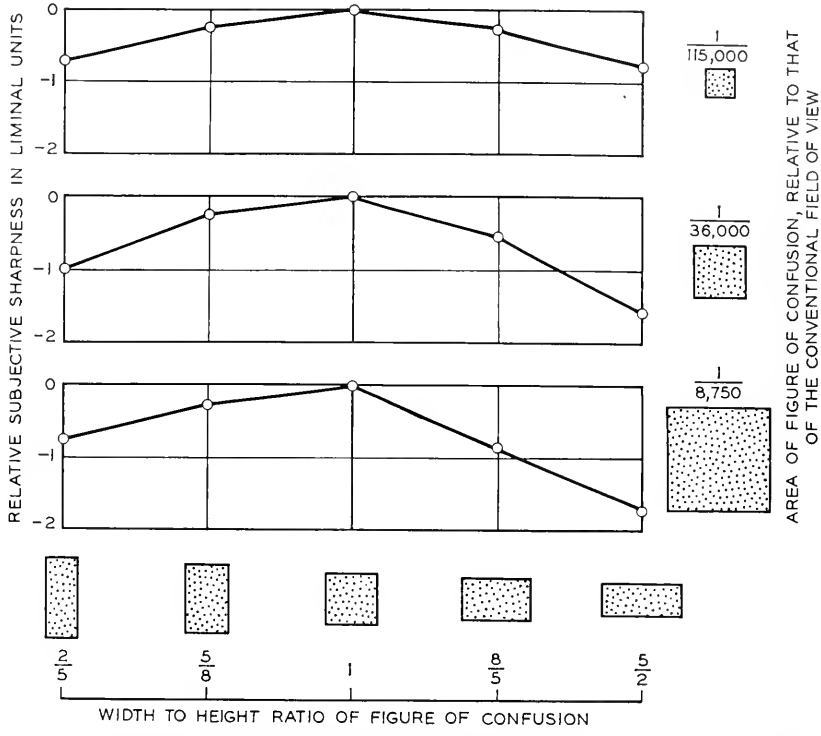


Fig. 5—Sharpness of small-sized motion pictures as a function of the relative values of horizontal and vertical resolutions. The conventional field of view is a rectangle whose height is 1/4 the viewing distance and whose width is 4/3 the height. Each point represents 150 observations at a viewing distance of 30 inches.

ently, are shown along the axis of abscissae, positioned according to the logarithm of the ratio of width to height, for the sake of symmetry. Three curves are shown, each for a different constant value of the area of the figure of confusion, which determines the sharpness for the central square shape (as in Fig. 1). At the right the relative areas are illustrated and specified in terms of the number of figures of confusion in the conventional field of view, whose width is 4/3 of its height and which is viewed at 4 times its height.

Sharpness, the subjective variable, is plotted along the axis of ordinates in liminal units. This unit denotes a difference in sharpness

which corresponds to a difference of one limen in the shape of the figure of confusion. When two images, characterized by different shapes of figure of confusion, are judged by a number of observers, the proportion of the observations in which one image is said to be sharper than the other affords a significant measure of the evaluation of the difference between them. When 25 per cent of the observations show that shape A yields a sharper image than shape B, we say that shapes A and B are different by one limen, and that the image A is less sharp than the image B by one liminal unit. All "no choice" or "equally sharp" judgments are distributed equally between the judgments for A and those for B.

In order to evaluate other than unit differences, we have assumed that a normal error curve describes accurately enough the distribution of sharpness differences in liminal units. Thus, image A is less sharp than image B by two liminal units when it is reported to be sharper than image B in 9 per cent of the observations. The difference is three liminal units when it is reported to be sharper in 2 per cent of the observations. Any difference larger than about three liminal units would indicate practically complete agreement that the one image is less sharp than the other, under our experimental conditions. A distribution of this nature was found to hold for sharpness differences resulting from changes in the area of the figure of confusion, as shown in Fig. 2.

Each shape of figure of confusion was compared with each of the four other selected shapes, and the sharpness differences were expressed in liminal units by the procedure just discussed. A fifth difference, corresponding to a null change, or a shape compared with itself, was presumed to be zero. The average value of these five sharpness differences, averaged in liminal units, measured the relative sharpness of that particular shape with respect to the average sharpness of all five shapes, an unvarying reference. In Fig. 5, the sharpness scales have been shifted so that zero denotes the most preferred one of the shapes, which happened in each case to be the square.

The sharpness curves are found to be slightly skewed with respect to the logarithm of the width : height ratio, there being a small preference for figures of confusion whose long dimension is vertical rather than horizontal. This is believed to be the first evidence of an asymmetric requirement on resolution. It suggests the possibility that the square figure might not have been the most preferred, had we tested other shapes nearer to the square than the ones we did use. With a more searching experiment we might have found that the eye prefers resolution in the horizontal direction to be just a little better than in

the vertical direction. Inasmuch as the effect is fairly small, and found only with the less sharp images, we shall leave it as another problem in physiological optics.

With an actual television image this small skewness would probably be reversed by the attendant coarsening of the scanning line structure. We do not know how much to allow for annoyance caused by visibility of the line structure. Taking our best estimate * of the height of the figure of confusion which would be equivalent in vertical resolution to a just noticeable pattern of scanning lines, we may say that for the uppermost curve in Fig. 5 the scanning line structure would not be noticeable except possibly for the shape marked 2/5. For the central curve the line structure would be noticeable for all shapes except possibly the one marked 5/2. It appears that the skewness and the line structure vanish together as the sharpness is increased.

Figure 5 demonstrates that equality of horizontal and vertical resolutions is a very uncritical requirement on the sharpness of an image, especially of a fairly sharp one. An image somewhat better than present television grade, exemplified by the uppermost curve in Fig. 5, shows a remarkably wide tolerance in this respect. Its figure of confusion could be three times as high as wide, or three times as wide as high, yet any intermediate shape between those two extremes would yield an equally sharp image to within one liminal unit. Under the ordinary conditions of television viewing the difference would be even less marked than that. This would imply that if the square figure of confusion simulates a television image of say 500 lines, then the number of lines could be changed to any value from about 300 to about 850 without altering the sharpness by as much as one liminal unit, under the condition, of course, that all the other pertinent factors, such as frequency band width and number of frames per second, remain unchanged.

The curves in Fig. 5 represent the averaged responses of fifteen observers each viewing five different motion picture scenes. Each one of the five selected shapes of figure of confusion was shown with each other one as a pair, a total of ten pairs. The observer was asked to identify which member of each pair he judged to yield the sharper image, or to report "no choice" if he judged them to be equally sharp. The pairs were scheduled in irregular order, and the observer could have the aperture shape shifted at will. The observers were instructed to consider the whole image area without undue regard for some features to the neglect of others.

* Engstrom [8] estimates that the scanning line structure becomes just noticeable when the spacing of the lines subtends an angle of 2 minutes at the observer's eye. In section 5 we show that the equivalent figure of confusion has a height 1.9 times as great as the spacing of the scanning lines.

5. Comparison of the Out-of-Focus Motion Pictures with a 240-Line Television Image of Known Characteristics

The motion picture machine was arranged to project out-of-focus pictures onto a screen set up beside the cathode ray receiving tube of a laboratory television apparatus [9] of excellent design. Duplicate films were run in the two machines, and the images were made equal in size and approximately equal in color and brightness. Special low-pass filters in the video circuit limited the frequency band without transient distortion, and permitted the trial of three different band widths. The conclusion was reached that the nominal band width of the video circuit, expressed in cycles per frame period, was equal to 1.3 times the number of figures of confusion in the frame area.

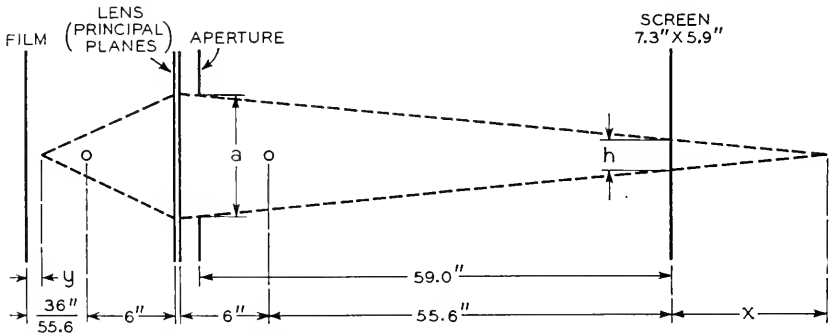


Fig. 6—Essential dimensions of the motion picture optical system as used for the correlation with a 240-line television image. For this case $a^2 = 1.00$ square inch.

$$y = \frac{36}{55.6} \cdot \frac{x}{55.6 + x}$$

$$\frac{h}{a} \doteq 1.45\, y.$$

A group of observers compared the two images, each observer being allowed to adjust the focus of the projection lens until he judged the images to be equal in sharpness. The distribution of lens positions, in terms of microscope scale divisions, was found to follow a normal error curve fairly well, and the median value for the group was used in computing the sizes of the figure of confusion. The external aperture shape was always square.

Since the television film scanner had been designed without regard for the unused space between frames on sound film, it became necessary to modify some of the dimensions of the out-of-focus projection system in order to make the two images equal in size. Figure 6 shows the modified dimensions. Comparison with Fig. 7, which gives the di-

mensions used in the main experiments, will show that the magnification was reduced from 12 to 9.3, the area of the external aperture was increased from 0.49 square inch to 1.00 square inch, and the aperture was mounted 2.6 inches instead of 1.3 inches from the principal planes of the projection lens.

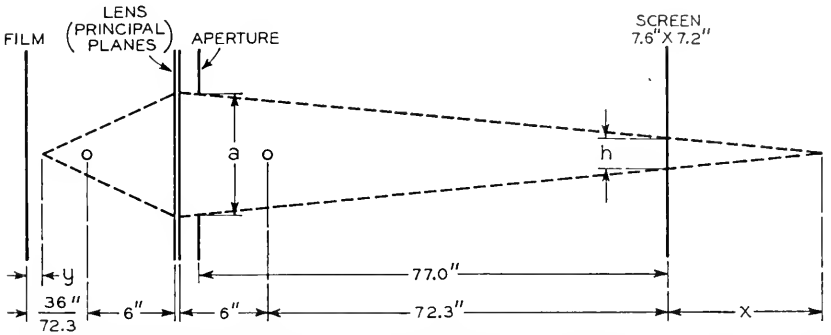


Fig. 7—Essential dimensions of the motion picture optical system as used for the subjective sharpness tests. For this case $a^2 = 0.49$ square inch.

$$y = \frac{36}{72.3} \cdot \frac{x}{72.3 + x}$$

$$\frac{h}{a} \doteq 1.88\, y.$$

The television apparatus was designed for 240 lines, 24 frames per second, and a width : height ratio of 7 : 6. Actually 20 per cent of the frame time was consumed in scanning the blank space between sound film frames, and 10 per cent of the line time was used up by the return sweep in the receiver. The television image, which was the same size as the projected motion picture, was 5.6 inches high. This dimension was 20 per cent less than the height of the entire 240-line field including the blank portion, which was, therefore, 6.9 inches. The width of the entire field including return trace was 6.9 × 7/6 or 8.1 inches, and the width of the television image was 10 per cent less than this, or 7.3 inches. Thus, the total area transmitted per frame period was 6.9 × 8.1 or 56 square inches; the useful image area was 5.6 × 7.3 or 41 square inches.

The three amplitude-frequency characteristics used in the video circuit are shown in Fig. 8. Curve *A* is for two square scanning apertures in tandem, one transmitting and one receiving, each having the height of one scanning line. No electrical band limitation was effective in this case. Curves *B* and *C* are for the addition of each of two special low-pass filters which were carefully phase-equalized and

designed for gradual cut-off. In each case the nominal band width was taken to be the same as that for the aperture effect alone, namely, the frequency at which the loss is 7.8 decibels greater than at low frequency. The addition of a low-pass filter could thus be considered

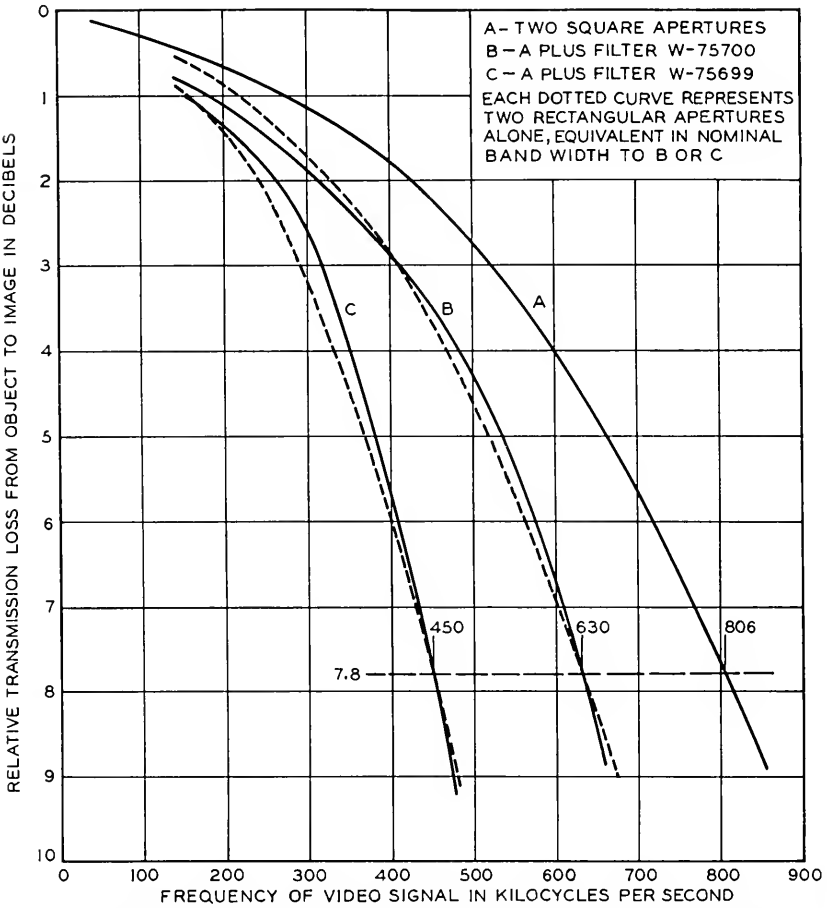


Fig. 8—Amplitude-frequency characteristics of the television system used for correlating the motion picture projection with a television image.

equivalent to an increase in the length of each of the scanning apertures in the ratio of the nominal band widths, as illustrated by the dotted curves.

The results of the comparison were as follows. The number of figures of confusion in the area of the frame was reckoned by dividing the entire area of the television frame, including the blank portion and

the return trace, by the observed area of one figure of confusion of the equally sharp motion picture. The number of cycles per frame period was the nominal band width of the video circuit, in cycles per second, divided by the number of frames per second, or 24.

	Case A	Case B	Case C
Figures of confusion per frame	22,400	18,900	14,800
Cycles per frame period	33,600	26,200	18,800
Ratio	1.50	1.38	1.27

The ratio in Case A was suspected to be too large because of unaccounted-for small defects in the film scanner which degraded the image sharpness more than was indicated by the aperture effect alone. The difference in ratio between Cases B and C was no larger than the measured probable error of each set of observations. Making allowance for these things, we concluded that the ratio between the number of cycles per frame period and the number of figures of confusion per frame area had been found to be 1.3.

This factor 1.3 gave us a basis for calculating the television aperture loss in the direction normal to the scanning lines, and enabled us to compute the nominal video frequency band required to yield an image having equal horizontal and vertical resolutions.

The stepped nature of the brightness variation across the scanning lines of a television image, in contrast to its continuous nature along the lines, gives rise to the requirement that for equal resolution in the two directions the scanning apertures must be longer in the scanning direction than they are across it. The extent of this departure from squareness has been estimated (see references 3 to 7) at from 1.2 to 1.9 mostly on theoretical grounds. Our comparison of a television image of known characteristics with a controlled out-of-focus motion picture furnished a subjective measurement of the effect which yielded the value 1.4 for the ratio of width to height of the scanning apertures for equal resolution. We take width to mean the dimension along the scanning lines, and height to mean the dimension normal to them.

We found that the nominal video band width of a television signal, in cycles per frame period, was 1.3 times the number of figures of confusion per frame area in the equally sharp motion picture. This meant that the area of each figure of confusion was 1.3 times as great as the area of one scanning line over a (scanned) length of one cycle. By the adopted definition of nominal video band width, the length of one cycle was just twice the length of each one of the pair of rectangular scanning apertures which were considered equivalent to the actual

square apertures plus the filter. According to the scanning theory of Mertz and Gray,[4] the pair of apertures in tandem was equivalent, in frequency limitation, to a single aperture 1.35 times as long as either one of the pair. Taking the width of this single aperture (1.35 times the length of one half cycle) equal to the width of the figure of confusion, the height of the figure of confusion was calculated from its area to be 1.9 times the height of one scanning line or one scanning pitch. This was the measure by subjective comparison of the resolution across the scanning lines.

Under the condition of equal resolution along and across the scanning lines, the figure of confusion would have to be square and its width would then also be 1.9 times the scanning pitch. The width of each one of the pair of equivalent tandem scanning apertures would be 1.9/1.35 or 1.4 times the scanning pitch. That is, two rectangular scanning apertures, each 1 line high and 1.4 lines wide, used in tandem without electrical band limitation, would yield an image having equal resolution along and across the scanning lines.

The nominal frequency band associated with such scanning apertures is 1/1.4 times that associated with square apertures. That is, the nominal video frequency band, in cycles per frame period, required for equal horizontal and vertical resolution is 0.70 times one half the number of square scanning elements per frame area, reckoning a square scanning element as an area of height and width equal to the scanning pitch, or spacing between scanning lines.

For comparison with the value 0.70 which we have just found, the following values of nominal band width coefficient have been lifted from their contexts in the references:

(*a*)	Kell, Bedford and Trainer (1934)	0.64
(*b*)	Mertz and Gray (1934)	0.53
(*c*)	Wheeler and Loughren (1938)	0.71
(*d*)	Wilson (1938)	0.82
(*d*)	Kell, Bedford and Fredendall (1940)	0.85

Acknowledgment

This work has been done under the direction of Dr. P. Mertz and with the extensive assistance of Mr. T. R. D. Collins. To them, and to my other colleagues who have given of their time and counsel, I wish to extend my appreciation of their help.

APPENDIX

1. Determination of the Size of the Figures of Confusion

The image was put out of focus by moving the projection lens nearer to the film gate, throwing the plane of sharp focus beyond the viewing

screen. Assuming for the moment that the optical imagery was perfect, each point of the film gave rise to a pyramidal volume of light whose base was the opening of the external aperture and whose apex was the point's image in the new focal plane beyond the screen. The intersection of this pyramid with the viewing screen was the geometrical figure of confusion for that point. The shape of the figure was geometrically similar to that of the aperture, and the side of the figure was to the corresponding side of the aperture as the distance from focal plane to screen was to the distance from focal plane to aperture.

The distance of the focal plane beyond the screen was related to the displacement of the lens from the "in focus" position by means of the simple lens formula, and this relation was verified by actual measurement of the distances. The geometrical area of the figure of confusion was thus known in terms of the lens displacement, as shown in Fig. 9.

Efforts to check this relationship by direct measurement of the dimensions of the figure of confusion in the plane of the screen were nullified by the aberrations of the optical system, especially by the residual chromatic aberration. A comparison method was therefore devised in which the out-of-focus image of a very thin vertical slit was compared with an actual slit in the plane of the screen. In the film gate was placed a glass plate bearing a sputtered layer of gold with a razor-blade scratch not wider than 0.0001 inch in selected portions. In the plane of the screen was placed a back-lighted slit made by cementing the two halves of a cut piece of thin black paper onto a piece of translucent white paper. This slit had sharp, parallel edges and uniform brightness over its width, which was easily made as small as 0.005 inch. A set of these slits was prepared, ranging in width up to 0.100 inch, and each one was observed, without optical aid, close beside the projected out-of-focus image of the scratch in the gold film. The apparent brightnesses were equalized by means of neutral-tint filters behind the paper slit.

The ranges of values of lens displacement and of external aperture shape which were used in the experiments were tested in this way, by adjusting the out-of-focus images to subjective equality with the sharp-edged slits. In every case the measured width of the comparison slit turned out to be about 15 per cent less than the calculated geometrical width of the projected image. This seeming a not unreasonable measure of the effect of the aberrations, it was adopted as a factor for converting geometrical sizes into subjective sizes of the figures of confusion.

Figure 9 shows both the calculated geometrical area and the observed subjective area of the figure of confusion in terms of the displacement

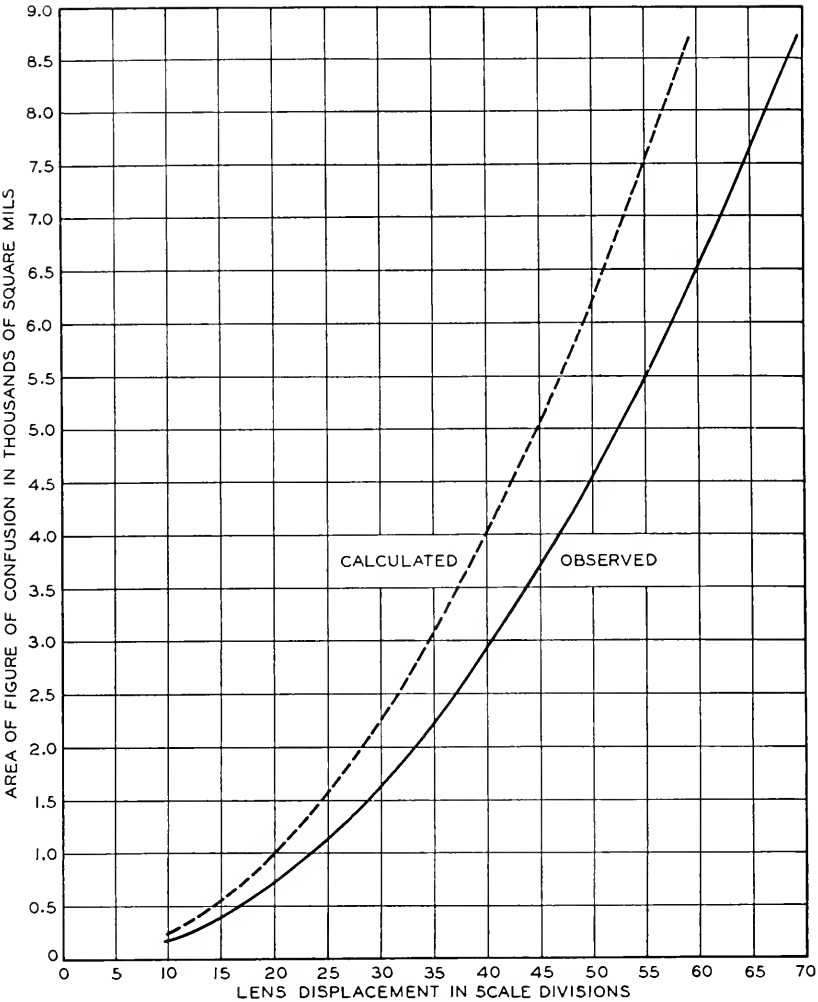


Fig. 9—Calibration curve for the motion picture optical system as used for the subjective sharpness tests.

of the lens from its sharp focus position. The lens displacement is expressed in microscope scale divisions, the working variable. Figure 7 shows the dimensions of the optical system.

2. Equipment and Conditions of the Experiment

Light Source

A ground glass screen ½ inch behind the film, illuminated by a 1,000-watt projection lamp and double condensing lens system. This served

to break up the image of the lamp filament which otherwise would have been formed near the principal planes of the projection lens and would have destroyed the uniformity of illumination within the figures of confusion of the out-of-focus image on the screen. The screen brightness was about 10 foot-lamberts with the projector running without film.

Projection Machine

Acme Portable, with two-bladed shutter. There was no provision for reproducing the sound track. The screen image, in sharp focus, was said by competent judges to represent very good motion picture projection practice.

Projection Lens

Bausch and Lomb Series "0", 6.00-inch focus. There was fitted over the lens barrel a brass ring with an extremely sharp turned edge to serve as an index for the measurement of lens displacement. The lens could be set to the nearest 0.0003 inch by means of the focusing mechanism. The image was put out of focus by moving the lens toward the film. At sharp focus the linear magnification was 12 times.

Measuring Microscope

Mounted rigidly on the frame of the projector, and fixed with respect to the film gate. The micrometer scale was focused on the index mark on the barrel of the projection lens. A lens displacement of 0.060 inch caused the index to traverse 50 divisions of the scale.

External Aperture

An adjustable black paper mask mounted $1\frac{1}{3}$ inches from the principal planes of the lens, on the screen side. The opening was rectangular, with sides horizontal and vertical, of constant area 0.49 square inch. The ratio of height to width could be varied continuously from 2.5 to 0.40 without changing the area. The opening was uniformly filled with light under all conditions.

Viewing Screen

White Bristol board, 7.2 inches high by 7.6 inches wide (the image size of an available television receiver to be used for comparison). The screen was hung at the back of a black-velvet-lined box 18 inches high, 22 inches wide and 12 inches deep. The viewing distance was always 30 inches.

The viewing room was completely darkened except for a little stray light from the projection machine.

Scene 1

Scene 2

Scene 3

Scene 4

Scene 5

Fig. 10

Scene 1 reproduced by courtesy of Loucks & Norling.
Scenes 2 and 3 reproduced by courtesy of Fox Movietone News.
Scenes 4 and 5 reproduced by courtesy of Paramount News.

The Observers

The observers were almost all Laboratories engineers associated with television research and transmission problems. The average observer devoted about one hour to the experiment on unequal horizontal and vertical resolutions, and about three hours (in two sessions) to the experiment on small differences in resolution. Each observer was carefully instructed with regard to the purpose and the mechanism of the experiments, and was allowed to examine trial pictures to see clearly the effects of changing the shape and size of the figure of confusion.

Motion Picture Film

Standard 35 mm. black-and-white sound film on safety base. The area projected onto the screen was 0.600 inch high by 0.633 inch wide.

For the experiment on unequal horizontal and vertical resolutions, five different scenes were used. Sample frames from them are shown in Fig. 10. For the experiment on small resolution differences, Scene 3 was selected as the most suitable on the basis of photographic excellence and picture content, and this alone was used. Each of the scenes was about one quarter of a minute in length, and was shown repeatedly. Brief descriptions follow:

Scene 1: A country-side landscape, with trees and fields. A center of interest is the tall steeple of a white church on the distant hillside. A concrete highway flanked by a white fence carries cars into and out of the picture. There is no fast motion.

Scene 2: A full-length view of a girl modeling an evening dress moving slowly against a dark, fluted backdrop. A large vase of flowers is a secondary center of interest.

Scene 3: A close-up view of a girl modeling a hat, turning slowly against a plain, neutral background.

Scene 4: A street scene of an Indian parade, with a background of store windows and signs. The parade moves rather rapidly, and there is some motion among the by-standers.

Scene 5: A closer view of some of the Indians in the parade. There is much fine detail in the costumes, and the motion is rapid.

References

1. M. Luckiesh and Frank K. Moss, "The Dependency of Visual Acuity Upon Stimulus-Distance," *Jour. Opt. Soc. Amer.*, **23**, 25 (Jan. 1933).
2. J. P. Guilford, "Psychometric Methods," McGraw-Hill, New York, 1936.
3. R. D. Kell, A. V. Bedford and M. A. Trainer, "An Experimental Television System," *Proc. I.R.E.*, **22**, 1246 (Nov. 1934).

4. P. Mertz and F. Gray, "A Theory of Scanning and its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television," *Bell Sys. Tech. Jour.*, **13**, 464 (July 1934).
5. H. A. Wheeler and A. V. Loughren, "The Fine Structure of Television Images," *Proc. I.R.E.*, **26**, 540 (May 1938).
6. J. C. Wilson, "Channel Width and Resolving Power in Television Systems," *Jour. Telev. Soc.*, **2**, 397 (June 1938).
7. R. D. Kell, A. V. Bedford and G. L. Fredendall, "A Determination of Optimum Number of Lines in a Television System," *RCA Review*, **5**, 8 (July 1940).
8. E. W. Engstrom, "A Study of Television Image Characteristics," *Proc. I.R.E.*, **21**, 1631 (Dec. 1933).
9. H. E. Ives, "The Transmission of Motion Pictures over a Coaxial Cable," *Jour. S.M.P.E.*, **31**, 256 (Sept. 1938).

Cross-Modulation Requirements on Multichannel Amplifiers Below Overload

By W. R. BENNETT

Interchannel interference caused by non-linearity of multichannel amplifier characteristics is analyzed in terms of second and third order sum and difference products of the bands of energy comprising the various channels. Methods of relating the resulting disturbance to discrete frequency measurements are described and means for arriving at modulation requirements on individual amplifiers thus established.

1. Introduction

WHEN a repeater is used to amplify a number of carrier channels simultaneously, departure from linearity in the response as a function of input amplitude tends in general to produce interference between the channels. The non-linear component of the amplifier characteristic in effect acts as a modulator, changing the frequencies in the input wave and producing components which fall in bands assigned to channels other than the original ones. This phenomenon has been called "interchannel modulation" or "non-linear crosstalk." In formulating the requirements which are imposed on a repeater to insure that the resulting interference between channels will not be excessive it is convenient to treat separately two aspects of the problem namely—the condition when the total load on the amplifier is within the range for which the amplifier is designed and the severely overloaded condition. Actually a transition region between these two cases must also exist but when a considerable amount of negative feedback is used the break in the curve of response vs. input is quite sharp so that for practical purposes the input may be said to be either below or above the overload value.

Below overload, the amplifier characteristic may in most cases be represented with sufficient accuracy by the first few terms of a power series and the interchannel modulation analyzed in terms of the resulting combination tones of the frequencies present in the different channels. The total interference resulting from the combination tones falling in individual channels must be kept below prescribed limits. Above overload on the other hand the resulting disturbance in all channels becomes quite large and requirements are based on making such occurrences sufficiently infrequent.

The load capacity requirement has been discussed in a paper by B. D. Holbrook and J. T. Dixon.[1] In the present paper we assume that the principles there developed have been used to fix the maximum load which the amplifier must deliver and proceed with the second phase of the problem—the determination of the modulation requirements which the amplifier should meet below overload. A secondary objective is to set up simple testing procedures by which the performance of the system may be assessed without incurring the complications attendant upon loading the system with talkers. One such procedure involves the measurement of distortion products by a current analyzer when sinusoidal waves are impressed upon the system. Another involves the measurement of noise in a narrow frequency band when a band of noise uniformly distributed over the transmission range is impressed upon the system.

The path followed in reaching these objectives starts in Section (2) with a demonstration of the way in which non-linearity leads to interference in specific cases. Section (3) then considers the magnitudes of typical modulation products and arrives at volume distributions for them, by which the fluctuating character of speech may be taken into account. The basis for treating interchannel modulation as noise is given in (4). Since one of the most convenient testing methods involves the use of sine waves the relationship of distortion measured with sine waves to the distortion observed with speech on the system must be set up as in (5). The number of products falling in any single channel is considered in (6) and the average noise in any channel may then be evaluated in (7) with the aid of the results of preceding sections. The effects introduced by multiple centers of distortion in the amplifiers of a transmission link are considered in (8). The paper concludes with a discussion of test methods presented in the ninth section.

2. Interchannel Modulation as a Source of Crosstalk

We shall consider specifically a single sideband suppressed-carrier multichannel speech transmission system of the four-wire type although much of the treatment is also applicable to other kinds of carrier systems. The carrier frequencies will be assumed to be adjacent harmonics of a common base frequency greater than the highest signal frequency. The amplifier characteristic will be assumed to be expressible with sufficient accuracy by means of linear square and cube law terms. That is if i_p represents the output current and e_g the input voltage we write

$$i_p = a_1 e_g + a_2 e_g^2 + a_3 e_g^3. \tag{2.1}$$

[1] *Bell System Technical Journal*, Oct. 1939, Vol. 18, pp. 624–644.

Multivalued characteristics such as associated with ferromagnetic materials and reactive characteristics in which the coefficients vary with frequency are not included. The mechanism by means of which the characteristic (2.1) gives rise to interchannel interference may be illustrated by assuming that a sinusoidal signal of frequency q radians per second is impressed on the voice frequency channel associated with the carrier frequency mp, where p is the base frequency in radians per second and m is an integer. The resulting wave impressed on the amplifier is of the form

$$e_g = Q \cos (mp + q)t, \tag{2.2}$$

if upper sidebands are transmitted; the plus sign would be replaced by a minus sign in a system using lower sidebands. Substituting the value of e_g given by (2.2) in the characteristic (2.1), we find:

$$i_p = a_1 Q \cos (mp + q)t + \tfrac{1}{2}a_2 Q^2 + \tfrac{1}{2}a_2 Q^2 \cos (2mp + 2q)t + \tfrac{3}{4}a_3 Q^3 \cos (mp + q)t + \tfrac{1}{4}a_3 Q^3 \cos (3mp + 3q)t. \tag{2.3}$$

Considering the terms which appear in the response (2.3), we note that the first term is the desired amplified signal. The second term is a direct current of trivial importance; if the system contains a transformer, this component is not transmitted. If q does not exceed $p/2$, the third term will be received in the channel associated with the carrier frequency $2mp$ and will there produce a detected frequency twice as great as the original signal frequency. In such a case, it represents interference produced in the $2mp$-channel when the mp-channel is actuated. If q exceeds $p/2$, the interference falls in the $(2m + 1)p$-channel. The fourth term of (2.3) is received in the mp-channel and represents a non-linearity in intrachannel transmission since its frequency is the same as that of the applied signal but its amplitude is proportional to the cube of the impressed signal amplitude. This term is of trivial interest in the study of transmission quality of individual channels of a well-designed multichannel system and is of no interest in the interference problem with which we are here concerned because it is received only in the originating channel. Finally, if q is less than $p/3$, the fifth term represents interference of frequency $3q$ received in the channel associated with the carrier frequency three times that of the originating channel; if q is greater than $p/3$, the interference falls in a higher channel.

Next suppose that a number of carrier channels are simultaneously transmitting signals. By substituting an expression representing the resulting carrier wave, which is a sum of several terms such as (2.2) with different values of Q, m and q, in the amplifier characteristic (2.1),

we find that in addition to interference in channels having twice and three times the fundamental carrier frequencies, there are modulation terms appearing in channels having carriers which are various combinations of sums and differences of the original carrier frequencies. The second order term gives rise to crosstalk products with carriers $(m + n)p$ and $(m - n)p$ as well as $2mp$. The third order term causes products with carriers $(2m + n)p$, $(2m - n)p$, $(l + m + n)p$, $(l + m - n)p$, and $(l - m - n)p$ as well as $3mp$. In the above l, m and n are integers. For convenience we represent the tones associated with the carriers lp, mp, np by A, B and C and designate the various types of products as $A + B$, $A - B$, $2A + B$, $2A - B$, $A + B + C$, etc. It will be noted that the resultant modulation falling in a particular channel at any instant depends on the particular loading conditions prevailing on other channels at the same instant, and that a wide variety of amplitudes, numbers, and types of products are possible. Detailed study of these possibilities is necessary for the solution of our problem.

3. Nature of Multichannel Speech Load and Resulting Modulation Products

Considering any individual channel of the system, we note that (1) it may be active or inactive and (2) if active, the signal power being transmitted may vary throughout a considerable range of values. With regard to (1), we may estimate from traffic data a probability τ that a channel is active.[2] With regard to (2), data are available on the distribution of volumes corresponding to different calls at the toll switchboard. By "volume" is meant the reading of a volume indicator of a standard type. The distribution is approximately normal and hence may be expressed in terms of the average value V_0 and standard deviation σ. In mathematical language, the probability that the volume from any subscriber is in the interval dV at V is given by

$$p(V)dV = \frac{1}{\sigma\sqrt{2\pi}} e^{-(V-V_0)^2/2\sigma^2} dV. \tag{3.1}$$

The value of V_0 is about 16 db below reference volume, or about $-$ 8 vu when measured on the new volume indicator recently standardized in the Bell System. The value of σ is about 6 db. It is to be noted that volume is proportional to the logarithm of average speech power and hence V_0 is not the volume corresponding to the mean of the average speech powers of different talkers. The latter quantity, which will

[2] A channel is said to be active whenever continuous speech is being introduced into it. See Reference (1).

be designated by V_{0p}, may be calculated by averaging the distribution according to power, thus

$$\begin{aligned} V_{0p} &= 10 \log_{10} \overline{\text{antilog}_{10}\, V/10} \\ &= V_0 + \frac{\sigma^2}{20} \log_e 10 \\ &= V_0 + .115\, \sigma^2, \end{aligned} \tag{3.2}$$

when V_0 and σ are expressed in db. The method of obtaining this result is indicated in Appendix A.

It will be convenient to extend the term "volume" to apply to modulation products by designating the modulation product produced by 0-vu talkers as a "zero volume modulation product" of its particular type. This is not its absolute volume as read by a volume indicator, but a reference value to which modulation products of the same type produced by talkers of other volumes may be referred. We assume on the basis of a power law of modulation that the volume of a product will increase one db for each one db increase in volume of a fundamental appearing once in the product, two db for each db increase in volume of a fundamental appearing twice, etc. Thus a $(2A - B)$-product should increase two db for one db increase in the volume of the A-component, and one db for one db increase in the volume of the B-component. If the fundamental talker volumes producing a particular product are normally distributed on a db scale, it follows from established relations concerning the distributions of sums [3] of normally distributed quantities that the volume of the product is also normally distributed. The relations between average and standard deviation for the modulation product and the corresponding quantities V_0 and σ for the fundamental are shown in Table I.

TABLE I

Modulation Product	Average in db Referred to Product from 0-vu Talkers	Standard Deviation in db
$2A$	$2V_0$	2σ
$A \pm B$	$2V_0$	$\sqrt{2}\sigma$
$3A$	$3V_0$	3σ
$2A \pm B, B - 2A$	$3V_0$	$\sqrt{5}\sigma$
$A + B \pm C, A - B - C$	$3V_0$	$\sqrt{3}\sigma$

That is, if the fundamental talker volumes are normally distributed with average value -8 vu, and standard deviation 6 db, the

[3] Multiplying amplitudes of fundamental components is equivalent to adding logarithms of amplitudes; hence the volumes of the fundamental components add to determine product volumes. For a derivation of the distribution function of the sum of two independent normally distributed quantities, see Cramer, Random Variables and Probability Distribution, Cambridge Tract No. 36, 1937, p. 50.

$(A + B - C)$-type products, for example, are also normally distributed in product volume with average value 24 db less than the product produced by three 0-vu talkers and standard deviation $6\sqrt{3}$ or 10.4 db. To obtain the product volume corresponding to the average power of the product distribution, $.115(10.4)^2$ or 12.4 db must be added. In general if V_{0p} of an x-type product is desired, it may be expressed as $\eta_x V_0 + .115\lambda_x\sigma^2$ where η_x is the order of the x-type product, and the value of λ_x is given by the square of the coefficient of σ in the third column of Table I.[4] We observe that $\eta_x V_0 + .115\lambda_x\sigma^2 = \eta_x V_{0p} + .115(\lambda_x - \eta_x)\sigma^2$, and that $\lambda_x = \eta_x$ for $x = A \pm B$ and $A \pm B \pm C$.

The frequencies present in a typical commercial speech channel extend over a range of approximately 3000 cycles. The spacing of carrier frequencies must be made somewhat greater than this to allow for filter cut-offs. Figures 1 and 2 illustrate the spectra of the various second and third order modulation products resulting from two and three fundamental channel spectra respectively which are flat from 10 per cent to 80 per cent of the carrier spacing. Actual speech channels would have peaked spectra but the results would be roughly similar. Each second order band of products occupies twice the frequency range of one original speech band, and a third order band of products spreads over three times the fundamental range. Portions of one product band may thus be received in different channels, but with one part usually much larger than the others. It is to be noted that a $2A$-type product band does not consist merely of the second harmonics of all tones in the band A, but includes all possible sums of the tones in the fundamental band. The spectrum of the $2A$-type product is similar in shape to that of an $(A + B)$-type product but has half as much total power because only half as many sum products can be formed from a single band as from two equal bands. The interfering effect of a $2A$-type product from a speech channel may of course be quite different in character from that of an $(A + B)$-type product since in the latter case the result depends on two independent talkers.

4. The Noise Resulting from Modulation Products

It will be noted that the interference produced as described above may be classified as unintelligible, since in products involving one channel, the wave form is distorted, and in products involving more than one channel, sums and differences of independent signal frequencies are heard. It may be said therefore that interchannel modulation

[4] In general for a $(m_1A \pm m_2B \pm m_3C \pm \cdots)$-product, $\lambda_x^2 = m_1^2 + m_2^2 + m_3^2 + \cdots$.

may be treated as noise, and the usual noise requirements apply. If a great many products are superimposed, the noise heard will be fairly steady, and the average weighted noise power is a sufficient indication of the interfering effect. In systems with a small number of channels, large variations in the noise may occur, and it may be necessary to consider the infrequent large bursts of modulation from exceptionally loud talkers as a limiting factor. The allowance to be made can be estimated by determining the complete distribution curve of modulation noise. Computation of the required distribution function may be carried out by methods similar to those described in the paper by Holbrook and Dixon.[1] The fact that the products are not independent introduces a difficulty which complicates the calculation. For systems with a large number of channels, the requirements may be based on average values with a considerable resulting simplification.

If in addition to the sidebands due to speech, "carrier leaks" (partially suppressed carrier waves) are present, modulation products are produced which are sums and differences of carrier frequencies and speech sidebands. Products of this sort may cause intelligible crosstalk. For example the carrier frequency mp modulating with the channel frequency $np + q$ causes an $(A + B)$-product of frequency $(m + n)p + q$, which is received in the channel with carrier frequency $(m + n)p$ as the original signal frequency q. Requirements on intelligible crosstalk are in general more severe than on unintelligible; hence it is important that the carrier leaks be suppressed well below the level of the speech channels. The intelligibility tends to disappear as the number of channels is increased, since the number of superimposed products becomes larger thereby producing masking effects. Carrier leak modulation is however more serious than modulation from speech channels having the same power since carrier leaks are present all the time, while speech sidebands occur only in active channels. Similar considerations apply to pilot and control tones.

Quantitatively, the various frequency components in modulation noise must be weighted in terms of their interfering effect on reception of speech. In practice the weighting is done by a noise meter designed for that purpose. The noise meter readings are expressed in terms of db above reference noise. A reading of zero, or reference noise, is produced by a 1000-cycle sinusoidal wave with mean power equal to one micromicrowatt. The weighting incorporated in the noise meter is determined by judgment tests of the relative interfering effects of single frequencies and other reproducible noises.

5. Relation Between Speech and Sine Wave Modulation

A goal of our investigations is to express the requirements finally in terms of measurements which can be made on amplifiers with sinusoidal testing waves. A means of relating modulation products produced by speech channels to those occurring when discrete frequencies are applied is therefore needed. For our purposes here we shall express the needed relation [5] in terms of a "Speech-Tone Modulation Factor," which we shall abbreviate as S.T.M.F. and define in terms of the following procedure: Apply the fundamental single-frequency test currents necessary to produce the product in question, which we shall designate as an x-type product. Adjust each fundamental to give mean power of one mw. at the zero level point of the system. Measure the resulting x-type product at the point of zero transmission level of the system. Suppose the product is H_x db below one fundamental. Next load the system with speech from the combination of fundamental talkers required to form the talker product of x-type. Each talker must produce speech volume of 0 vu at the transmitting toll switchboard or point of zero transmission level. The product is then received from the appropriate channel and a comparison is made between it and the speech from one talker with both talker and product at the same transmission level point in the system. The comparison should be made on the basis of relative interfering effect. Suppose it is determined that an x-type product is L_x db below one 0-vu talker. Then s_x, the S.T.M.F. for an x-type product, is defined by

$$s_x = L_x - H_x. \tag{5.1}$$

The sign of the S.T.M.F. has been assigned here to be positive when the difference in db between effect of talker and talker product is greater than the difference between sine wave fundamental power and sine wave product power.

It is to be noted that not only does each type or product possess its own S.T.M.F., but also that the several portions of a product appearing in different channels have different S.T.M.F.'s. This may be clearly seen from Figs. 1 and 2. We note also that the S.T.M.F. is a property of the system on which the measurements are made, since it varies with the band width of the channels, the spacing of carrier frequencies, and the extent of departures from the simple square and cube law representation of the amplifier characteristic. It also varies with the type of transmitting and receiving instruments used. Theo-

[5] The quantity here defined is related to what has been called "staggering advantage" of modulation products. Since the term "staggering advantage" has been applied to various kinds of interference including linear crosstalk, its use here might lead to confusion and is avoided.

retically it should be possible to compute the S.T.M.F. for any particular product if sufficient information concerning the properties of speech, the transmitting and receiving instruments, the carrier system

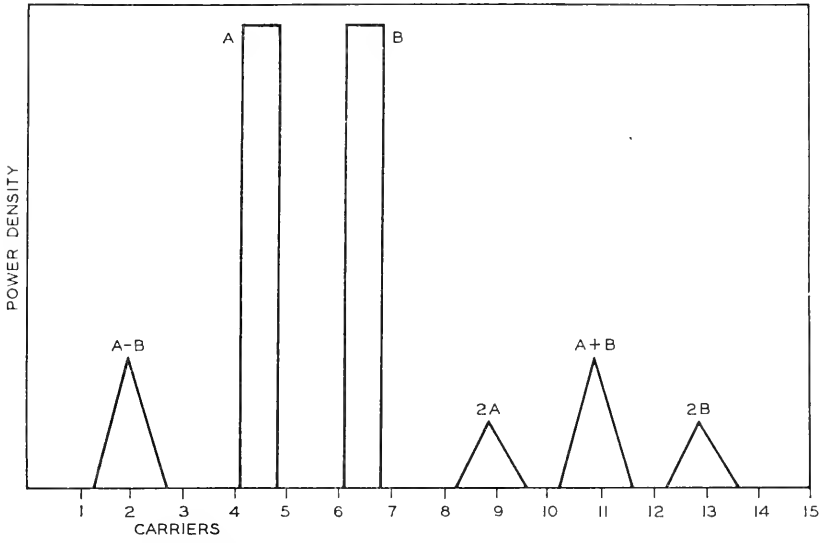


Fig. 1—Spectra of second order modulation products from two fundamental channels.

itself, and the ear were known, but in practice it is found best to use experimental determinations. In the case of new systems for which experimental data are not available, estimates based on known systems of similar type would be used.

6. Number of Products Falling in Individual Channels

In the appendix it is explained how the total number of possible products of each type falling in individual channels may be counted. Table II shows the result of counting all second and third order type products.[6] Results for products falling both within and without the fundamental band are given. In certain of the $(A + B - C)$- and $(A - B - C)$-type products, the channel in which the product occurs also is the source of one of the fundamentals. Since this would give a type of interference heard only when the disturbed channel is also carrying signal, it is not in general as serious a form of crosstalk as the cases of independent fundamental and product frequencies. Therefore the number of these special kinds of products has also been evaluated

[6] Mr. J. G. Kreer collaborated in the derivation of these formulæ.

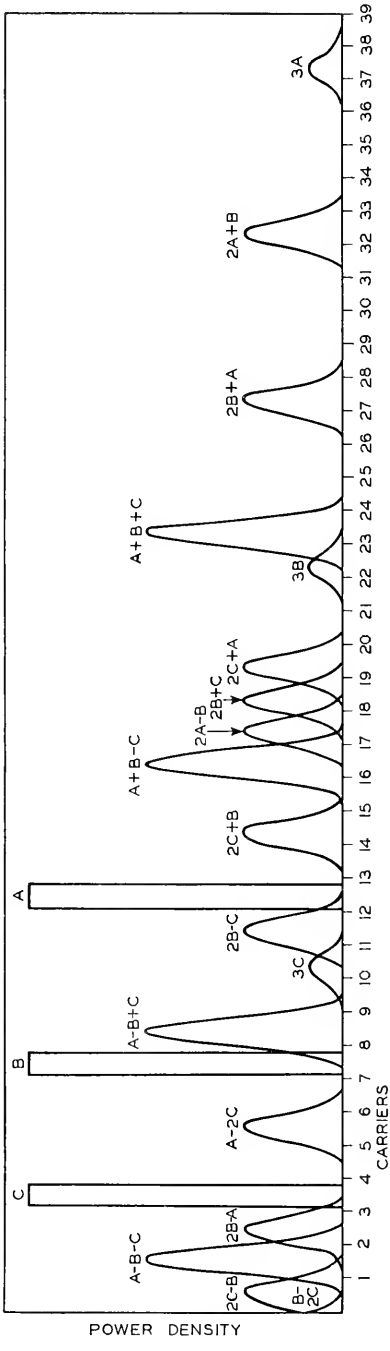


Fig. 2—Spectra of third order modulation products from three fundamental channels.

TABLE II

NUMBER OF PRODUCTS FALLING IN kTH CHANNEL OF MULTICHANNEL CARRIER SYSTEM WITH HARMONIC CARRIER FREQUENCIES n_1p, $(n_1+1)p$, ... $(n_1+N-1)p$

n_1p = Lowest Carrier Frequency.
$n_2p=(n_1+N-1)p$ = Highest Carrier Frequency.
kp = Carr. Freq. Associated with Mod. Product.
N = Number of Channels.
$I(x)$ = "Largest Integer $\leq x$."
No. of Products is 0 Outside Ranges Indicated.

NUMBER OF PRODUCTS

Type

Second Order

$2A$ ..... 1, $n_1 \leq \frac{k}{2} \leq n_2$ if $\frac{k}{2}$ is an integer

$A+B$
$$\begin{cases} I\left(\frac{k+1}{2}\right)-n_1, & 2n_1-1 \leq k \leq n_1+n_2 \\ n_2-I\left(\frac{k}{2}\right), & n_1+n_2 \leq k \leq 2n_2 \end{cases}$$

$A-B$ ..... $N-k$, $0<k<N$

Third Order

$3A$ ..... 1, $n_1 \leq \frac{k}{3} \leq n_2$ if $\frac{k}{3}$ is an integer

$2A+B$ ....
$$\begin{cases} I\left(\frac{k-n_1}{2}\right)-I\left(\frac{k}{3}\right)+I\left(\frac{k-1}{3}\right)-n_1+1, & 3n_1 \leq k \leq 2n_1+2n_2 \\ I\left(\frac{k-n_1}{2}\right)-I\left(\frac{k}{3}\right)+I\left(\frac{k-1}{3}\right)-I\left(\frac{k-n_2+1}{2}\right)+1, & 2n_1+n_2 \leq k \leq n_1+2n_2 \\ n_2-I\left(\frac{k}{3}\right)+I\left(\frac{k-1}{3}\right)-I\left(\frac{k-n_2+1}{2}\right)+1, & n_1+2n_2 \leq k \leq 3n_2 \end{cases}$$

$2A-B$ ....
$$\begin{cases} I\left(\frac{n_2+k}{2}\right)-n_1+1, & 2n_1-n_2 \leq k \leq n_1-1, \quad k \geq 0 \\ I\left(\frac{n_2-k}{2}\right)-I\left(\frac{n_1+k+1}{2}\right), & n_1 \leq k \leq n_2 \\ n_2+1-I\left(\frac{n_1+k+1}{2}\right), & n_2+1 \leq k \leq 2n_2-n_1 \end{cases}$$

$A-2B$ .... $I\left(\frac{n_2-k}{2}\right)-n_1+1$, $0<k \leq n_2-2n_1$, $n_2 \geq 2n_1$

$A+B+C$ ..
$$\begin{cases} I\left(\frac{k-3n_1+3}{6}\right)+I\left[\frac{(k-3n_1-1)^2}{12}\right], & 3n_1+3 \leq k \leq 2n_1+n_2+1 \\ (N-1)\left[k-2n_1+1-\frac{N}{2}\right]+I\left(\frac{3n_2-k+3}{6}\right)+I\left[\frac{(3n_2-k-1)^2}{12}\right]-I\left[\frac{(k-3n_1)^2}{4}\right] \\ \qquad -\frac{1}{2}I\left(\frac{k-3n_1+2}{3}\right)I\left(\frac{k+3n_1+5}{3}\right)-\frac{1}{2}I\left(\frac{3n_2-k}{3}\right)I\left(\frac{3n_2+k+5}{3}\right), \\ \qquad\qquad 2n_1+n_2+2 \leq k \leq n_1+2n_2-2 \\ I\left(\frac{3n_2-k+3}{6}\right)+I\left[\frac{(3n_2-k-1)^2}{12}\right], & n_1+2n_2-1 \leq k \leq 3n_2-3 \end{cases}$$

$A+B-C$ ..
$$\begin{cases} I\left(\frac{N-n_1+k}{2}\right)I\left(\frac{N-n_1+k+1}{2}\right), & 2n_1-n_2+1 \leq k \leq n_1-1 \\ I\left(\frac{k-n_1}{2}\right)I\left(\frac{k-n_1-1}{2}\right)+(k-n_1)(n_2-k)+I\left(\frac{n_2-k}{2}\right)I\left(\frac{n_2-k-1}{2}\right), & n_1 \leq k \leq n_2 \\ \qquad \text{Note: Number of products included in which } k\text{-channel signal is one fund. component} = \begin{cases} k-n_1, & n_1 \leq k \leq n_1+\frac{N-1}{2} \\ n_2-k, & n_1+\frac{N-1}{2} < k \leq n_2 \end{cases} \\ \left[n_2-I\left(\frac{k+n_1}{2}\right)\right]\left[I\left(\frac{k+n_1}{2}\right)-k+N\right], & n_2<k \leq 2n_2-n_1-1 \end{cases}$$

$A-B-C$ .. $I\left(\frac{N-k-n_1}{2}\right)I\left(\frac{N-k-n_1-1}{2}\right)$, $0<k \leq n_2-2n_1-1$, $n_2 \geq 2n_1+1$

Note: Number of products included in which k-channel signal is one fund. component =
$$\begin{cases} N-2k-1, & n_1 \leq k \leq \frac{n_2-1}{3} \\ k-n_1, & \frac{n_2-1}{3} \leq k \leq \frac{n_2+1}{3} \\ N-2k, & \frac{n_2+1}{3} \leq k \leq \frac{N}{2}, \end{cases}$$

and shown in the table. The number of these is to be subtracted from the total of the $(A + B - C)$- or $(A - B - C)$-types to obtain the number of products not involving the listening channel. It should be pointed out also that kp is the derived carrier frequency throughout and that the products may extend over into adjacent channels. The principal component of the product usually falls in the channel with carrier frequency kp, but in some cases the amount of energy falling in adjacent channels may be quite considerable, as may be seen from Figs. 1 and 2.

The average number of products falling simultaneously in one channel is found by multiplying the total possible number by τ^2 for two-frequency products and τ^3 for three-frequency products, these factors being the probability that any particular product is present. The average number present is not affected by the dependence of the products arising from the fact that one talker may participate in the formation of more than one of the products falling in a channel. For convenience in making use of the results of Table II in evaluating the amplifier requirements, we shall represent the number of x-type products falling in channel number k when it is idle and all other channels are active by the symbol ν_{xk}. We shall also let $\mu(x)$ represent the number of distinct fundamental components required to produce an x-type product, e.g., $\mu(A + B) = 2$, $\mu(2A + B) = 2$, $\mu(A + B - C) = 3$, etc. It follows that the probability that any particular product is present is $\tau^{\mu(x)}$, since τ is the probability that any one required component is present. The average number of x-type products present in the k-channel is therefore $\nu_{xk}\tau^{\mu(x)}$, and may be considered as determined since ν_{xk} is the quantity tabulated in Table II.

7. Modulation Requirement in Terms of Average Total Noise Permissible in a Channel

From Section 3 we have a result for the volume of one product of arbitrary type, averaged on a power basis for a distribution of fundamental talker volumes, referred to the product of the same type produced by zero volume talkers. From Section 6 we have the average number of products of each type appearing in a channel. Combining these two results should give the average total modulation of each type present in a channel. A difficulty occurs however inasmuch as it is not certain how the interfering effect of superimposed modulation adds. The noise caused by one modulation product is of an irregular nature and it is probable that its most disturbing effect is associated with infrequent peak values. When two products are superimposed their individual peaks are not apt to coincide and hence the resultant dis-

turbance may not be much greater than that of one alone. We shall introduce here the concept of "plural S.T.M.F." Suppose ν products of x-type are superimposed and comparison of the resulting noise with one fundamental talker shows that the difference is $L_{x\nu}$ db. If interfering effects add as mean power we should expect $L_{x\nu}$ to be equal to $L_x - 10 \log_{10} \nu$. Hence it seems logical to write

$$s_{x\nu} = L_{x\nu} + 10 \log_{10} \nu - H_x, \tag{7.1}$$

where $s_{x\nu}$ is the "plural S.T.M.F." to be used when ν products are superimposed in order that power addition of products may be valid. Combining (5.1) and (7.1),

$$L_{x\nu} - L_x = s_{x\nu} - s_x - 10 \log_{10} \nu, \tag{7.2}$$

which shows that the correction to be subtracted from power addition is

$$\rho_{x\nu} = s_{x\nu} - s_x. \tag{7.3}$$

The value of $\rho_{x\nu}$ is best determined by experiment. Superposition of a large number of products without using an excessive number of talkers can be accomplished by making phonograph records of individual products and combining their outputs in subsequent recordings.

The average total modulation of x-type in a channel is found by multiplying the average value for one product by the average number of products, and subtracting the quantity $\rho_{x\nu}$, which may be called the "plural S.T.M.F. correction," thus

$$V_x = \eta_x V_{0p} + .115(\lambda_x - \eta_x)\sigma^2 + 10 \log_{10} \nu_{xk} \tau^{\mu(x)} - \rho_{x\nu}. \tag{7.4}$$

where V_x is the volume averaged on a power basis of the x-type modulation in the k-channel referred to the volume of one x-type product from 0-vu talkers. We next wish to express V_x in terms of db above reference noise.

Let T_a represent the "noise" produced by a 0-vu talker in db above reference noise. This is an experimentally determinable quantity and is about 82 db. Let T_x represent the noise from an x-type product from 0-vu talkers. Then L_x, the quantity appearing in (5.1), is given by

$$L_x = T_a - T_x. \tag{7.5}$$

The average total noise produced by all x-type products in db above reference noise is given by

$$W_x = V_x + T_x = V_x + T_a - s_x - H_x. \tag{7.6}$$

If we assume that the total modulation noise allowable for an x-type product is X db above reference noise at zero level, we may equate X to W_x in (7.6) and solve for H_x, giving

$$H_x = V_x + T_a - s_x - X. \tag{7.7}$$

Substituting the value of V_x from (7.4) in (7.7), we get for the system requirement in terms of allowable ratio of fundamental to product when there is one mw. of each fundamental at the point of zero transmission level:

$$H_x = T_a + \eta_x V_{0p} - s_x + .115(\lambda_x - \eta_x)\sigma^2 + 10 \log_{10} \nu_{xk} + 10\mu(x) \log_{10} \tau - \rho_{x\nu} - X. \tag{7.8}$$

For the convenience of the reader, the following recapitulation of significance of the symbols used in (7.8) is given:

H_x = Ratio in db of power of each single frequency fundamental to power of resulting x-type product when each fundamental has power of one mw. at point of zero transmission level.

T_a = Reading of 0-vu talker on noise meter in db above reference noise.

η_x = Order of x-type product.

V_{0p} = Volume in vu corresponding to the average power of the talker volume distribution at the point of zero transmission level $= V_0 + .115\sigma^2$ where V_0 is average talker volume in vu.

σ = standard deviation in db of talker volume distribution.

s_x = Speech-Tone Modulation Factor of x-type product as defined in Section 5.

λ_x = $\sqrt{m_1^2 + m_2^2 + \cdots}$ for $(m_1A \pm m_2B \pm \cdots)$-type product.

ν_{xk} = total number of x-type products which can fall in channel with carrier frequency kp. See Table II.

$\mu(x)$ = number of distinct fundamentals required to produce x-type product.

τ = fraction of busiest hour that a channel is active.

$\rho_{x\nu}$ = correction to be applied to S.T.M.F. when ν products are superimposed. Defined in (7.1)–(7.3).

X = Allowable modulation noise in channel in db above reference noise at zero transmission level point.

The allowable noise may be divided equally between second and third order purely on a power basis by setting the requirement for each 3 db more severe than the total value allowed, or it may turn out that the noise from one order is much more difficult to reduce than that of the other in which the full allowance may be given to the more

difficult one, and the other made to contribute a negligible amount. Usually one type of modulation product will predominate for a given order and the allowable noise for the order may be assigned to this type. If such is not the case division of the noise between the various types may be estimated.

8. Addition of Products in a Multirepeater Line

When a number of amplifiers are used in a carrier system, contributions to interchannel interference occur at each repeater. The considerations previously discussed have a bearing on the modulation requirements on the system as a whole, but it is evident that the individual amplifiers may have to meet much more severe requirements. The relation between total system modulation and single amplifier modulation depends to a considerable extent on the phase angles between products originating in the various repeater sections.

A discussion of the general problem of addition of modulation products from multiple sources is to be given in a forthcoming paper by J. G. Kreer. A point of particular interest in connection with broad band systems is the effect of a linear phase characteristic on the phase shift between modulation products originating in different repeater sections. The curve of phase shift vs. frequency throughout the frequency range occupied by a considerable number of adjacent channels will in general depart but little from a straight line, but the intercept of this straight line if produced to zero frequency is in general not zero or a multiple of 2π. The intercept of such a linear phase curve is effective in producing phase difference between contributions to modulation from successive repeater sections of all the second order products and of some of the third order products, namely the types $3A$, $2A + B$, $B - 2A$, $A + B + C$, and $A - B - C$. The phases of third order products of types $2A - B$ and $A + B - C$ however are unaffected by the value of the intercept and the contributions from the different repeaters of these types of modulation will add in phase to give the maximum possible sum whenever the phase curve is linear throughout the channels involved. Third order modulation requirements on individual repeaters of a system may therefore have to be based on the very severe condition of in-phase or voltage addition of separate contributions.

Experimental verification of in-phase addition of third order modulation products from the repeaters of a 12-channel cable carrier system are included in the paper by Kreer previously mentioned. Corroborating data obtained on an experimental system capable of handling 480 channels are shown in Fig. 3. The measurements there shown were

made [7] on a loop approximately 50 miles in length with repeaters spaced 5 miles apart. The band transmitted extended from 60–2060 kc. Fundamental frequencies of 920 and 840 kc. were supplied from

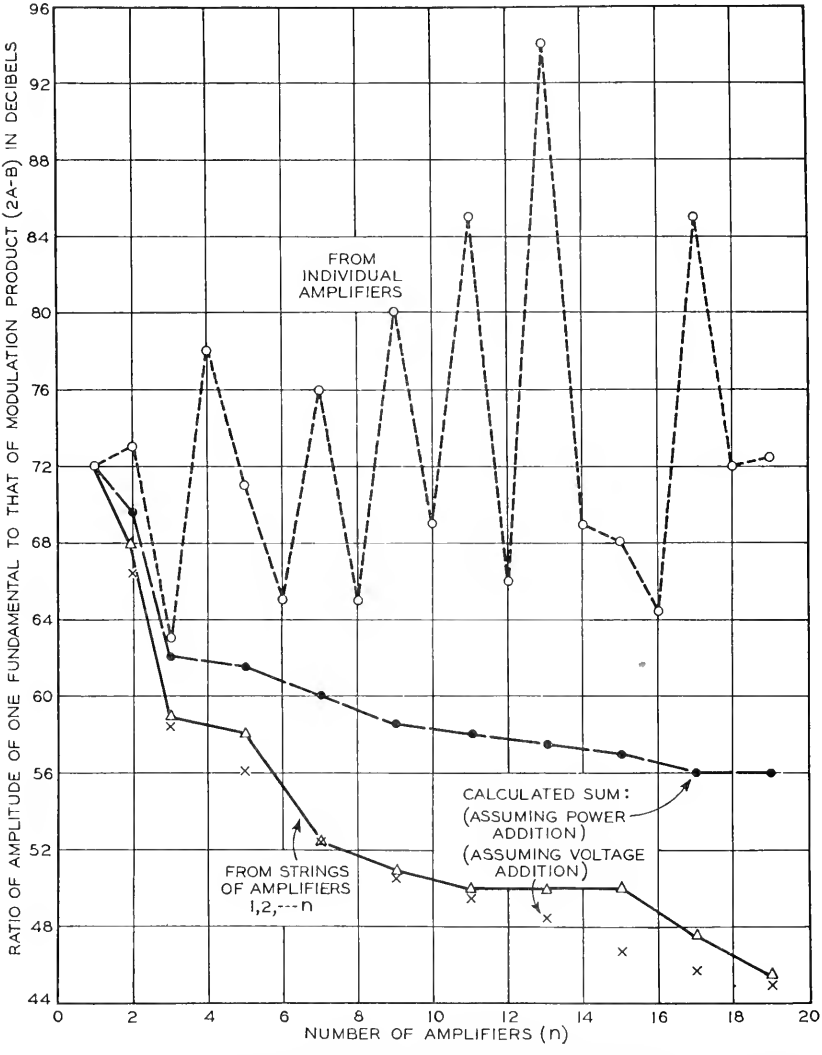


Fig. 3—Experimental data on addition of third order modulation from a multi-repeater line. Fundamental test tones A = 920 kc, B = 840 kc. Modulation product $2A-B$ = 1000 kc.

two oscillators at the sending end, and measurements with a portable current analyzer were made at each repeater to determine the ratio of

[7] Messrs. M. E. Campbell and W. H. Tidd collaborated in these measurements.

amplitude of fundamental to that of the $(2A - B)$-product falling at 1000 kc. A band elimination filter having more than 100 db loss at 1000 kc. and suppressing a band approximately from 940 to 1070 kc. was inserted in the line at the repeater station to remove all contributions to the modulation product originating ahead of the station at which measurements were made. In this way, the modulation contributed by each amplifier and by various combinations of amplifiers could be measured without disturbing the operating levels throughout the system.

The data shown on Fig. 3 include measured modulation from individual amplifiers, and from tandem amplifiers with intervening cable sections. The summation of amplifiers proceeds in the same order as the plotted individual amplifier values. The crosses show the calculated sums of the individual contributions assuming in-phase addition. Agreement between these values and the measured sums is well within the accuracy of the measurements, considering the difficulties involved and the length of time required to complete the run. The dots show the resultant modulation which would be obtained by adding the power in the individual components instead of the voltages, which would be the expected result for a large number of components with random phase angles. The modulation thus calculated is much smaller than the measured values indicating that a hypothesis of random phasing is untenable for this product.

In actual systems both the magnitude and phase shift of modulation products in the different repeater sections exhibit variations because of non-uniform output levels, differences in tubes and other amplifier parts, and unequal repeater spacings. The addition factor for converting system requirements to single amplifier requirements should therefore contain a marginal allowance for these irregularities in performance.

Summarizing our conclusions on addition of modulation from multiple sources, we may state that the third order requirement invokes the most severe condition—that of in-phase addition. Second order products on the other hand will have enough phase shift, either inherent or from simple reversals of terminals at alternate sections, to make the addition no more rapid than on a power basis. In fact if there is a high degree of similarity with respect to both amplitude and phase increment of products from successive amplifiers throughout the system, the total second order modulation may be much less than calculated from addition of power. In setting the requirements which each amplifier must meet, marginal allowances should be made for differences in lineup throughout the system and aging effects which may take place after the amplifier is put in service.

Let A_x represent the ratio expressed in db between the total x-type modulation received from the system and the contribution of x-type from one amplifier, assuming the amplifiers contribute equally. For example, if there are K amplifiers in the system and if the contributions to the product add in phase $A_x = 20 \log_{10} K$. If power addition occurs, $A_x = 10 \log_{10} K$. A favorable set of phase angles may reduce this factor by an amount depending on the uniformity of the repeaters. If the system is divided into K_1 links having K_2 amplifiers in each link, with phase shifts and changes of frequency allocations of individual channels present at the link junctions such that amplitude addition occurs within links and power addition from link to link, $A_x = 10 \log_{10} K_1 + 20 \log_{10} K_2$. We shall also introduce a lineup factor F_x defined as the number of db by which the x-type product requirement must be increased to allow for irregularities in lineup operating levels of the amplifiers. These may be due to initial differences in repeaters or cable sections and to subsequent changes which may occur because of aging effects. If H_{x0} represents the requirement on the ratio of fundamental to x-type product in the output of a single amplifier when one mw. of test tone power is delivered at zero level,

$$H_{x0} = H_x + A_x + F_x, \tag{8.1}$$

where H_x is the system requirement given by (7.8).

9. Testing Methods

It is difficult to test a broad band carrier system under conditions simulating normal operation because of the large number of independent conversations required to load the channels. We have seen that much information applicable to speech load can be deduced from current analyzer measurements of modulation products when discrete frequencies are applied. Since rather extensive calculations are required to evaluate the performance of the system from single frequency data, an overall test under conditions comparable to actual operation has considerable value as a check. A convenient method of simulating the speech load in the high frequency medium by means of a uniformly distributed spectrum of energy such as thermal noise or the output of a gas tube applied through a narrow band elimination filter has been developed [8] for this purpose. The band elimination filter suppresses the energy which would fall in several adjacent channels, hence anything received in these channels at the receiving end of the line is introduced by the system. We can thus measure interchannel modulation if it exceeds the background of noise from other sources. A

[8] E. Peterson, *Bell Laboratories Record*, Nov. 1939, Vol. 18, No. 3, pp. 81–83.

summation is obtained over all types of products but the predominant order may be distinguished by the power law followed. Effects not simulated are the frequency distribution of speech energy within individual channels and the idle periods which occur in various channels during normal operation, but these effects are minor in a system with a large number of channels. The noise loading method is particularly valuable on a multirepeater line, in which modulation measurements made with discrete frequencies may show large variations with frequency because of phasing between the various sources. Loading with a flat band of energy secures an average of these variations over the frequency range used.

Acknowledgments

Contributions to the solution of the various problems discussed here have been made by many members of the Bell Telephone Laboratories. The author wishes to take this opportunity to acknowledge his indebtedness to those of colleagues who have participated in the development of the ideas here discussed.

Other References

Following is a list of published papers (excluding those to which reference has already been made in the text) relating to various aspects of the general problem discussed here:

F. Strecker, "Nichtlineares Nebensprechen bei der gemeinsamen Übertragung mehrerer modulierten Trägerwellen," *Hoch. u. Elek.*, **49** (1937), 5, pp. 165–171.

H. Jacoby and G. Günther, "Über die Wahrscheinlichkeit der in Trägerfrequenz-Vielfachsystemen auftretenden linearen und nichtlinearen Spannungen," *Hoch. u. Elek.*, **52** (1938), 6, pp. 201–209.

E. Hölzler, "Das Nichtlineare Nebensprechen in Mehrfach-Systemen mit Übertragenen Trägern," *Hoch. u. Elek.*, **52** (1938), 4, pp. 137–142.

H. F. Mayer and D. Thierbach, "Über den Einfluss von Nichtlinearität und Wärmerauschen auf die Reichweite von Trägerfrequenz-Vielfach-Systemen," *E. F. D.*, **48** (1938), pp. 6–12.

H. Tischner, "Zur Berechnung und Messung nichtlinearen Verzerrungen in Trägerfrequenten Übertragungs-systemen," *Zeitschr. f. techn. Physik*, **19** (1938), 11, pp. 425–429.

Yonezawa and Hirayama, "Relation Between the Non-Linear Distortion and the Crosstalk on Multiplex Transmission," *Nippon Elec. Comm. Eng.*, June, 1938.

Yonezawa, "Non-Linear Distortion and Crosstalk Frequency Groups in Multiplex Carrier Transmission," *Nippon Elect. Comm. Eng.*, September, 1938.

Matzumae and Yonezawa, "Equipments for Multiplex Carrier Telephony on Ultra Short Wave," *Nippon Elect. Comm. Eng.*, Feb. 1939.

APPENDIX A

Evaluation of Volume Corresponding to Mean Power When Volume Distribution is Normal

The successive steps in the evaluation of (3.2) are as follows:

$$\overline{\text{antilog}_{10} V/10} = \overline{10^{V/10}} = \overline{\exp\left(\frac{V \log_e 10}{10}\right)}$$

$$= \int_{-\infty}^{\infty} \exp\left(\frac{V \log_e 10}{10}\right) p(V) dV$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{V}{10}\log_e 10 - \left(\frac{V-V_0}{2\sigma^2}\right)^2} dV$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\log_e 10}{10}\left(V_0 + \frac{\sigma^2}{20}\log_e 10\right)} \int_{-\infty}^{\infty} e^{-\frac{\left(V - V_0 - \frac{\sigma^2}{10}\log_e 10\right)^2}{2\sigma^2}} dV$$

$$= e^{\frac{\log_e 10}{10}\left(V_0 + \frac{\sigma^2}{20}\log_e 10\right)} = 10^{\frac{1}{10}\left(V_0 + \frac{\sigma^2}{20}\log_e 10\right)}.$$

Hence

$$V_{0p} = 10 \log_{10} \overline{\text{antilog}_{10} \frac{V}{10}} = V_0 + \frac{\sigma^2}{20} \log_e 10.$$

APPENDIX B

Counting of Modulation Products

Consider an N-channel system with carrier frequencies $n_1 p$, $(n_1 + 1)p$, $\cdots$, $n_2 p$, where $n_2 = n_1 + N - 1$. Let q_n represent the frequency of the signal impressed on the voice frequency channel associated with the carrier frequency np. The wave to be amplified is then

$$e_g = \sum_{n=n_1}^{n_2} Q_n \cos\left[(np + q_n)t + \theta_n\right]. \tag{1}$$

The phase angles are of no consequence if the signal frequencies are incommensurable; hence we shall simplify the notation by setting $\theta_n = 0$. The square of the single series in n may be written as a double series in m and n, thus:

$$a_2 e_g^2 = \frac{a_2}{2} \sum_{n=n_1}^{n_2} Q_n^2 + \frac{a_2}{2} \sum_{n=n_1}^{n_2} Q_n^2 \cos (2np + 2q_n)t$$

$$+ a_2 \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m Q_n \cos [(m + n)p + q_m + q_n]t$$

$$+ a_2 \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m Q_n \cos [(m - n)p + q_m - q_n]t. \quad (2)$$

Similarly the term $a_3 e_g^3$ may be written as a triple series in l, m, n as follows:

$$a_3 e_g^3 = \frac{3a_3}{4} \sum_{n=n_1}^{n_2} Q_n^3 \cos (np + q_n)t$$

$$+ \frac{a_3}{4} \sum_{n=n_1}^{n_2} Q_n^3 \cos (3np + 3q_n)t$$

$$+ \frac{3a_3}{2} \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m^2 Q_n \cos (np + q_n)t$$

$$+ \frac{3a_3}{2} \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m Q_n^2 \cos (mp + q_m)t$$

$$+ \frac{3a_3}{4} \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m^2 Q_n \cos [(2m + n)p + 2q_m + q_n]t$$

$$+ \frac{3a_3}{4} \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m^2 Q_n \cos [(2m - n)p + 2q_m - q_n]t$$

$$+ \frac{3a_3}{4} \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m Q_n^2 \cos [(m + 2n)p + q_m + 2q_n]t$$

$$+ \frac{3a_3}{4} \sum_{m=n_1+1}^{n_2} \sum_{n=n_1}^{m-1} Q_m Q_n^2 \cos [(m - 2n)p + q_m - 2q_n]t$$

$$+ \frac{3a_3}{2} \sum_{l=n_1+2}^{n_2} \sum_{m=n_1+1}^{l-1} \sum_{n=n_1}^{m-1} P_l P_m P_n$$

$$\times \{\cos [(l + m + n)p + q_l + q_m + q_n]t$$

$$+ \cos [(l + m - n)p + q_l + q_m - q_n]t$$

$$+ \cos [(l - m + n)p + q_l - q_m + q_n]t$$

$$+ \cos [(l - m - n)p + q_l - q_m - q_n]t\}. \quad (3)$$

It is now a straightforward, though somewhat tedious process, to count the total number of possible products of each type falling in individual channels. The arrangement of terms above is such that $l > m > n$; this forms a convenient way of insuring that no product is counted twice. We shall illustrate by taking a simple case—the second order sum product, or $(A + B)$-type. Let kp represent the carrier frequency of the channel in which we wish to determine the number of possible $(A + B)$-type products. This means that in (2) we wish to find the number of terms in the third summation in which $m = n = k$. The resulting sum becomes:

$$\sum_{m=n_1+1}^{n_2} (n_1 \leq k - m \leq m - 1). \tag{4}$$

That is, there are as many terms as there are integer values of n satisfying the simultaneous inequalities,

$$\left\{ \begin{array}{c} n_1 + 1 \leq m \leq n_2 \\ \dfrac{k+1}{2} \leq m \leq k - n_1 \end{array} \right\} \tag{5}$$

The number of terms is zero if $k > 2n_2 - 1$, because the lower limit of the second inequality exceeds the upper limit of the first. If $n_1 + n_2 \leq k \leq 2n_2 - 1$, the upper limit of the first inequality and the lower limit of the second inequality are governing, and the number of terms is $n_2 - I(k/2)$, where $I(x)$ is a symbolic representation for the largest integer $\leq x$. If $2n_1 \leq k \leq n_1 + n_2$, the second inequality is governing and the number of terms is $I\left(\dfrac{k+1}{2}\right) - n_1$. If $k \leq 2n_1$, the number of terms is zero.

In a similar manner the more complicated sums representing the number of third order products can be evaluated. It is to be noted that contributions to a particular type can come from more than one of the sums listed. For example, the $(A + B - C)$-type is made up of the summations from the $l + m - n$, $l - m + n$, and $l - m - n$ terms. In fact all these are of $(A + B - C)$-type except those in which $l - m - n$ is negative. The latter, since only positive values of frequency are significant, are of $(A - B - C)$-type. An $(A - B - C)$-type product differs from $(A + B - C)$-type not only in S.T.M.F., but also in manner of addition of contributions from a multi-repeater line as discussed in Section 8.

As an alternative to an actual count of the products falling in a channel, it is possible to approximate the sum by an integration

process when the number of channels is large. This is an especially valuable simplification for products of high order for which the counting becomes very tedious. Suppose there are D components in unit band width in the range a to b. Let $x_1, x_2, \cdots x_n$ be n frequencies such that $a \leq x_1 < x_2 < \cdots < x_n \leq b$. The number of products of form $m_1x_1 + m_2x_2 + \cdots + m_nx_n$ which can be formed by selecting fundamentals from the n bands of width dx_1 at x_1, dx_2 at x_2, $\cdots dx_n$ at x_n is $D^n dx_1 dx_2 \cdots dx_n$. To count the total number of such products which can be formed in the band a to be such that the resultant frequency lies in the band x_0 to x, we form the integral

$$g(x, x_0) = D^n \int_a^b dx_n \int_a^{x_n} dx_{n-1} \cdots \int_a^{x_2} \varphi(x, x_0, x_1, x_2 \cdots x_n) dx_1, \quad (6)$$

where

$$\varphi(x, x_0, x_1, x_2 \cdots x_n) = \begin{pmatrix} 1, & x_0 \leq m_1x_1 + m_2x_2 + \cdots + m_nx_n \leq x \\ 0, & \text{otherwise} \end{pmatrix}. \quad (7)$$

A suitable representation of φ is furnished by

$$\varphi(x, x_0, x_1, x_2 \cdots x_n) = \frac{1}{2\pi i} \int_C \left(e^{-ix_0 z} - e^{ixz} \right) e^{tz \sum_{r=1}^{n} m_r x_r} \frac{dz}{z}, \quad (8)$$

where C is a contour going from $z = -\infty$ to $z = +\infty$ and coinciding with the real axis except for a downward indentation at the origin. To obtain the number of products $\nu(x)dx$ falling in band of width dx at x, we write

$$\nu(x) = \operatorname*{Lim}_{\Delta x = 0} \frac{g(x, x - \Delta x)}{\Delta x}$$

$$= \frac{D^n}{2\pi} \int_C e^{-ixz} dz \int_a^b dx_n \int_a^{x_n} dx_{n-1} \cdots \int_a^{x_2} e^{tz \sum_{r=1}^{n} m_r x_r} dx_1. \quad (9)$$

If the spacing between carrier frequencies is used as the unit band width, we may set $D = 1, a = n_1, b = n_2, x = k$, in (9) and obtain the limiting forms of the results given in Table II as the number of channels is made large. Evaluation of (9) is easily carried out by means of the relation:

$$\int_C \frac{e^{ixz} dz}{z^m} = \begin{pmatrix} \dfrac{2\pi i^m x^{m-1}}{(m-1)!}, & x \geq 0 \\ 0, & x < 0 \end{pmatrix}. \quad (10)$$

As an example, suppose it is desired to calculate the approximate number of $(2A + B)$-type products falling in channel number k when the number of channels is large. If $x_2 > x_1$, this type of product may be either of form $2x_1 + x_2$ or $2x_2 + x_1$. Both are included by the expression

$$\begin{aligned}
\nu_{2A+B}(k) &= \frac{1}{2\pi}\int_C e^{-kz}dz \int_{n_1}^{n_2} dx_2 \int_{n_1}^{x_2} \left[e^{iz(2x_1+x_2)} + e^{iz(x_1+2x_2)} \right] dx_1 \\
&= \frac{D^n}{4\pi}\int_C \frac{e^{-ikz}}{i^2z^2} \left[e^{2in_2z} - e^{i(2n_1+n_2)z} - e^{i(n_1+2n_2)z} + e^{3in_1z} \right] dz \\
&= \begin{cases} 0, & k < 3n_1 \\ \frac{1}{2}(k - 3n_1), & 3n_1 < k < 2n_1 + n_2 \\ \frac{1}{2}(n_2 - n_1), & 2n_1 + n_2 < k < n_1 + 2n_2 \\ \frac{1}{3}(3n_2 - k), & n_1 + 2n_2 < k < 3n_2 \\ 0, & k > 3n_2. \end{cases} \qquad (11)
\end{aligned}$$

The method may be generalized to include the calculation of the modulation spectra produced by fundamentals having specified arbitrary spectra by inserting the appropriate function of the power in unit band width at $x_1, x_2, \cdots x_n$ in the integrand of (6).

Radio Extension Links to the Telephone System

By R. A. HEISING

TO the general public, the word *radio* means broadcasting, and a *radio set* means a radio receiver for listening thereto. The average man never has any direct contact with a radio transmitter, nor with the radio telegraph which preceded the radio telephone and so utilizes the all inclusive word "radio" for that one part of it which he sees, buys, and uses.

The radio engineer is not quite in that class. There is, however, much in radio with which he is unacquainted. The radio field has become so broad and extensive that it is physically impossble for anyone to keep abreast of the whole art. Since the application of radio to telephone links is a specialized field, undoubtedly much of its history and technical developments is known only sketchily or not at all to many engineers. This paper, therefore, is planned to cover this field briefly, show its general development, and describe in principle a number of devices developed for use therein, most of which are seldom if ever used in the field of broadcasting.

The radio telephone was not one of those devices that an inventor springs upon an unexpecting world. On the contrary it was expected, and was the object of search and investigation for years before a practical form appeared. Because the wire telephone followed the wire telegraph, technical men expected the radio telephone to follow the radio telegraph as soon as the latter had been practically demonstrated. Telephone men developed an interest in it as soon as it was suggested. Telephony over large bodies of water, over difficult terrain, and to moving conveyances was difficult or impossible for wire telephony, and the telephone man was intrigued by the possibility of providing his circuits without the use of conducting wires.

During the first few years of this century, several radio telephone systems were technically demonstrated but were found impractical. In 1912 an important step occurred. The audion, invented by DeForest, was brought to the attention of the American Telephone and Telegraph Company. It appeared to have possibilities making it superior to the mechanical and the arc repeaters for wire telephone lines. A telephone repeater, or amplifier, was a main object of search at that time by telephone men. The audion became the subject of study in the Research Department of the American Telephone and

Telegraph Company and Western Electric Company immediately, and within a year and a half went through some rapid transformations. A high vacuum and increased electron emission were provided by H. D. Arnold and A. M. Nicolson, while a practical circuital theory was provided by H. J. Van der Bijl. The internal arrangement was engineered and a socket and base developed. This improved vacuum tube was put into use on the commercial telephone lines in the latter half of 1913 as a telephone amplifier and was the first commercial use of the high vacuum tube. This vacuum tube amplifier contributed to the establishment of the original transcontinental wire telephone line which carried its first messages in July 1914.

The improved vacuum tube, during its period of development, appeared to have possibilities as a generator of sustained oscillations and suggested to telephone engineers that it might be much more useful in radio than it had been up to that time. With this in mind the A. T. & T. Company decided to start work in that direction and as one result a number of new engineers, including the writer, were employed and began work in the Research Department of the Western Electric Company in the middle of 1914. Developments on the radio telephone moved rapidly. Early in 1915 plans were made and active work was started for field trials. A transmitting station was established at Montauk, L. I., and a receiver located on Hotel DuPont in Wilmington, Delaware. On April 4, 1915, speech was transmitted from Montauk to Wilmington, a distance of 220 miles. Connections were made with telephone lines at both ends to show its possibilities as a link in a telephone circuit.

There followed tests to Jekyl Island, off the coast of Georgia, about 800 miles, and then work was started for a transoceanic test. To transmit across the ocean required more power and a larger antenna. In order to avoid the antenna expense, arrangements were made with the Navy Department to use the Arlington antenna for transmitting, and to use Naval radio stations at San Francisco, San Diego, Panama and Honolulu for receiving locations. Observers with radio receivers were dispatched to these four receiving locations while a fifth expedition was sent to Paris where in spite of the war the French Government kindly allowed listening on the Eiffel Tower antenna. At Arlington the Western Electric Research Department (now part of Bell Telephone Laboratories, Inc.) installed a vacuum tube transmitter, and proceeded to make one-way tests. In August 1915 speech was understood at Panama, and in September a one-way demonstration was made across the continent, receiving at San Francisco. Within a few days speech was heard in Honolulu and then in Paris. The tests

showed that transoceanic telephony was possible and indicated some of the difficulties that had to be overcome.

The radio transmitter in these tests deserves a few words because of its novelty and because in one respect it has never been equaled. The carrier was modulated at a relatively low level and then amplified. The final stage of amplification contained 550 tubes in parallel which in number appears to be an all time record. Each tube was capable of delivering only 15 or 20 watts peak h.f. power which would give a power rating on a telephone basis of about $2\frac{1}{2}$ kw.

With these tests completed, transoceanic telephony withdrew into the laboratory for almost eight years while further intensive work was carried out. The second step in public occurred in January 1923

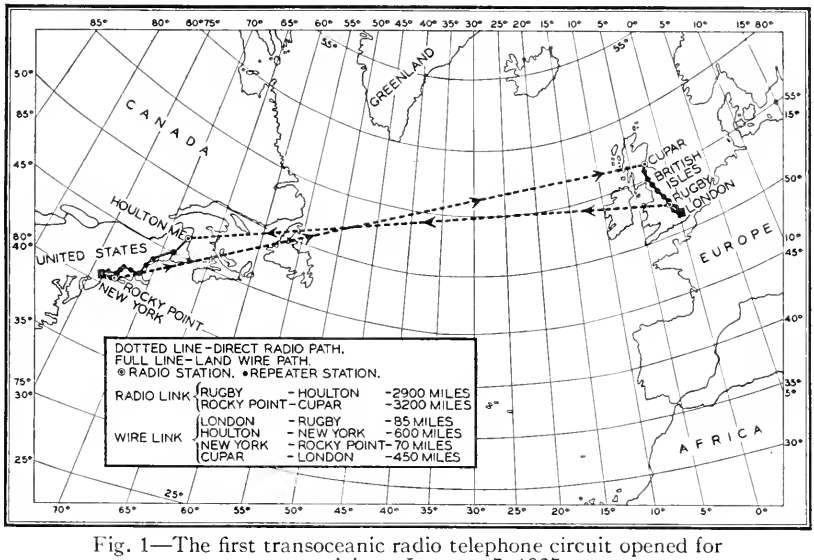


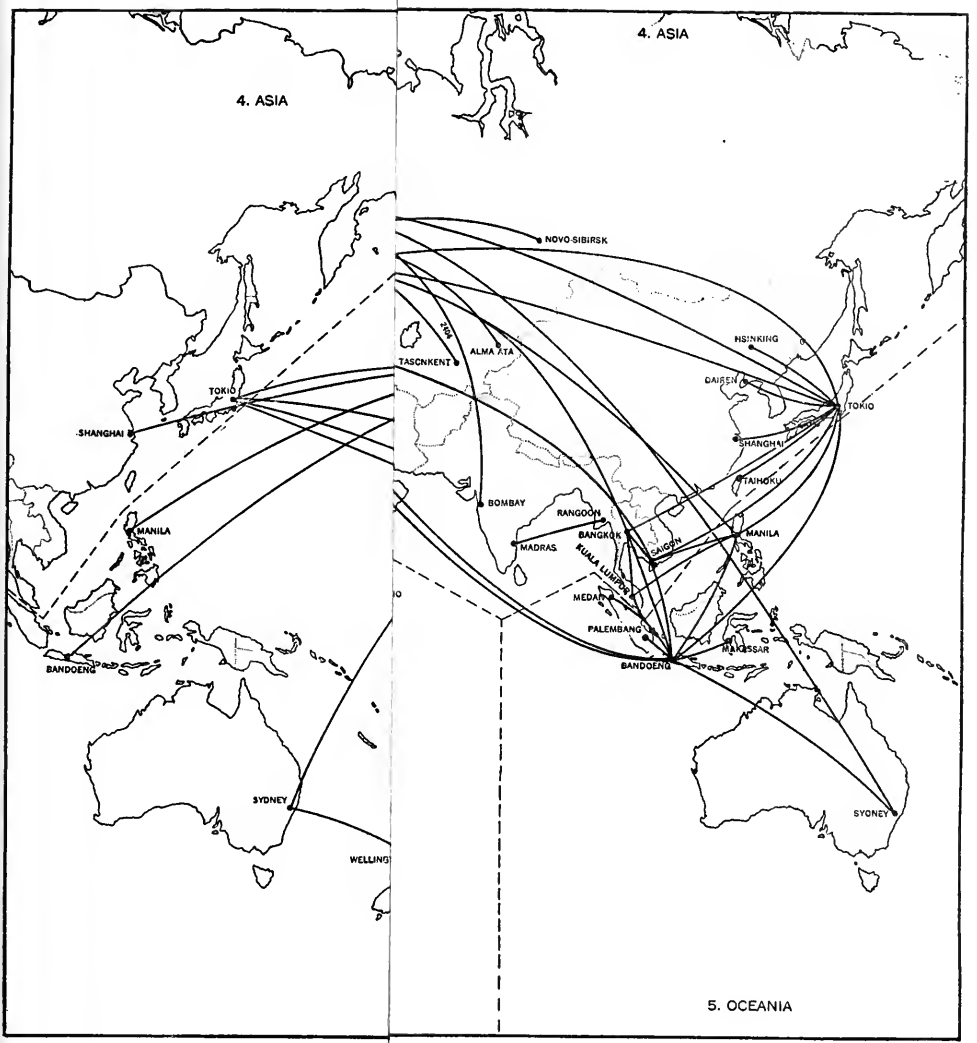
Fig. 1—The first transoceanic radio telephone circuit opened for commercial use January 7, 1927.

when a second transoceanic test was made. A 200 kw. single sideband transmitter had been constructed and installed at Rocky Point, Long Island, while engineers with receiving equipment journeyed to London. A demonstration was given to government engineers and to newspaper reporters over there to show that practical transoceanic telephony was possible and to interest them in constructing a return circuit. The British government was interested and with our assistance took up the matter of providing a transmitter and receiver for their end. Three years were required for this third step and in February 1926 the first two-way radio telephone conversations were held between the United States and England. Commercial service opened in January 1927. See Fig. 1.

With the first transoceanic circuit established, further circuits followed rapidly. For a number of years prior to 1927 investigations had shown that short waves could travel enormous distances with very much less attenuation than the long waves. Telephone engineers conducted a series of long distance tests which laid the foundation for developing circuits on these short waves with much less power. As a result a short wave transoceanic telephone circuit, the first of its kind, was opened on June 6, 1928, between the United States and England. The Germans followed by establishing a circuit to Buenos Aires in December of that year. The Dutch established one to Bandoeng in January of the following year, 1929. Then another circuit was opened up between the United States and London in June of 1929. The circuit from Madrid to Buenos Aires was established that same year, and there followed very rapidly circuits from London to various British Colonies, a circuit from New York to Bermuda, and one from San Francisco to Honolulu, so that as of Jan. 1, 1939 the world was covered by a multitude of circuits as indicated in Fig. 2.* However, the radio circuits of greatest interest to us are those circuits extending from this continent to other continents. These appear in Fig. 2. There are several channels to London, one to Paris (temporarily suspended), one to Rome, one to Australia (temporarily suspended), one to Berlin, one to Switzerland, two channels to Honolulu and a number of circuits to South and Central American countries, circuits to Manila, Bandoeng, Tokyo and Shanghai. There is a circuit from Montreal to London operated by the Canadian Marconi Company and British Post Office. The facilities are now such that from almost any telephone in the United States it is physically possible to talk to almost any telephone in the rest of the world, although due to censorship some of the circuits are not actually in use.

Another use for radio as a link in telephone service is for providing service where wire circuits would be unusually expensive and difficult to maintain. Such a circuit indicated in Fig. 2 runs from Seattle to Juneau, Alaska. It is operated by the Signal Corps and connects with the general telephone system at Seattle. Another circuit is shown in Fig. 3 which runs from Green Harbor to Provincetown, Massachusetts, a distance of 24 miles across Cape Cod Bay. This circuit supplements the wire lines which reach Provincetown by a roundabout path by land around the south side of the Bay. The radio link provides a very desirable alternate route to points on the Cape and has been useful in a number of emergencies in maintaining

* Explanation.—On account of unsettled conditions, has not been brought up-to-date.


4. ASIA
4. ASIA
NOVO-SIBIRSK
TASCHKENT
ALMA ATA
HSINKING
DAIREN
TOKIO
SHANGHAI
SHANGHAI
TAIHOKU
BOMBAY
RANGOON
BANGKOK
SAIGON
MANILA
MADRAS
KUALA LUMPUR
MEDAN
PALEMBANG
MAKASSAR
BANDOENG
BANDOENG
SYDNEY
SYDNEY
5. OCEANIA

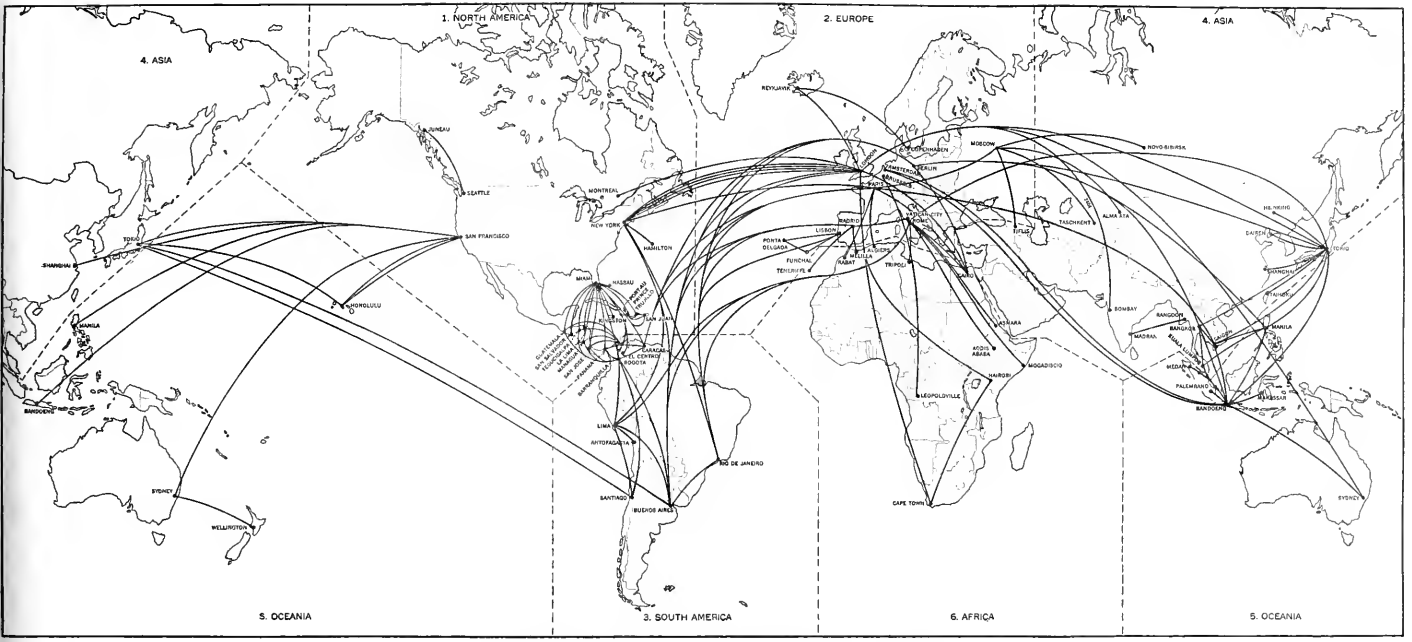


Fig. 2—Transoceanic radio telephone circuits in operation January 1, 1939.

service. During the hurricane of 1938 it provided the only route to Cape Cod for a time. The Provincetown radio link is different from any of the transoceanic links mentioned previously in that it operates in a third region of the radio spectrum known as the ultra-short-wave region while the transoceanic circuits are in the short-wave and long-wave regions. This circuit operates on 63 and 65 megacycles, 4.75 and 4.61 meters, respectively.

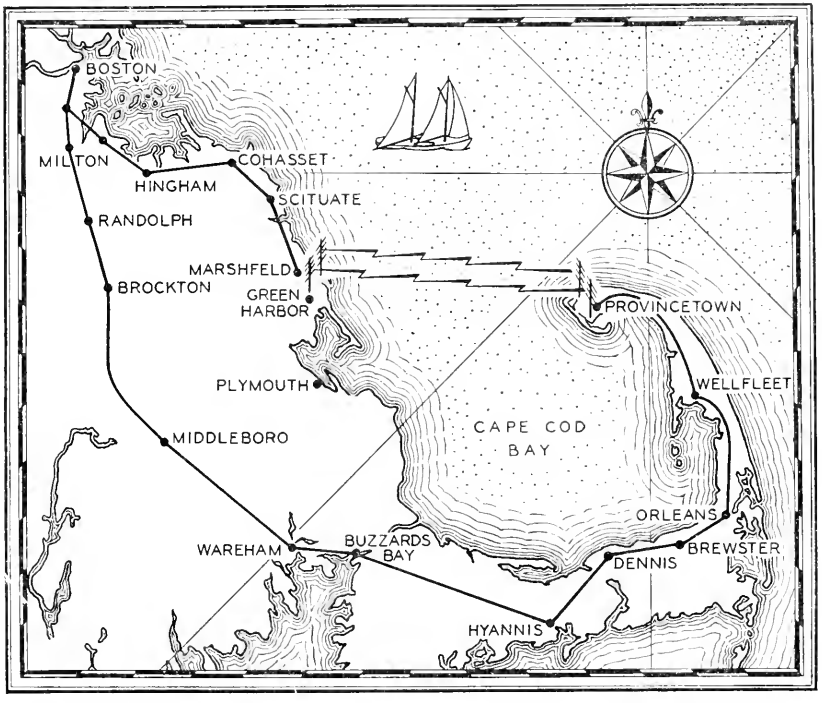


Fig. 3—Radio telephone circuit from Provincetown to Green Harbor, Massachusetts, connecting with the telephone line to Boston. Land wire route between Provincetown and Boston is also shown.

Transoceanic and other point-to-point services were not the only services envisaged by radio engineers prior to the era of broadcasting. Service to ships was considered an important use for radio. The first commercial telephone service to ships was the service established with transatlantic liners in December of 1929. Most of the larger transoceanic liners are now equipped for radio telephone communication with both shores. A few years after this service was established, a ship radio telephone service of a more local character was initiated to serve fishing fleets off the New England coast. The radio station

service. During the hurricane of 1938 it provided the only route to Cape Cod for a time. The Provincetown radio link is different from any of the transoceanic links mentioned previously in that it operates in a third region of the radio spectrum known as the ultra-short-wave region while the transoceanic circuits are in the short-wave and long-wave regions. This circuit operates on 63 and 65 megacycles, 4.75 and 4.61 meters, respectively.

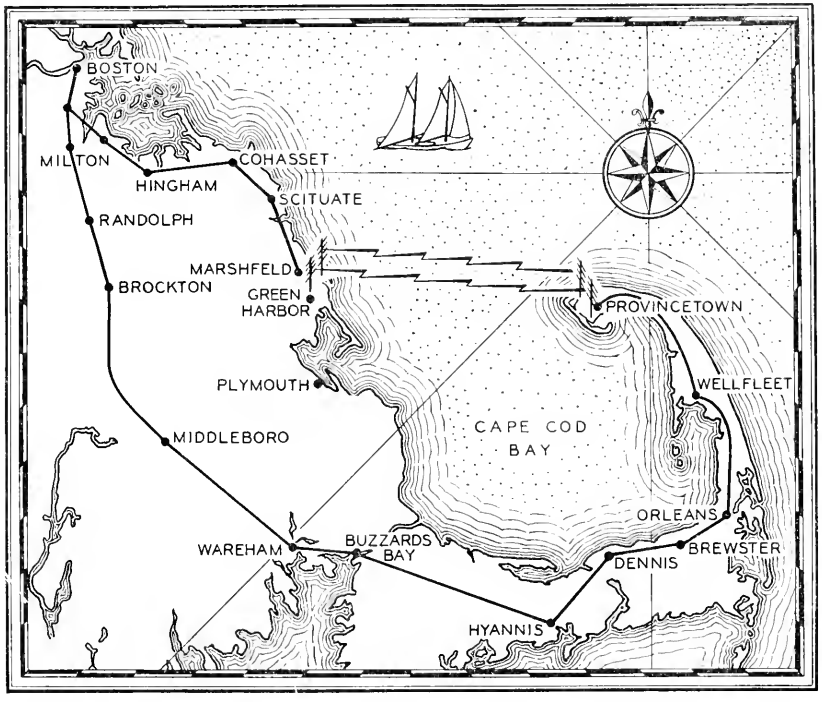


Fig. 3—Radio telephone circuit from Provincetown to Green Harbor, Massachusetts, connecting with the telephone line to Boston. Land wire route between Provincetown and Boston is also shown.

Transoceanic and other point-to-point services were not the only services envisaged by radio engineers prior to the era of broadcasting. Service to ships was considered an important use for radio. The first commercial telephone service to ships was the service established with transatlantic liners in December of 1929. Most of the larger transoceanic liners are now equipped for radio telephone communication with both shores. A few years after this service was established, a ship radio telephone service of a more local character was initiated to serve fishing fleets off the New England coast. The radio station

for this service was established near Boston. The necessary equipment for boats was also developed. The fishing boats by means of this service can keep in touch with the fish markets and can take advantage of rises in prices. They also find it convenient for communicating with each other when schools of fish are found and that has also been a help in their operations. There have been a number of occasions also in which it has resulted in the saving of lives at sea, as the radio was used to notify the shore station in case of accident and the shore station called other vessels and sent them to the rescue. This service to fishing vessels was then extended to other coastwise vessels, yachts, tugs, etc., so that there has gradually developed an extensive radio service of this type on both of our coasts. Radio stations are located not only in Boston now but as indicated in Fig. 4 there are stations at New York, Ocean Gate, N. J., Wilmington, Del., Norfolk, Charleston, S. C., Miami, New Orleans, Galveston, Los Angeles (San Pedro), San Francisco and Seattle. Stations are under construction at Tampa, Fla., Astoria and Portland, Oregon. Service is now given to more than 2,000 vessels, there being 200 tugs, 1,100 yachts, 100 steamships, 400 fishing vessels and numerous others, police boats, pilot boats, barges, launches, etc. The largest number of vessels so equipped for communication with shore are grouped around New York and San Pedro, there being about 600 in each of these areas.

In this type of service each shore transmitter and each shore receiver is assigned a frequency. Any ship may provide itself with frequency control crystal elements for communicating with as many of the shore stations as it desires. Coastwise vessels in traveling along the coast may thus keep in touch with their nearest shore station. In New York there are two such circuits provided with transmitters located on Staten Island, and there are receivers located at four places around the harbor for each of the circuits so that the low-powered ship transmitters may reach the nearest receiver while the higher-powered shore transmitters reach the ship receivers directly.

On the United States side of the Great Lakes, connecting telephone companies operate coastal harbor radio telephone stations at Lorain, Ohio, Duluth, Port Washington, Wis., Lake Bluff, Ill. and Mackinac Island.

The use of radio in the telephone system brings forth a number of problems. First of all, to provide a radio circuit for a telephone conversation there are required two radio transmitters and two radio receivers. The transmitters and receivers must be so located and so designed and operated that one person at one end of the circuit may speak to a second person at the other end and the second one to the

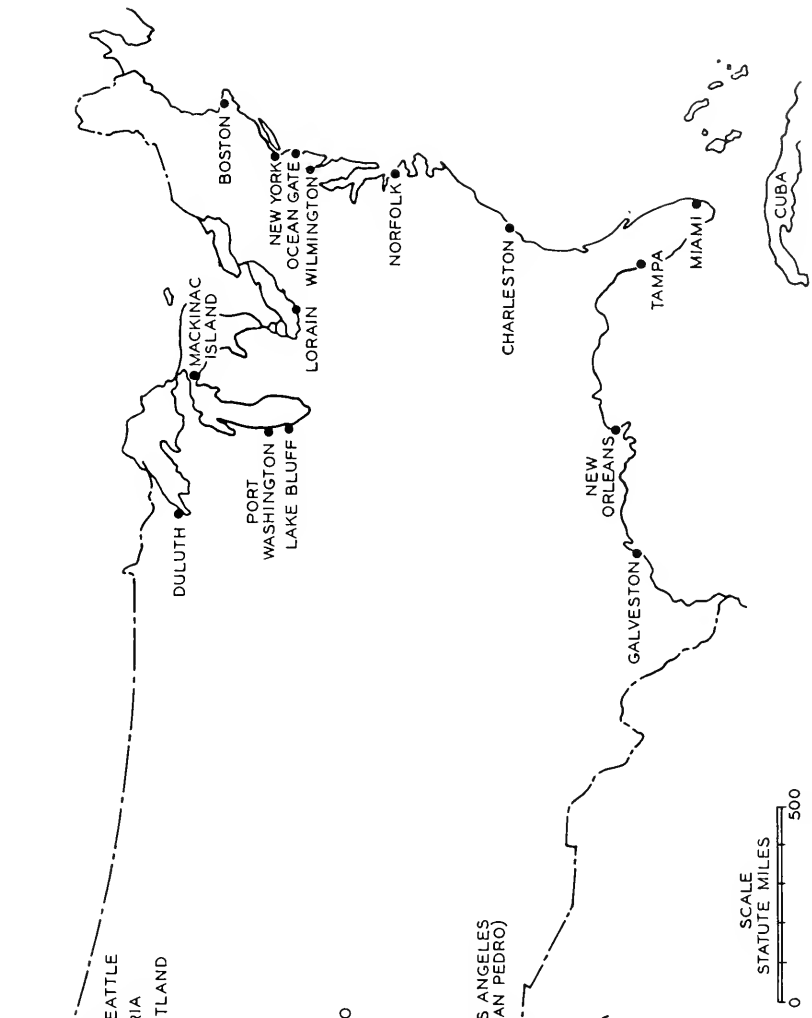


Fig. 4—Coastwise boat radio service, and other radio links.

first even though either or both of the radio equipments be entirely outside of the manual control of the speakers and be many miles away. The problems involved have given rise to the development of many pieces of apparatus which are seldom used in the broadcasting field. It is the intention, therefore, in this paper to review some of these devices and tell briefly why they are used and what they do.

To begin with, attention is called to three diagrams in Fig. 5. These three diagrams indicate three of the many ways in which a transmitter

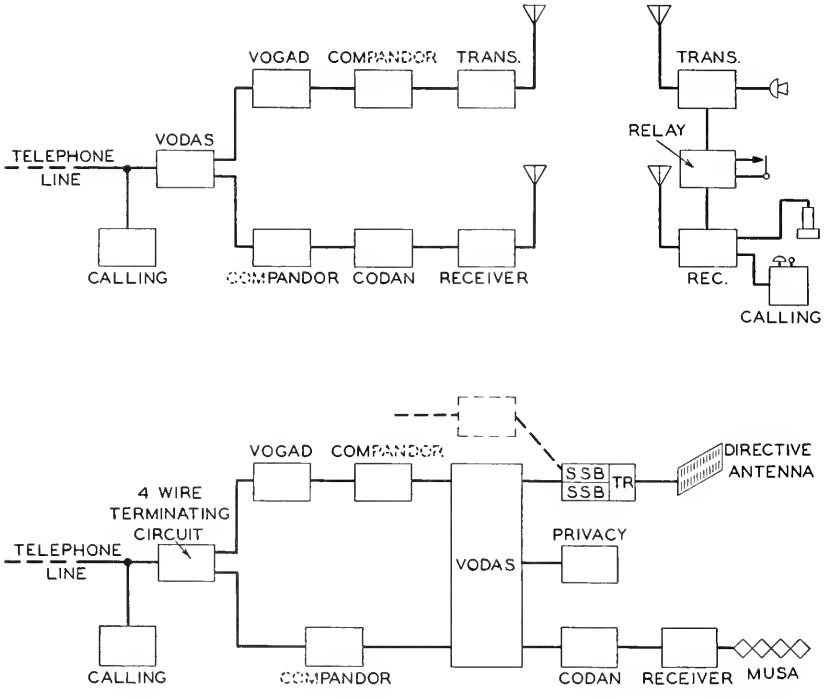


Fig. 5—Three arrangements of radio terminal apparatus are shown herein.

and a receiver might be connected with associated apparatus for one end of a radio link. Two of the diagrams indicate connection to a telephone line which may extend more than 3,000 miles to a subscriber. The third circuit indicates an arrangement which is customary on small boats, sometimes in aircraft, and other places where a partially trained operator is available and is the one using the device.

It is to be observed first of all that in two of these diagrams the input to the transmitter and the output of the receiver are connected to the same telephone line. This necessary connection leads to difficulties. Signals from a subscriber on the telephone line operate the

radio transmitter. Some of the radiated energy from the transmitter impinges on the receiver. If the receiver should be tuned to the same wave-length the signals will then get back onto the telephone line and some will again go to the transmitter, producing by this circular path a singing circuit. A circuit so constructed will be entirely useless due to the singing produced. Now it appears possible to use a hybrid coil to connect a balancing circuit to the telephone line, with conjugate connections to the transmitter and to the receiver so that the incoming signals from the receiver will not go to the transmitter. Such a hybrid circuit will work provided it can be balanced and maintained in balance. However, anyone who has tried balancing such circuits knows that it is generally not practicable to provide a balanced circuit suitable for all wire line connections and for the variable gains in the radio link. Additional means must therefore be used. Now, of course, it is possible to operate the incoming radio circuit and, therefore, the receiver on a different frequency as is usually done, in which case the signals from the local transmitter will be tuned out. However, if a similar system is used at the other end of the radio link the signals from the near end transmitter will come in on the far end receiver, will again go out on the far end transmitter, will come back into the near end receiver whence they get back into the near end transmitter, thereby making a loop circuit again which will produce singing even though the round trip path of such a circuit may be 6,000 miles. It is therefore found necessary, when connecting with telephone lines, to provide a system which will at all times keep the incoming energy of the receiver from going out on one's own transmitter.

To accomplish the foregoing is the function of the "Vodas" [1] as indicated on the diagram, a device which connects the telephone line to either the transmitter or the receiver but not to both simultaneously. It must, however, connect them at proper and suitable times so that a two-way conversation can take place. A simple system comes immediately to mind to accomplish this purpose. It is that of a voice-operated relay which throws the telephone line from the receiver to the transmitter whenever the speaker on that end speaks, with the relay making the reverse connection when he stops speaking. Such a simple circuit has been used in some cases but has been found not to be adequate for general use. To begin with, the line is not switched until part of what the speaker has said has arrived to actuate the relay. Some clipping, therefore, occurs. To make things worse many words begin with sounds of small energy like f's and s's, which may not be sufficient to actuate the relay. The relay will then not operate until the vowel sound following arrives and when the relay does operate the

entire preceding consonant is clipped off. The clipping is sometimes disconcerting and may impair the intelligibility of the transmitted speech. If the relay is made sensitive enough to operate on the f's and s's another difficulty arises. Relatively low values of room noise, noise induced on wire lines, or the speaker breathing into the microphone may produce enough energy to actuate the relay during listening periods, thus interrupting the conversation. In other words, if the relay is sensitive only to the loud sounds clipping will occur, while if sensitive to the weaker sounds it may be actuated by noise. These difficulties are surmounted by using the more complex circuit termed the "Vodas" as indicated in Fig. 6.

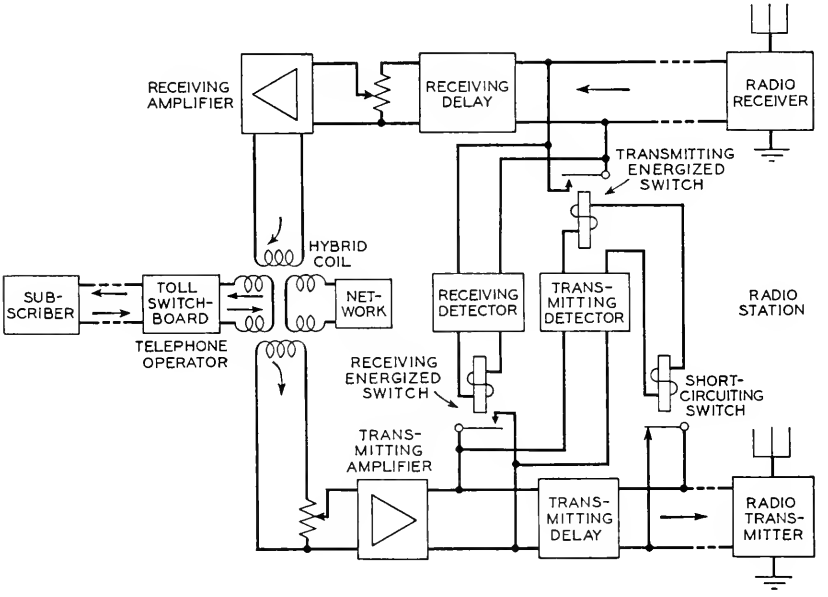


Fig. 6—Vodas (voice operated device anti-singing).

In this diagram the transmitter is located on the bottom branch in which there is interposed a transmitting delay circuit. The incoming speech after passing through the transmitting amplifier operates relays by means of the transmitting detector. The relay is so adjusted as to be operated by the louder sounds in the voice but not by noise. While it is operating upon the vowel sound following a consonant, the consonant may be on its way through the delay circuit so that the relay which normally shuts off signals to the transmitter will actually clear the path to the transmitter in time for the first sounds of a word in most cases. The utilization of the delay circuit can therefore practi-

cally eliminate the clipping and allow of relays being adjusted so as not to be operated by small noises from the telephone line.

This circuit also includes arrangements to prevent other difficulties. It will be observed there are two sets of relays, one operated by the transmitting branch rectifier and one by a receiving branch rectifier. These are so arranged that when speech signals operate the transmitting branch rectifier, the receiving line is short-circuited to prevent any signals, such as noise, from the radio receiver reaching the talker or from going out on the transmitting branch to interfere with the transmitted speech. When no talking is occurring at this end of the circuit signals coming in on the radio receiver operate a receiving relay which short-circuits the transmitter circuit so that the received speech will not be retransmitted by the transmitter and so set up a singing condition. This particular diagram indicates the hybrid coil and balancing network which are used to assist in operation but not to provide the main means for preventing the received speech from reaching the transmitter.

This circuit as indicated is about as simple as a satisfactory circuit can be made. Figure 7 indicates a more complex circuit which has a number of advantages, among which is that of connecting in privacy equipment. This circuit allows of using one piece of privacy equipment which is used in the transmitter branch for outgoing signals and is switched to the receiver branch for incoming signals.

Returning now to Fig. 5, attention is called to another device in the first diagram labeled "Vogad."[2] This word comes from the initial letters of the words "voice operated gain adjusting device."[2] This is a type of device which is very useful in telephone practice but is seldom, if ever, used with a broadcasting transmitter. Every telephone user is cognizant of the fact that different people with whom he speaks over the telephone use different intensities of voice; also different lengths of telephone line introduce different amounts of attenuation. If the incoming telephone signals are to operate the radio transmitter to its full modulation capacity some means must be provided to equalize these signals of various levels. It is therefore desirable to have a device which will maintain the output level nearly constant regardless of the variation in the input level due to different speakers and different lengths of telephone lines. This operation is provided by the Vogad indicated in Fig. 8. The channel across the top is the direct path of the speech signals. Within the dotted lines marked "vario repeater" are some elements including the amplifier whose gain is varied to make up for variation in intensities at the input. It is not sufficient to construct an amplifier which will give a large gain

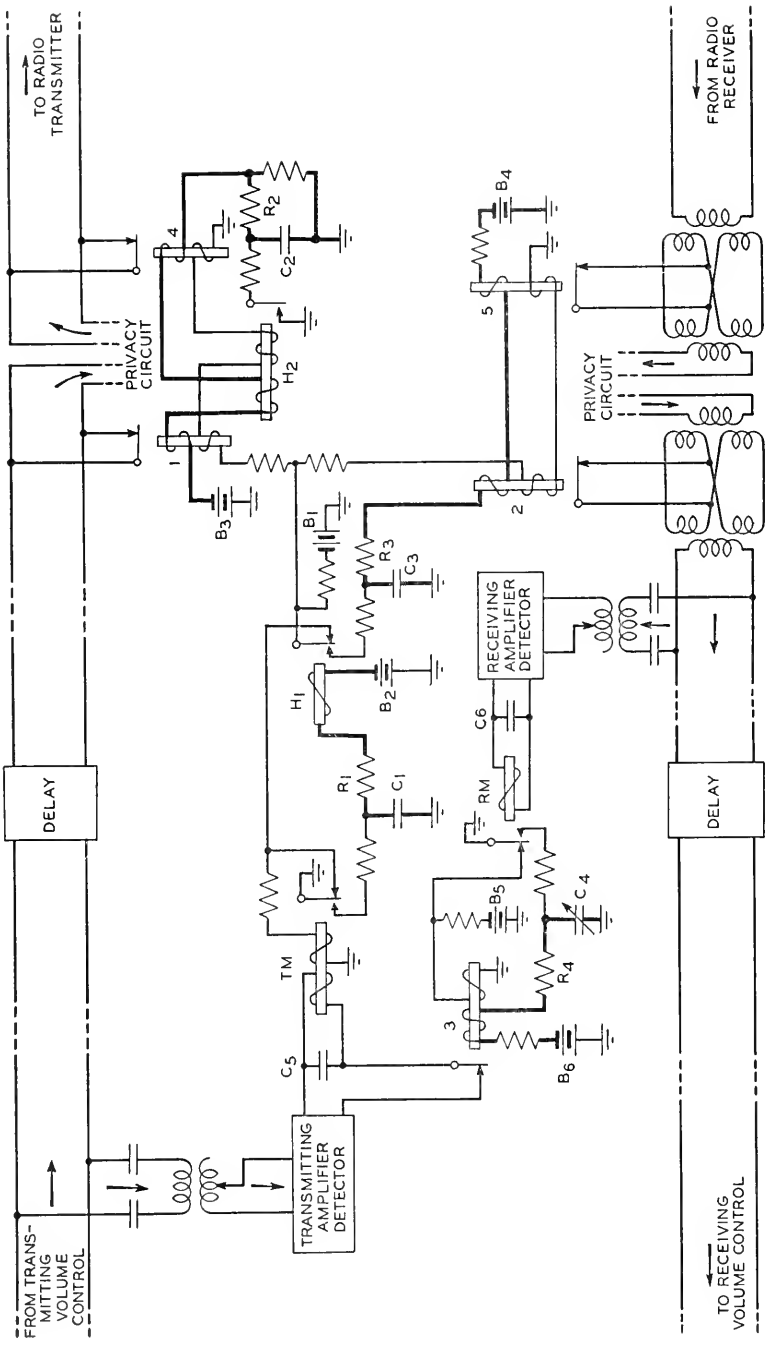


Fig. 7—Improved vodas circuit.

with a small input and small gain with a large input as the use of as simple a circuit as this will cause the amplifier to adjust itself for maximum gain when no input signal is coming in and in that case any noise on the line may be amplified sufficiently to be troublesome. Also, with such an amplifier the gain will become larger whenever the speaker stops or hesitates and will momentarily overload the transmitter with the first syllable on resumption which may result in dis-

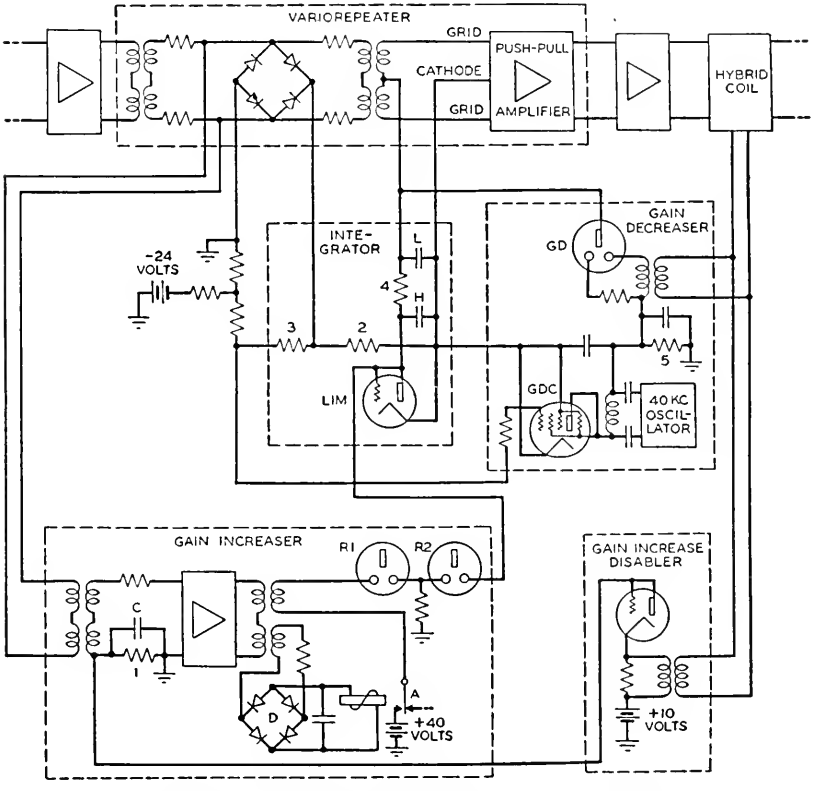


Fig. 8—Vogad (voice operated gain adjusting device).

tortion of considerable consequence. It is therefore desired that this circuit have a maximum gain which will not make noise troublesome and also be so constructed that the gain during any conversation will remain constant even during hesitation and listening periods, and in which the gain will increase only when speech signals become weaker or the gain decrease when speech signals becoming stronger. To accomplish this one element of the circuit marked "Gain Increaser" receives energy from the input and operates through the gas tubes

R1 and R2 causing the gain integrator to increase the gain of the main amplifier when the speech input levels decrease. However, this increase in gain must continue only until sufficient speech volume is going out to modulate the transmitter in an approximately satisfactory manner. At this instant another part of the circuit called the "Gain Increaser Disabler" operated by the output signal comes into play and disables the gain increaser. If the input signal and, therefore, the output signal become louder, then a fourth element, called the "Gain Decreaser," comes into play and begins to reduce the gain of the amplifier. The combination of these control circuits with the main amplifier therefore causes the volume of output signal to be reasonably constant with wide variations in the volume of the input signal and at the same time to hold the gain substantially constant as long as speech signals of the same volume are coming or while no speech signals are coming.

Returning to Fig. 5, note a situation which may be troublesome. Anyone who has operated the earliest broadcast receivers with automatic gain control will remember that when the incoming signal became weak or disappeared the gain of the receiver climbed to such a point as to produce disconcerting noise in the loud speaker. The receivers in all of the systems indicated in this figure contain automatic volume control and if the transmitter from the remote station stops momentarily or if the signals fade out this automatic volume control boosts the gain to such an extent that noise is delivered by the receiver to the telephone line. Such noise may be sufficient to operate the Vodas and in doing so will seize control of the circuit and not allow the signals from the subscriber at this end to reach the transmitter. This would lock up the circuit and put an end to the conversation.

Such a contingency is avoided by the use of the device indicated in the block marked "Codan."[3] The word Codan comes from the initials of the words "carrier operated device anti-noise." The Codan is a device which is operated by the carrier picked up by the receiver and connects the receiver to the telephone line only while a carrier is present. Under these conditions the volume control will go up and down as the carrier goes down and up but if the carrier disappears the Codan disconnects the receiver so that noise will not operate the Vodas and prevent the speech from the subscriber at the near end from being transmitted.

Specifically, circuits operating with ships at sea must have the Codan or its equivalent because the ships usually employ a system in which the carrier is cut off when the ship stops speaking so that the disappearance of the carrier at the receiving station on shore is the

period during which the shore subscriber is expected to talk. The noise at this time in the receiver on shore must not actuate the Vodas or the transmitting branch will be interrupted. A suitable rectifier with a relay can, under certain conditions, be a satisfactory Codan but not in all cases. There is always a certain amount of static, strays and so forth, reaching an antenna and if the relay is adjusted to operate on a very weak carrier the strays and the static may also operate the relay. Now, of course, it is possible to adjust the relay so it will not operate on the noise occurring at a particular time but will operate on a carrier which is slightly stronger. If that adjustment is made during the day on short waves when the noise is low, when night comes the noise level rises and the noise may then be able to operate the relay. It would be necessary with a simple Codan of this type to have the operator continually adjust and readjust the Codan for different parts of the day and night.

However, it is impracticable for an operator to be continuously on watch and continuously and satisfactorily adjust the sensitivity of such a relay, with the consequence that a satisfactory Codan necessarily involves automatic adjustment as provided in the circuit of Fig. 9. In this diagram the part above the middle dividing line is the receiver while the part below is the Codan. The Codan here consists of two parts. It consists of a part which selects the carrier by a crystal filter for operating the relay, and instead of a spring to hold back the armature an electrical arrangement is provided whereby the noise coming through the second part of the circuit will produce current in the relay in the opposite direction. The noise is picked out by another crystal filter Y4 which selects all the energy in the two sideband positions minus the carrier position. This noise is amplified and rectified so that whenever the noise is high it will require a large carrier coming through filter Y3 to operate the Codan relay S2 and when the noise is low through the noise branch a smaller carrier coming through the carrier branch can operate the Codan relay. This Codan therefore automatically adjusts itself to the noise level in the ether so that the carrier can connect the receiver to the telephone line whenever the carrier appears.

Since the development of a successful Codan it has been found practical to dispense with the Vodas at terminals that connect with radio stations which radiate their carrier during transmitting periods only. This is brought about by using the Codan to operate the relay that switches from transmitting to receiving. Since the Codan is operated by the incoming carrier and not by outgoing voice signals,

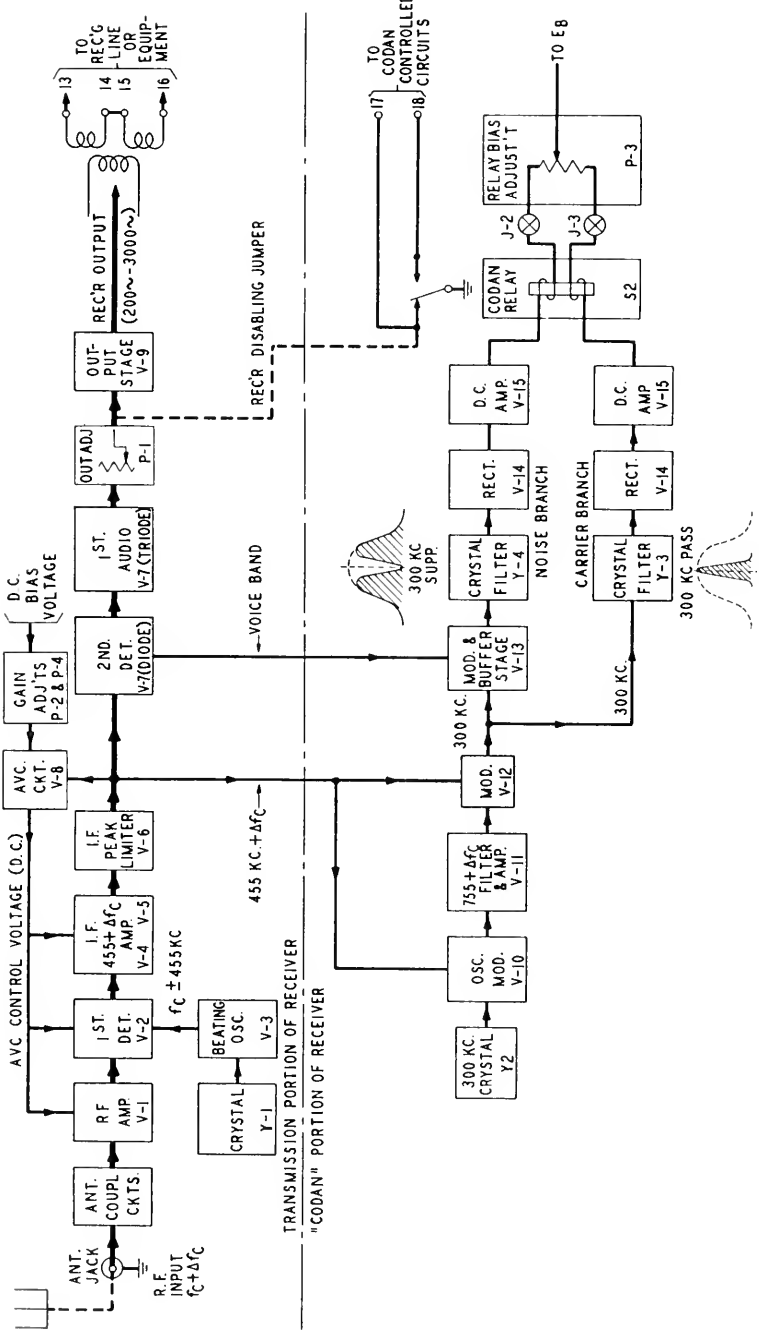


Fig. 9—Codan (carrier operated device anti-noise).

the delay circuits of the Vodas are unnecessary. This arrangement is finding increased application in ship-shore terminals.

Referring again to Fig. 5, note two squares in the first diagram labeled "Compandor." [4] Each of these squares has part of the name dotted to indicate that the two circuits are different but together form the entire Compandor. The Compandor is another device to assist in making signals more intelligible in the presence of noise at the receiver. It accomplishes this by the peculiar method of distorting the signal going out and then restoring it at the receiver. The reason for such a device and its mode of operation are as follows. Ordinary speech contains loud as well as weak signals. Most of the consonants and some of the vowels do not contain much energy. They therefore

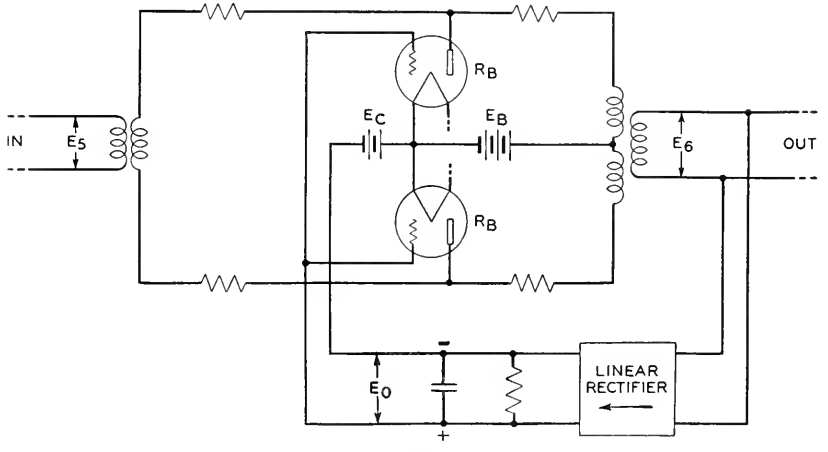


Fig. 10—Compressor part of the compandor.

will not modulate the transmitter fully and are the ones whose reception will be interfered with by noise at the receiver. The Compandor reduces this effect by making the weak parts of the transmitted signal larger than normal.

The part of the Compandor in the transmitter branch distorts the speech signal by reducing the energy variations between the loud and weak sounds a certain amount. It does not wipe out all variation as it is necessary to leave a certain variation which is made use of at the receiver to restore the original variation. The circuit used at the transmitter end is given in Fig. 10. It is called the Compressor. A speech signal comes in on the left-hand side and goes out on the right. Between the input and output circuits are connected two vacuum tubes. Although superficially this circuit looks as though these tubes are amplifiers, actually they are connected to absorb

energy. The operation therefore involves absorbing part of the energy in the louder signals and a lesser amount from the weaker signals so that the output contains speech which has been distorted in such a fashion that the variations in energy may be only one-tenth as much as they originally were. These two absorbing vacuum tubes are controlled by potential built up across a circuit containing capacitance and resistance. This circuit has a potential E_0 produced on it by a linear rectifier which secures its energy from the output circuit. The strong signals appearing in the output produce a larger voltage on the resistance-condenser combination, thereby causing the grids of the two vacuum tubes to be more positive than with weak signals or no signals, and the two tubes, then acting as conductive resistances, reduce the intensity of the loud signals. The resistance-condenser combination is so proportioned that the charge on the condenser rises and falls with syllabic frequency. It must not have such a short time constant as to wipe out individual cycles. It is to operate upon groups of cycles only. With this Compressor between the telephone line and the transmitter, and the amplifiers properly adjusted, the transmitter can still be fully modulated with the louder sounds in the voice but it will be modulated very much more than it would normally be by the weaker sounds in the voice.

At the receiving end the signal delivered by the receiver and transmitted towards the telephone line will be the same distorted signal which modulated the transmitter. Such a distorted signal, although scarcely discernible from the original, is not in all situations the desirable one to put upon a telephone line, so there are reasons for restoring this distorted signal to its original form. This distorted signal contains the weaker parts of speech amplified many times with respect to what would occur without the Compandor and therefore these weaker parts of speech will be many times above the noise which would have interfered with reception under ordinary conditions. This distorted speech now goes into the part of the Compandor in the receiving branch which is called the "Expandor," as shown in Fig. 11. The Expandor contains many elements similar to those in the Compressor but they are arranged in a slightly different form. Two vacuum tubes instead of absorbing energy are now used as amplifiers. The signal comes in on the left and goes out on the right but the output is not a true amplified picture of the input because as the signal goes through the amplifier the amplification of the two tubes is varied so as to restore the original signal. To do this use is made of the remaining amplitude variations within the signal to operate a linear rectifier and put a variable voltage E_x upon similar condenser

and resistance, and by virtue of connections to vary the amplification of the two amplifier tubes. The louder signals therefore put more positive bias on the grids of the amplifiers than the weaker ones and the amplifier tubes will amplify the stronger tones more than the weaker. There is thus delivered to the output the original signal but with very much improved signal-to-noise ratios.

In the use of the Compandor on circuits having noise it has been found possible to produce signal-to-noise improvements as high as 30 db. Average improvements are 15 to 20 db. The improvement depends upon the amount of noise present.

Returning again to Fig. 5, there are certain elements labeled "Calling." [3] The particular configuration indicated with calling devices attached to the telephone line in the first diagram and to a receiver

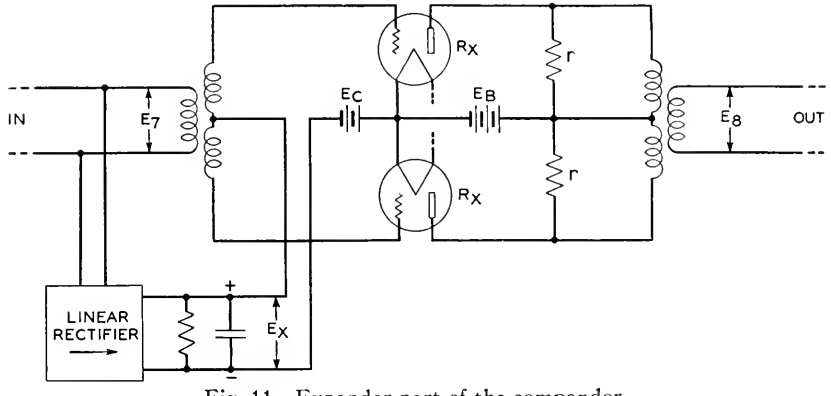


Fig. 11—Expandor part of the compandor.

in the second diagram are the particular calling arrangements used for communicating with boats at sea. The fishing boats, such as mentioned previously, and private yachts, do not find it expedient to have an operator listening at all times for calls nor do they like to have a loud speaker operating continuously delivering all and sundry communications to the people on a boat. It is desirable that means be provided so that the boat may be called by having a calling mechanism available. The system is indicated in Fig. 12. In this system the receiver on the boat must be operated continuously and must be connected to the selector and bell circuit so that whenever the correct calling signal comes in the bell will ring. At the shore station the calling is accomplished by sending out certain combinations of 600 cycles and 1500 cycles as indicated in this figure, the various combinations being chosen by the telephone dial which actuates a relay switching one or the other audio frequency onto the transmitter.

A certain combination which consists of certain sequences of the 600 and 1500-cycle tones is assigned to each boat. The 600 and 1500-cycle tones selected by band-pass filters are rectified and actuate a polar relay which then delivers to the part marked "selector" signals corresponding to those made by the telephone dial. The selector is a standard train dispatcher selector which can be set for various combinations of signals and when the correct combination arrives it will close a switch and ring the bell.

For calling the shore operator from the ship, the Codan mentioned previously connects the receiver to the line and also operates a relay to light the shore operator's switchboard light whenever a boat operator starts his transmitter and puts on his carrier.

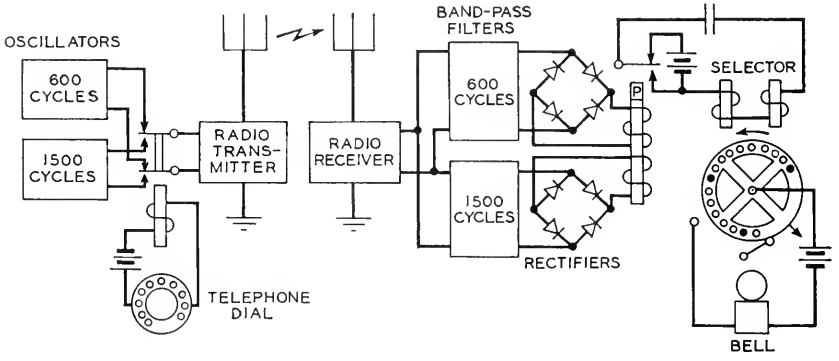


Fig. 12—Calling system for boats.

These two methods of calling in the two directions are not the only ones which are used on radio links. For most transoceanic and service to the large liners, prearranged schedules or continuous watch allow of calling by voice. In the Green Harbor-Provincetown circuit of Fig. 3 and some transoceanic circuits calling is accomplished by transmitting 1000 cycles interrupted 20 times a second which is a standard means of ringing over telephone lines.

Referring again to Fig. 5, note the third diagram and the element marked "SSBTR" [5] which means single sideband transmitter. Single sideband transmission has been used on transoceanic radio telephone circuits since the first circuit was opened. It is not used in broadcasting, at least in this country, although it has been proposed a number of times. Single sideband communication was first used on wire line carrier circuits with their inception in 1918. It was utilized when the long-wave radio circuit was tested experimentally in 1923

and opened in 1927. Single sideband has since been applied on short wave circuits to London and to Honolulu.

Single sideband has the theoretical advantage of 9 db in signal-to-noise ratio over double sideband with carrier transmission; 6 db is secured from the utilization of all of the energy in one sideband and 3 db comes from reduced noise in the reduced band width of the receiver. Tests have shown that the 9 db signal-to-noise improvement is secured in practice.

The application of single sideband to the short-wave circuits encountered a number of difficulties. One of the bigger difficulties was that of resupplying the eliminated carrier. In order that speech received over single sideband circuits be truly normal and be recognizable as the voice of the talker, the carrier must be resupplied within 20 cycles. If it is supplied more than 20 cycles out of position the speech will be intelligible in varying degrees but it is impossible to recognize the voice even of one's best friend. To resupply the carrier within 20 cycles when the radio frequency is 20 MC means that the transmitter and the beating oscillator at the receiver individually should not vary more than 10 cycles, and 10 cycles out of 20 MC is one part in 2 million. The frequency of either oscillator must therefore remain constant to better than one part in 2 million if voices are to be recognized. It is of course possible to build oscillators which are more stable than this. However, such oscillators at present appear to be in the laboratory rather than the commercial class and so it has been found desirable to adopt a different means for maintaining this frequency of the resupplied carrier at the right value. This is accomplished by transmitting a small part of the original carrier and then at the receiver use this small or vestigial carrier as it is known to actuate a mechanism which will supply a local frequency exactly in synchronism with it. The resupplied carrier can therefore be maintained well within one part in 2 million, in fact it can be maintained within one cycle in 20 megacycles.

In producing a single sideband other difficulties are encountered. In the present state of the art it is difficult to eliminate one sideband and leave the other, except at relatively low frequencies. In the long-wave transoceanic circuits which operate on 60 to 70 kc the elimination occurs at 30 kc and a second modulation shifts the remaining sideband to the desired position. This gives good selection of the desired sideband and provides flexibility in the final positioning of the sideband.

In operating at high frequencies, which may be as high as 22 MC, it has been found desirable to reach the desired point not with two steps in modulation but with three steps. This is indicated in Fig. 13.

Signals come in on input circuit A to modulator 1A which modulates 125 kc from which crystal filter A selects the upper sideband. This upper sideband then goes into modulator 2 where it modulates 2500 kc. The sideband in this case will be located more than 125 kc away from the carrier and can be selected by a relatively inexpensive filter which now delivers the single sideband in position 2625.1 to 2631 kc to modulator 3. In modulator 3 this sideband modulates a suitable

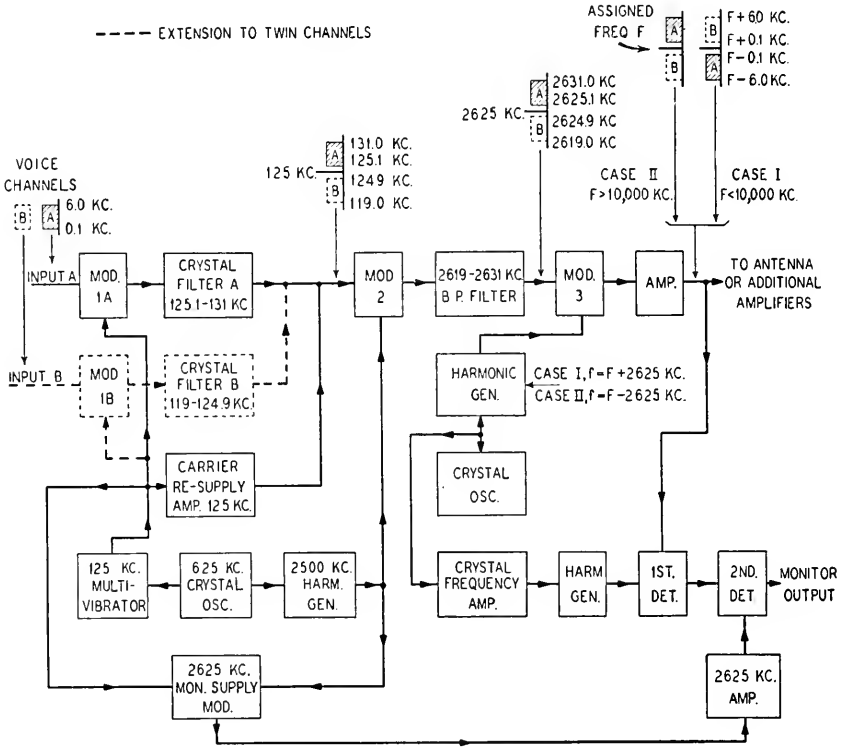


Fig. 13—Single sideband transmitter circuit.

frequency which may be anywhere from 6 MC to 20 MC which is secured from a suitable source such as the harmonic output of a crystal oscillator. The output of this third modulation places the single sideband in a desired position in the ether whereupon it passes through an amplifier to an antenna and is radiated. This desired position in the ether may be either above or below the final carrier frequency as may be found desirable, depending upon the frequency range in which operation is to occur at that particular hour.

Following the success in applying single sideband to one of the short-wave channels consideration has been given to utilizing the position of the vacated sideband or the contiguous position for a second channel. The resultant method is called "twin channel single sideband." Referring to the diagram in the upper right-hand corner of Fig. 13, Channel *A* is the single sideband which has just been discussed while the second single sideband is now placed in *B* position so as to give two separate conversations with the same transmitter. This immediately brings to mind the question, "Is the advantage of single side band lost by using the two sideband positions for two separate channels?" Odd as it may seem it is only slightly affected. If those two sideband positions are used for a single channel the frequencies in the two sidebands appear simultaneously, and in corresponding positions. The transmitter must handle both simultaneously and the receiver must be broad enough to receive both bands. However, when two separate conversations are placed in the two separate sideband positions similar frequencies do not appear simultaneously in both bands except at such remotely occasional times that their mutual interference is small or not noticeable, and at the same time the receiver for each channel is tuned for only one sideband, thereby keeping down the noise. By using the two positions for two separate channels it is possible to get on a statistical basis two single sideband circuits each 8 db better in signal-to-noise ratio than would be obtained using the same two sideband positions for one channel alone. It produces a remarkable increase in efficiency of use of a circuit.

Now it also happens that addition of the second channel requires a surprisingly small amount of apparatus. Looking at this same figure, Channel *B* is indicated on the left providing input *B* to modulator 1*B*. This modulator modulates the same frequency as modulator 1*A* but crystal filter *B* selects the lower sideband in this case, which sideband is now delivered to modulator 2 along with that from filter *A*. These two parts are all the apparatus necessary to add to this transmitter to convert it from one channel to two channels. It is thus to be observed that by suitable application of single sideband to the short-wave channels it has been possible to multiply their number by two and increase each one 8 db in signal-to-noise ratio. This twin channel single sideband has been applied to two of the three short-wave transoceanic circuits, to the San Francisco-Honolulu circuit, and undoubtedly will be applied to other circuits in the future.

Referring again to Fig. 5, there will be seen an element marked "Directive Antenna." Directive antennas have been used very little in broadcasting. They are coming into greater use with short-wave

broadcasting and with efforts to produce less interference between stations in the United States operating on similar wave-lengths, but their use has been much smaller than their use in radio links of the telephone system. Directive antennas are of great importance in telephone links for the reason that in operating over great distance, where weak signals must be received all or most of the time, much power may be saved if directive antennas at the transmitter are used

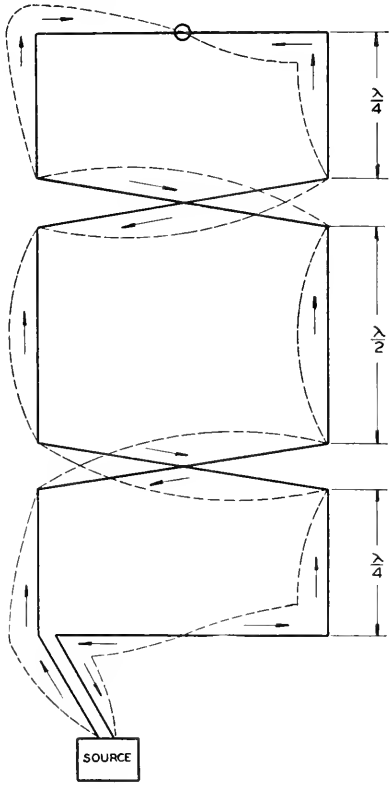


Fig. 14—Element of the Sterba directive antenna.

to send as much energy as possible in the desired direction, and used at the receiver to exclude as much noise from undesired directions as can be done.

In connection with the directive antennas used in the telephone system, quite a variety have been developed. Figure 14 indicates the principle of what is known as the Sterba [6] antenna, invented by the late Mr. E. J. Sterba. This is an elemental section, the complete antenna being built up of a number of elements of this kind. The

element consists of a conductor, whose length is an integral number of half wave-lengths, bent in such a way that the currents in all of the vertical elements will be in phase while in the horizontal elements they will largely balance out. The vertical elements are all placed in the same plane and the horizontal elements depart from that plane only enough for crossing without short-circuiting. With this arrangement the energy radiated in a direction perpendicular to the paper will be a maximum. It will be a minimum in the plane of the paper in side or vertical directions. Any suitable number of these elements may be arranged in the same plane and connected together so as to increase the energy in the desired direction. Inasmuch as the length of the wire and the configuration into which it is bent are associated with the frequency, an antenna constructed for one frequency is not usable at another. This is true in any antenna where standing waves exist.

In the antennas of this type which have been used on our short-wave circuits for operating across the Atlantic, as many as 8 elements were connected in parallel in the same plane so as to radiate in the desired direction producing a sharp directivity pattern. At the same time another set of 8 elements were located one-half wave-length away parallel to the first as indicated in Fig. 15 so as to eliminate the radiation in one of the two directions perpendicular to the screen. This causes all the energy to be radiated in the desired direction. For operating one radio transmitter on a transoceanic circuit it is necessary to have three or four antennas for each transmitter, one for each wave-length. These antennas were strung between towers. Figure 16 shows the antennas as used formerly at Lawrenceville, New Jersey. One antenna occupied two inter-tower spacings. The direction of transmission is perpendicular to the line of the towers.

At the receiving end directive antennas are also used. An elemental picture is shown in Fig. 17 of the receiving antenna devised for this purpose by Mr. E. Bruce.[7] The third diagram shows the shape into which a long conductor is bent and the arrows show the instantaneous directions of current flow at a moment of maximum. It is observed here also that in all the vertical elements the currents flow in the same direction while in the horizontal elements enough flows in the two directions so the effect is neutralized. This antenna will therefore also receive or transmit in the directions perpendicular to the plane of the conductor and not in directions within the plane. This type of antenna can also be constructed with a reflector behind it to reduce the direction of transmission or reception to a single direction.

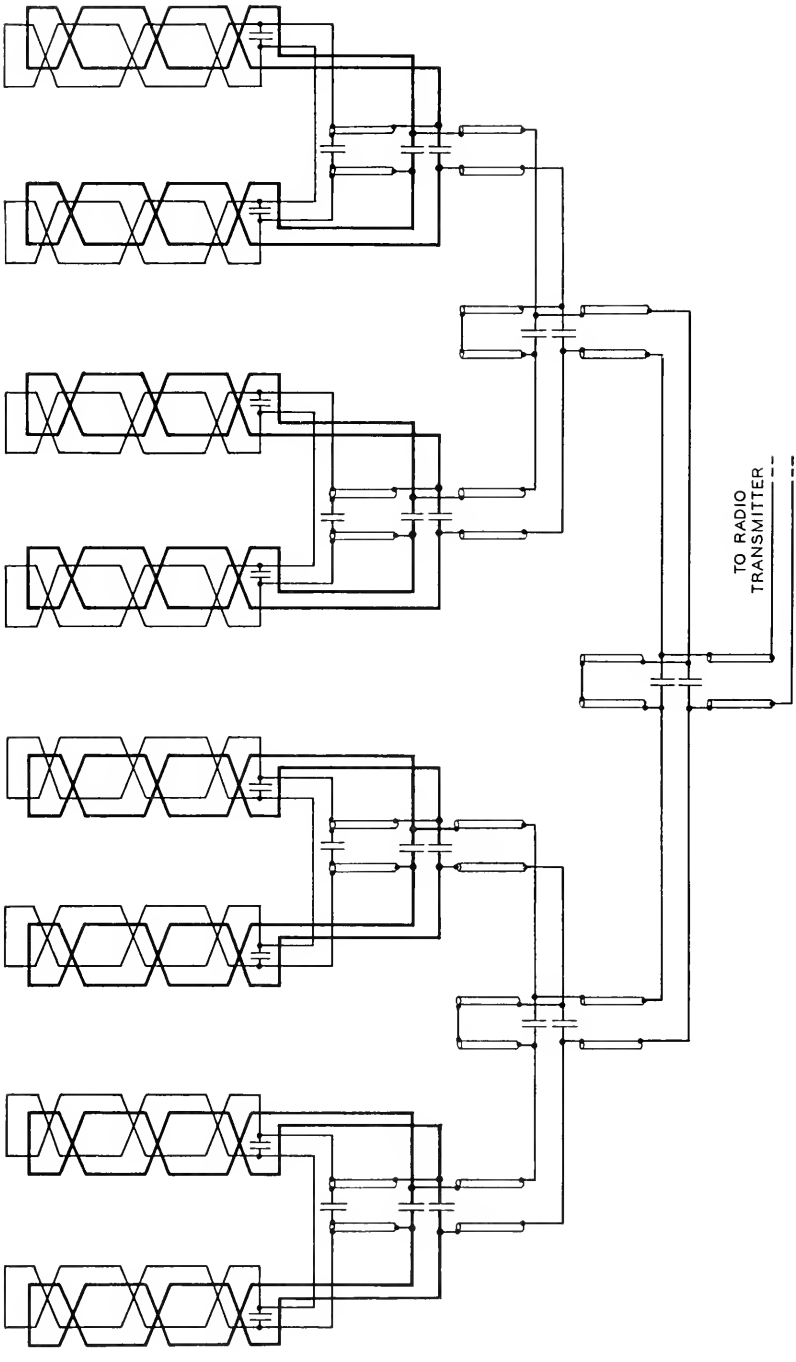


Fig. 15—A complete Sterba antenna with reflectors for one wave-length.

Fig. 16—Towers at Lawrenceville, New Jersey, used to support a number of Sterba antennas for transoceanic communication. These antennas have since been superseded by rhombic antennas.

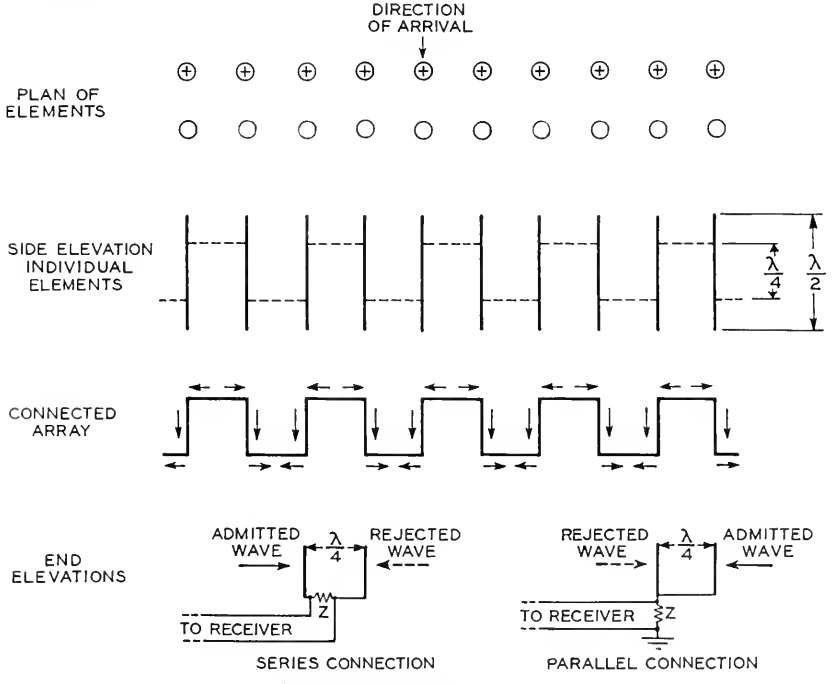


Fig. 17—Bruce antenna array.

Such an antenna as used at Netcong, New Jersey, for reception is shown in Fig. 18.

In the course of time further improvements in directive antennas have been made. One improvement is indicated in Fig. 19 and is known as the "rhombic antenna" [8] because it consists of a wire or wires supported by four poles in the shape of a rhombus elevated some distance above the ground and parallel to the ground. As usually used two contiguous sides of the rhombus form one branch of the antenna and the other two sides form the other branch. At one end is connected the receiver (or transmitter). At the end opposite the connection to the receiver are connected resistances of suitable value.

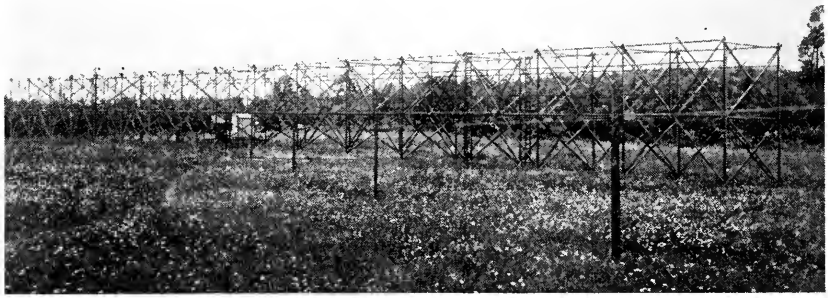


Fig. 18—Bruce receiving antenna at Netcong, New Jersey, used for transoceanic communication. This type of antenna has been superseded by the rhombic type at Netcong.

This antenna differs radically from the previous ones and most other directive antennas in one respect and that is it does not usually operate with a standing wave thereon. The purpose of the resistors is to absorb all the energy reaching such resistors. This antenna when arranged as in Fig. 19 receives from the right. The energy strikes the wires and is thence transmitted to the receivers. Energy coming from the left-hand side travels away from the receivers and when it strikes the resistors is absorbed. If the resistors are not used and the two terminals are either connected together or kept insulated, energy reaching this end will be reflected, in which case this antenna will operate with a standing wave thereon and will receive or transmit from either the right or the left. Inasmuch as this antenna as preferably used is unidirectional and does not operate with a standing

wave, a single antenna may be used for a number of wave-lengths without readjustment.

When the rhombic antenna is used for a number of frequencies without change in size or form, the directivity is different for each frequency. The maximum directivity for the higher frequencies will be a lower angle than for the lower frequencies. This works in very well for long distance operation since the angles at which the high and low frequencies come in tend to agree with this characteristic of the antenna.

The rhombic antenna has come into extensive use for transoceanic short-wave links during the last few years. Due to its multi-wave-

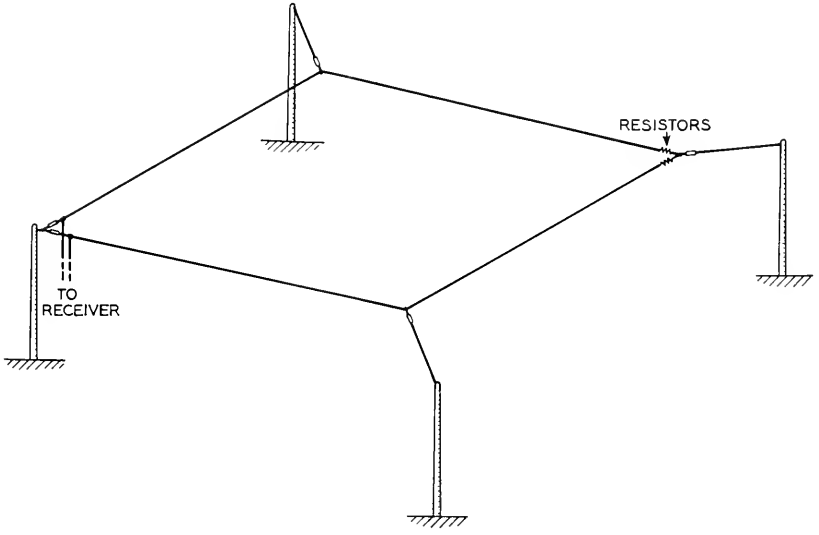


Fig. 19—Rhombic antenna invented by E. Bruce.

length characteristics it is replacing the other types of antennas described previously. The initial cost is much less than the type supported on steel towers, the land required is less, and the upkeep is also less.

In Fig. 20 is shown a photograph of another type of directive antenna. Two antennas are indicated, one for transmitting and one for receiving. These antennas are known as "pine tree" [9] antennas because of the connections of the radiators to a transmission line passing up from below. This particular antenna is for ultra-short-wave operation around 60 MC and is one end of the Provincetown-Green Harbor circuit. Each antenna contains 8 radiators in a plane

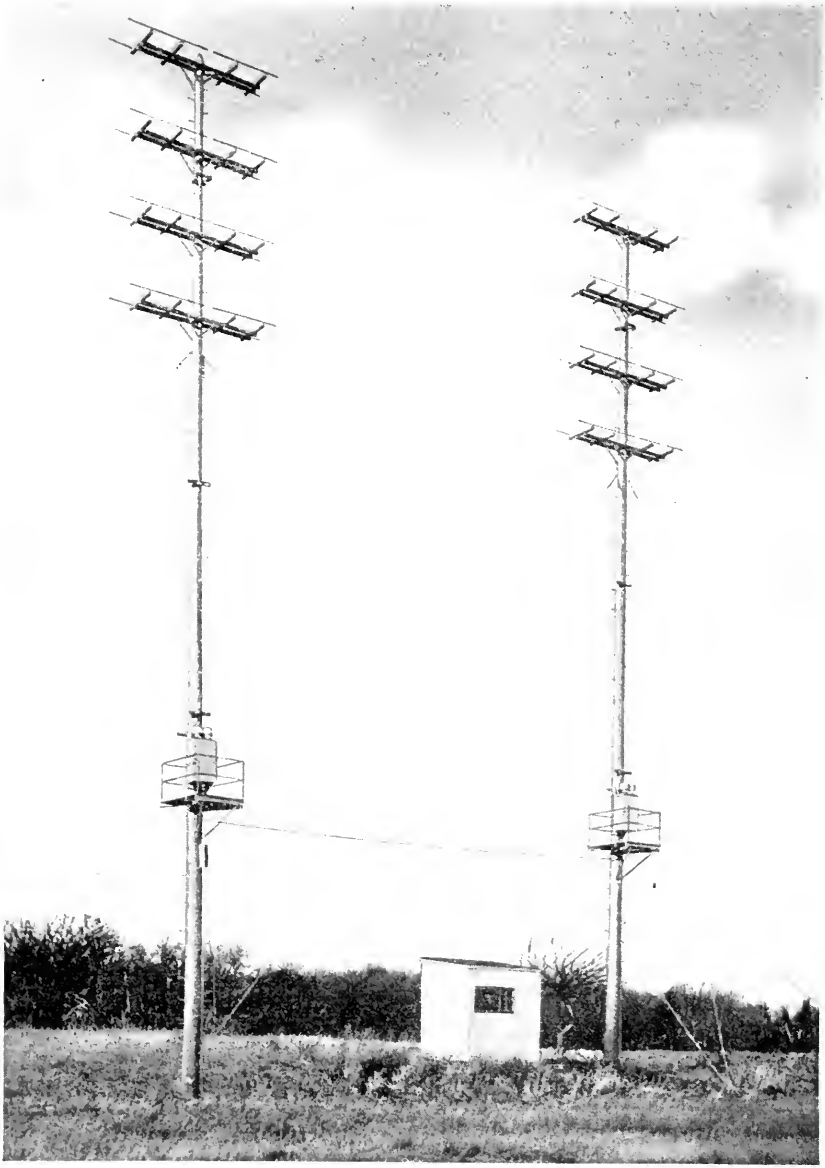


Fig. 20—Pine tree antennas used on Provincetown-Green Harbor ultra-short-wave circuit.

and 8 reflectors in a parallel plane. This type lends itself very well to construction for ultra-short-wave operation. The directivity of any antenna is a function of its breadth measured in wave-lengths and when the wave-length is short it is more economical and easier to construct the antenna with the larger dimensions up and down the pole or at right angles to the pole than to stretch it along the ground.

Figure 21 indicates one of the first directive antennas to be used. This is known as the "Beverage" [10] antenna and is used on long-wave telephone circuits which operate between 60 and 70 kc. Such an antenna is located at Houlton, Maine, for receiving from England. The antenna structure appearing in this diagram may be several miles in length. It receives best from the upper right-hand direction, as indicated by the arrow. The incoming wave produces currents in the antenna which experience gradual build-up along the line to the receiving end where they operate the receiver and deliver the signal to the telephone line. This type of antenna has a horizontal directional pattern as indicated in the figure. Its maximum direction as used is northeast, as that is the direction signals arrive from England. It also happens that more static and strays reach this part of the country from the southwest than from other directions and with the antenna so oriented there is what is sometimes called a "blind eye" faced in the southwest direction so as to receive a minimum of interference from that direction.

Some of the difficulties involved in transoceanic short-wave reception may be explained by reference to Fig. 22; this shows a diagram of the earth and the ionized region of the atmosphere called the "ionosphere." Signals from the transmitting station in England may reach the receiving station in the United States by more than one path. One path indicated has two reflections from the ionosphere and the other has three reflections. Careful measurements on this diagram indicate that these two paths are not equal in length, with the result that signals received in the United States from across the ocean coming over the two paths may be out of phase. Not only may there be two paths but sometimes there are three, four or more so that the interference caused by signals coming over the several paths can give rise to bad fading and distortion. The lower diagram in this figure shows the vertical directive pattern of an ordinary directive antenna. It shows this directive pattern to be large enough to receive simultaneously both incoming signal components from the two paths. If it is desired to eliminate the undesirable effects produced by the two signals coming in out of phase, one should be eliminated. This can be done provided an antenna is constructed having a sharp directive

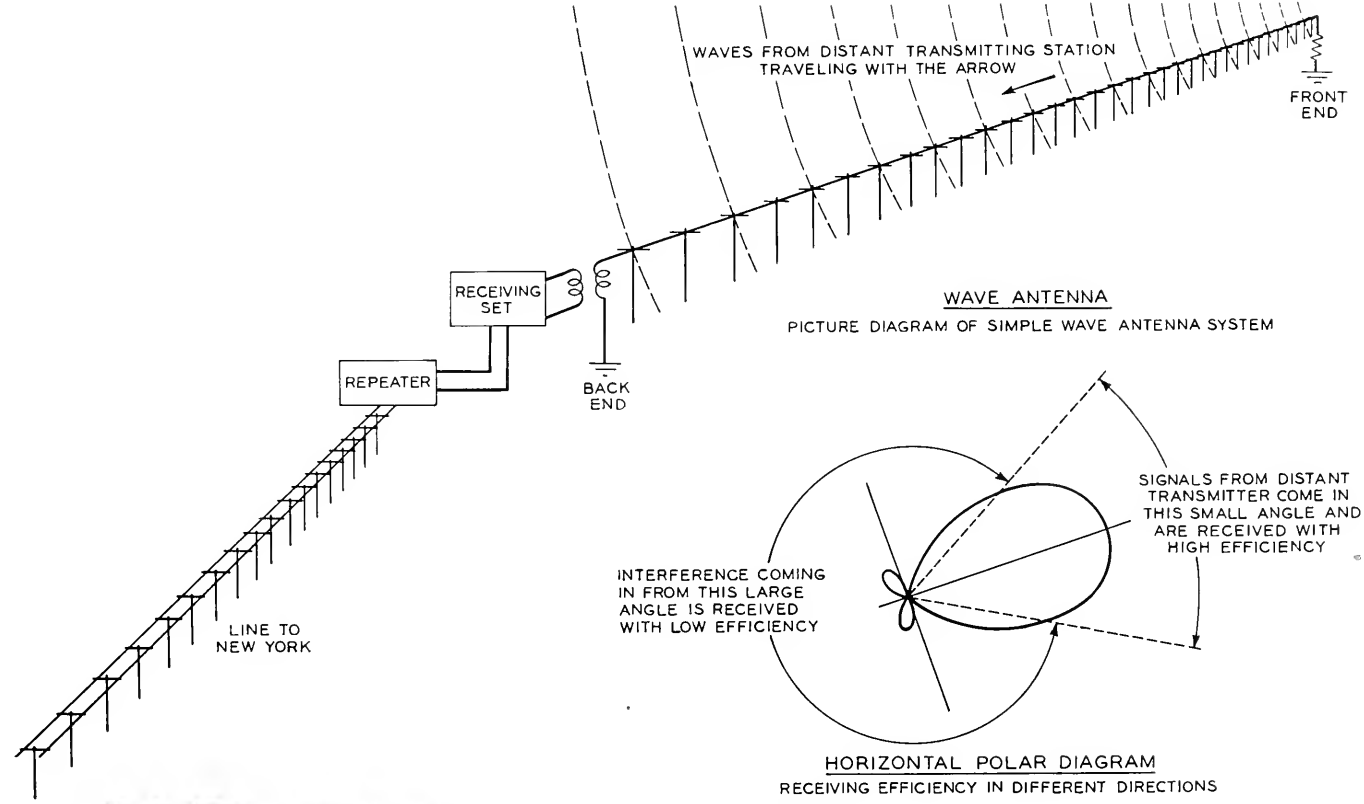


Fig. 21—Beverage antenna.

pattern as indicated in the same sketch and such a directive pattern is produced by an antenna system called the "Musa." [11] The name "Musa" comes from the initials of the words "multiple unit steerable antenna." The Musa is one of the latest developments in directive antennas and possesses not only the important characteristic of a very sharp directivity pattern but also is steerable so it may be altered to receive a desired component, and if such desired component changes its angle of arrival alteration may be made to accommodate such change.

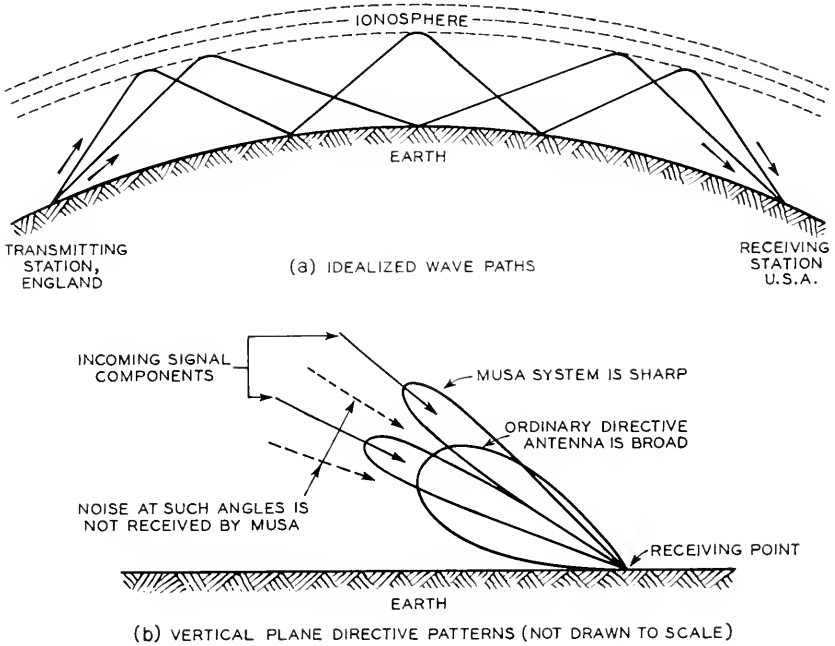


Fig. 22—Paths of short waves over long distances as determined by the ionosphere.

Figure 23 indicates the elements of the Musa. It consists of a row of rhombic antennas lined up in the direction from which signals are expected. Each rhombic antenna is connected by a transmission line to a phase shifter and the outputs of the phase shifters are connected to a receiver. These phase shifters may be so adjusted as to cause any desirable phase additions from the separate antennas. By changing the adjustments on the respective phase shifters the direction of reception may be altered. In this diagram a row of antennas is shown as connected through phase shifters to receiving branch A with the phase shifters adjusted to produce a directive pattern as indicated in the neighboring diagram marked branch A and drawn dotted. Into branch A will therefore come the signals which arrive from one trans-

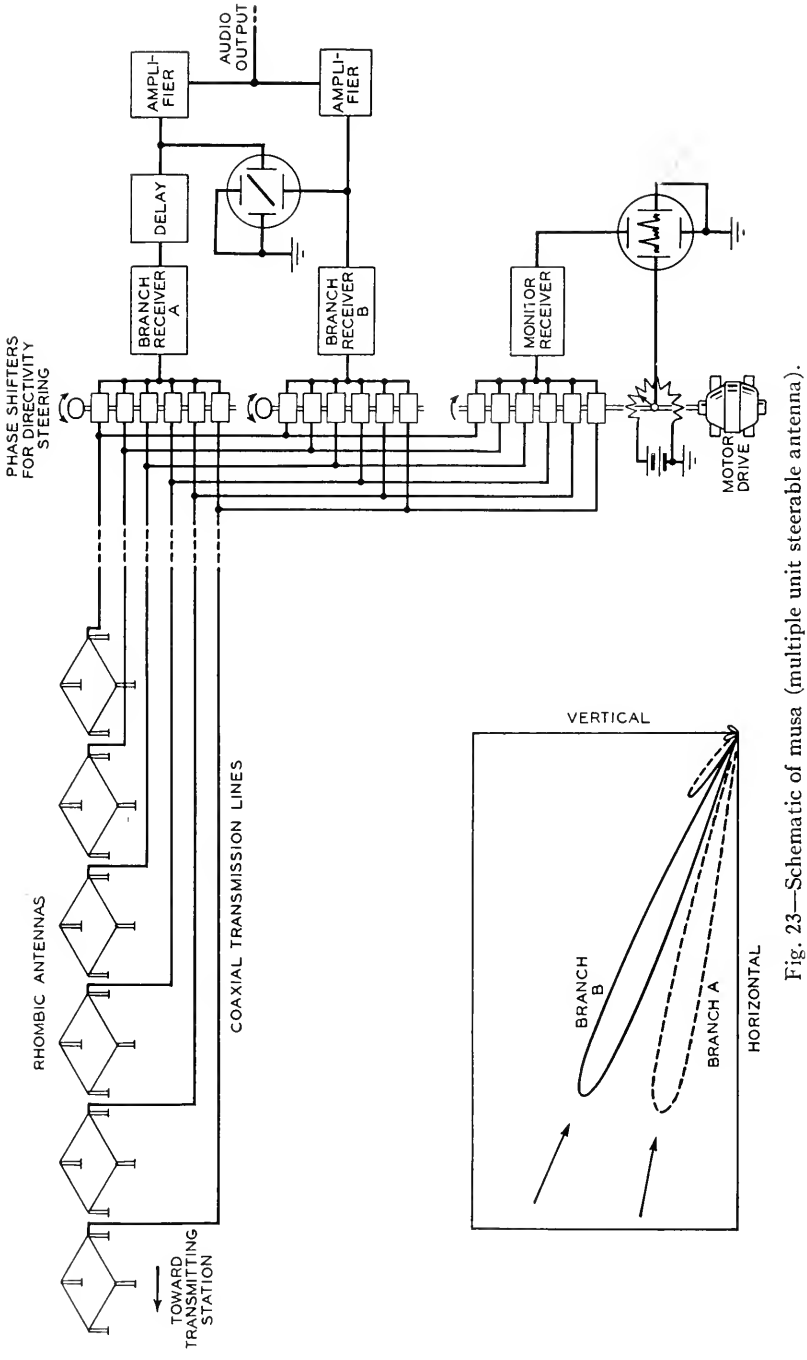


Fig. 23—Schematic of musa (multiple unit steerable antenna).

oceanic path at the lower angle. Beneath this row of phase shifters is indicated a second set connected in parallel with the first set. The second set is adjusted somewhat differently so that the receiver in branch *B* will receive signals over the path with the higher angle as indicated by the solid line directive pattern in the neighboring diagram. The two separate signals may therefore be separately received. Since, however, they do not arrive at identical times they cannot be added without producing distortion. By interposing a delay in the output branch of the one arriving earlier, which in this case is branch *A*, they may be added directly into a single audio output. In this case if one branch fades the other will continue to receive the signal, and while both are receiving a better signal-to-noise ratio obtains.

A cathode ray oscillograph is connected on the outputs of these receivers as indicated in the diagram so as to indicate when the delay is satisfactory for adding the two signals in phase.

As mentioned previously the direction from which a signal comes may change from time to time. In this case a fixed adjustment of the phase shifters may allow the signal to disappear. The operator needs to be ready to change the phase shifters when necessary but it is also desirable that he change them in the desired direction without interfering with the conversation arriving over the circuit, which means he should be able to adjust them without using cut-and-try methods of adjustment. To accomplish this a third set of phase shifters is provided which are continually driven in rotation by a motor so that as they rotate they cause the direction of the signal received by a monitoring receiver to shift its angle of reception over the entire angle range at which signals might arrive. The output of this receiver is connected to two plates of a cathode ray oscillograph while to the other two plates are connected potentials from a rheostat on the phase shifter drive shaft so that the potential will be indicative of the position of the phase shifter. If the phase shifter continually rotates the varying signals will cause the varying deflection in the vertical direction and so draw a pattern as indicated in Fig. 23, which will show immediately by suitable calibration the angle between the two different signals. The operator may thereby adjust his phase shifters by calibration and without cut-and-try tuning.

Inasmuch as the rhombic antenna does not operate with standing waves and may be used for a variety of frequencies it may be used simultaneously for this variety of frequencies. The transmission lines from the rhombic antennas to the phase shifters also operate without standing waves so they likewise may operate simultaneously at a variety of frequencies. It is thus possible by connecting other phase shifters and other receivers to utilize a single row of rhombic antennas

to receive simultaneously a number of radio signals on different frequencies. Each one of these receivers with phase shifters will provide its own directivity pattern and may be adjusted independently of the others.

A Musa system has been constructed for transoceanic reception on short waves at Manahawken, New Jersey, and is now in operation. In this system 16 rhombic antennas are placed in a row approximately a mile and a half long.

The radio terminals for all services have not attained standardized final forms but are in a slow state of flux as better circuits and methods for handling existing problems are devised. The circuits and devices indicated in the figures in principle or in detail are not the only ones that have been tried or used, but represent steps that at one time or another were considered to be advancements suitable to be put into use while the attentions of the development engineers were directed toward more pressing problems. These various devices have augmented the reliability of radio circuits enormously. Distances covered have been enlarged. To accomplish similar results by power increase alone would have in most cases rendered it uneconomic to construct and operate the radio systems. In each case peculiarities either in speech, in radio circuits, or in static and noise characteristics are taken advantage of in making a design to aid the signal and reduce the effect of noise. It is believed that the limit has not been reached but that further improvements and other devices will in due time give increased reliability.

References

1. "The Vodas," S. B. Wright, *Electrical Engineering*, August 1937, or *Bell System Technical Journal*, October 1937.
2. "A Vogad for Radiotelephone Circuits," S. B. Wright, S. Doba and A. C. Dickieson, *I.R.E. Proceedings*, April 1939.
3. "A Radio Telephone System for Harbor and Coastal Service," C. N. Anderson and H. M. Pruden, *I.R.E. Proceedings*, April 1939.
4. "The Compandor, an Aid Against Static in Radio Telephony," R. C. Mathes and S. B. Wright, *Electrical Engineering*, June 1934, and *Bell System Technical Journal*, July 1934.
5. "A Short Wave Single Sideband Radio Telephone System," A. A. Oswald, *I.R.E. Proceedings*, December 1938.
6. "Theoretical and Practical Aspects of Directional Transmitting Systems," E. J. Sterba, *I.R.E. Proceedings*, July 1931.
7. "Transoceanic Telephone Service," A. A. Oswald, *Bell System Technical Journal*, vol. IX, p. 270.
8. "Developments in Short Wave Antennas," E. Bruce, *I.R.E. Proceedings*, August 1931, and "Horizontal Rhombic Antennas," E. Bruce, A. C. Beck and L. R. Lowry, *I.R.E. Proceedings*, January 1935.
9. "An Unattended Ultra-Short-Wave Radiotelephone System," N. F. Schlaack and F. A. Polkinghorn, *I.R.E. Proceedings*, November 1935.
10. "The Wave Antenna," H. H. Beverage, C. W. Rice and E. W. Kellog, *Trans. A. I. E. E.*, vol. 42, p. 215, 1923.
11. "A Multiple Unit Steerable Antenna for Short-Wave Reception," H. T. Friis and C. B. Feldman, *I.R.E. Proceedings*, July 1937.

Abstracts of Technical Articles by Bell System Authors

Lodgepole Pine Poles—Full Length Treatment Under Pressure—Butt Treatment in Open Tanks.[1] C. H. AMADON. Lodgepole pine ("Pinus contorta") forms extensive forests in Colorado, Wyoming, Idaho and Western Montana. The timber has been used widely for mine props, railway ties and telephone and telegraph poles by the various industries in the general region in which it grows. Like most of the pines, lodgepole pine in its natural state is classed as a non-durable wood in contact with the soil, and where a relatively long service life is desired the timber has been treated with creosote or some other wood preservative.

Lodgepole pine poles are exceptionally straight, free from knots of objectionable size, fairly soft and when well seasoned weigh about 30 lb. per cu. ft.

The purpose of this paper is to present information on the behavior, under actual service conditions, of lodgepole pine poles that had been pressure-treated in closed cylinders or butt-treated in open tanks, and to describe the development of a process for the pressure treatment of lodgepole pine poles to meet specific penetration and low retention specification.

Sound Measurement Objectives and Sound Level Meter Performance.[2] J. M. BARSTOW. The standardization of sound level meters is shown to have improved conditions in the field of sound measurement, although several characteristics thought to be desirable in visual indicating sound measuring devices are not fully realized in instruments conforming to the present standards. The extent to which certain sound measurement objectives have been realized in present sound level meters is discussed. Further work will undoubtedly be necessary before some of these objectives may be more completely realized. Present indications are that sound level meter limitations in regard to the approximation of sound jury loudness levels will be difficult to remove and at the same time retain reasonable apparatus simplicity. Some consideration is given to possible courses of action in regard to such limitations.

Coordination of Power and Communication Circuits for Low-Frequency Induction.[3] J. O'R. COLEMAN and H. M. TRUEBLOOD. Where power

[1] *Proc. Amer. Wood-Preservers' Assoc.*, 1940.
[2] *Jour. Acous. Soc. Amer.*, July 1940.
[3] *Electrical Engineering*, July 1940.

and communication facilities are in proximity, electromagnetic induction from the power system may cause disturbances in the communication system. The avoidance or minimizing of such disturbances, with due regard to the service and other needs of both systems, is a problem of coordination, which is conveniently divided into two parts, one dealing with low-frequency inductive coordination and the other with noise-frequency coordination.

The present paper undertakes a general examination of the problem of low-frequency inductive coordination in the light of developments during the past decade. The situation as it existed at the beginning of the decade is to be found well set forth in a paper presented in 1931 at the A.I.E.E. winter convention by R. N. Conwell and H. S. Warren. The present paper, like its predecessor, derives from the work of the Joint Subcommittee on Development and Research of the Edison Electric Institute and the Bell System. It is largely concerned with induction from currents due to power system ground faults and the transients which accompany such faults. It gives relatively little attention to continuous low-frequency effects since, up to the present at least, such effects have not been a primary concern in the low-frequency coordination of commercial power circuits and Bell System communication circuits.

A further object of the paper is to outline the various factors that require consideration in practical situations and to discuss their significance under present-day conditions. To provide necessary background for this, recapitulations of fundamentals are included at appropriate points. Detailed discussions necessarily omitted from the paper itself are to be found in the papers listed in the bibliography, many of which, particularly the Conwell-Warren paper, contain further references.

Insulating Paper in the Telephone Industry.[4] J. M. FINCH. This article discusses briefly a few of the more important types of paper insulations used by the telephone industry, and shows the relation the manufacturing procedures bear to the initial properties, the permanence, and the uses of the product. Special emphasis is placed on chemical properties as criteria of permanence. The specification control of paper is discussed with emphasis on the simplification of chemical test methods and on minimizing the number of such tests. Finally, mention is made of some of the modified forms of cellulose, which possess insulating characteristics superior to paper and which are already replacing it for some uses.

[4] *Indus. and Engg. Chemistry*, August 1940.

Rectilinear Electron Flow in Beams.[5] J. R. PIERCE. Electrodes are devised by means of which rectilinear electron flow according to well-known space charge equations can be realized in beams surrounded by charge-free space. It is shown how these electrodes can be used in the design of electron guns having desirable characteristics.

High-Gain Amplifier for 150 Megacycles.[6] G. RODWIN and L. M. KLENK. An ultra-high-frequency amplifying system is described which operates at about 150 megacycles with an over-all gain of 114 decibels and transmitted band of over 2 megacycles. An output power of 2.5 watts is available with a signal-to-distortion ratio of 60 decibels. By a frequency-shifting modulator in the amplifier chain the input and output are made to differ by 10 megacycles. A filter-type circuit is used as the interstage coupling to give the necessary band width.

Room Noise at Subscribers' Telephone Locations.[7] D. F. SEACORD. The effect of room noise on the ability to hear speech is roughly equivalent to a partial deafening of the listener; hence the study of room noise conditions at telephone locations is of considerable interest to the telephone engineer since these conditions have an important bearing on the degree of satisfaction with which speech is received over a telephone connection. As a consequence, various studies of room noise have been made from time to time and information of increasing value has been obtained over a period of years with the development of improved measuring equipment and technique. This paper is based on the results of recent room noise surveys carried out in the Bell System and gives a broad picture of the magnitude of room noise at subscribers' telephone locations under present-day conditions. The data presented are a part of the information required in the work of devising and applying methods for taking into account the effects of room noise on telephone transmission in the design of the telephone plant.

Temperature Effects in Secondary Emission.[8] D. E. WOOLDRIDGE. Measurements have been made on the effects of temperature changes on the emission of secondary electrons from iron, nickel, cobalt, and molybdenum. Abrupt changes of one or two per cent were observed to accompany the $\alpha-\gamma$ transition of iron, while the hexagonal to face-centered cubic transformation of cobalt was accompanied by a change in secondary emission of only about 0.4 per cent. The magnetic trans-

[5] *Jour. Applied Physics*, August 1940.
[6] *Proc. I. R. E.*, June 1940.
[7] *Jour. Acous. Soc. Amer.*, July 1940.
[8] *Phys. Rev.*, August 15, 1940.

formation was found to alter the secondary emission coefficient of nickel by less than 0.3 per cent. The temperature coefficient of secondary emission, in the cases of nickel, cobalt, and molybdenum, was found to be much less than the volume coefficient of expansion of the metal. The smallness of the temperature coefficient and the effect of the magnetic transformation are shown to lend support to the view that the secondary electrons are scattered or "absorbed" by an excitation process similar to that whereby they are originally produced.

Contributors to this Issue

MILLARD W. BALDWIN, JR., E.E., Cornell, 1925; M.A., Columbia, 1928. Bell Telephone Laboratories, 1925–. Mr. Baldwin's work has been in the field of telephotography and television.

W. R. BENNETT, B.S., Oregon State College, 1925; A.M., Columbia University, 1928. Bell Telephone Laboratories, 1925–. Mr. Bennett has been engaged in the study of the electrical transmission problems of communication.

G. K. BURNS, S.B., S.M., Massachusetts Institute of Technology, 1935. Western Electric Company, 1935–1940; Bell Telephone Laboratories, 1940–. Mr. Burns has been engaged in the development of methods and facilities for the manufacture of crystal filters. At present he is on a temporary assignment with the Laboratories' filter design group.

HOMER DUDLEY, B.S. in Electrical Engineering, Pennsylvania State College, 1921; M.A., Columbia University, 1924. Bell Telephone Laboratories, 1921–. Mr. Dudley has worked on transmission problems. In recent years, the work has been in research on circuits for speech analysis and synthesis.

MARK B. GARDNER, B.S., Brigham Young University, 1930. Bell Telephone Laboratories, 1930–. Mr. Gardner has been engaged in various studies in physiological acoustics.

RAYMOND A. HEISING, E.E., University of North Dakota, 1912; M.S., University of Wisconsin, 1914. Western Electric Company, 1914–25; Bell Telephone Laboratories, 1925–. Mr. Heising has been engaged in research and development on all phases of radio and is best known for his work on radio transmitters.

H. C. MONTGOMERY, A.B., University of Southern California, 1929; M.A., Columbia University, 1933. Bell Telephone Laboratories, 1929–. Engaged at first in studies of hearing acuity and related problems in physiological acoustics, Mr. Montgomery has been occupied more recently with the study and analysis of speech.

J. C. STEINBERG, B.Sc., M.Sc., Coe College, 1916, 1917. U. S. Air Service, 1917–19. Ph.D., Iowa University, 1922. Engineering Department, Western Electric Company, 1922–25; Bell Telephone Laboratories, 1925–. Dr. Steinberg's work since coming with the Bell System has related largely to speech and hearing.