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 experimental 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

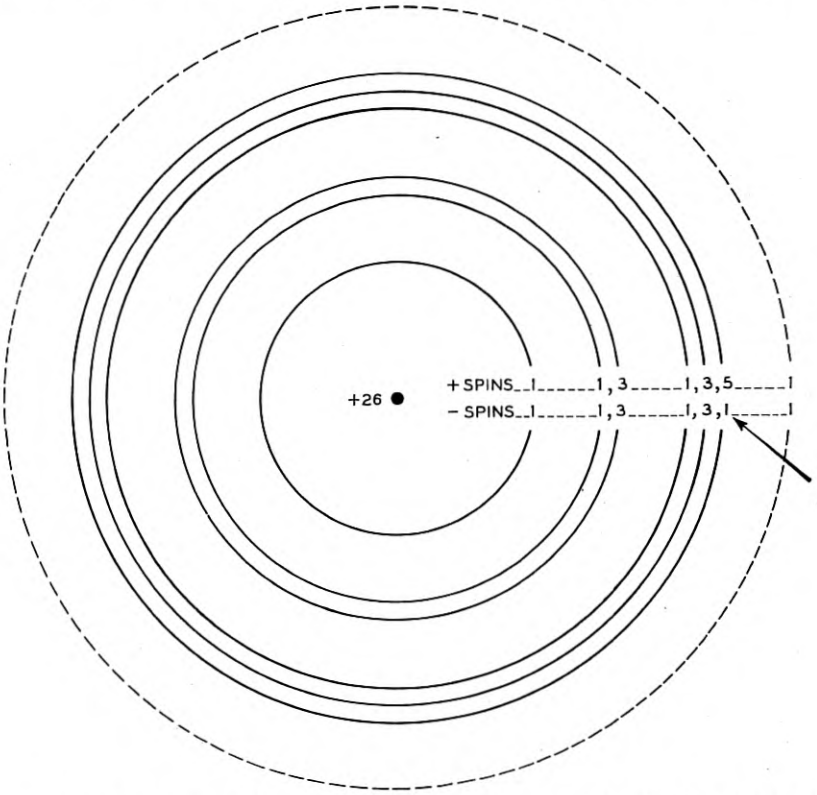


Fig. 1—Electron shells in an atom of iron. The arrow indicates the incomplete sub-shell that is responsible for ferromagnetism. The numbers specify how many electrons with each spin are in the corresponding sub-shells.

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.

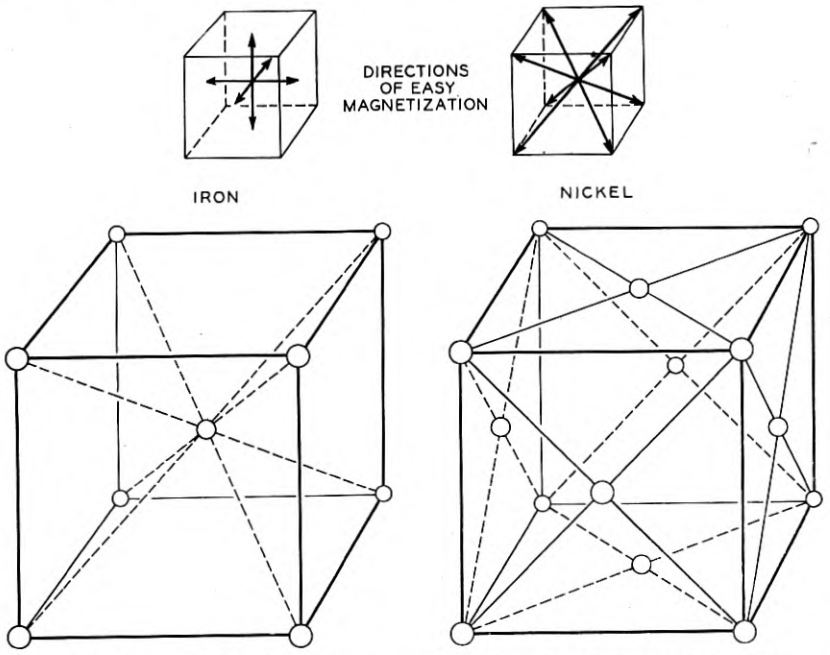


Fig. 2—The positions of the atoms and the directions of easy magnetization in crystals of iron and of nickel.

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

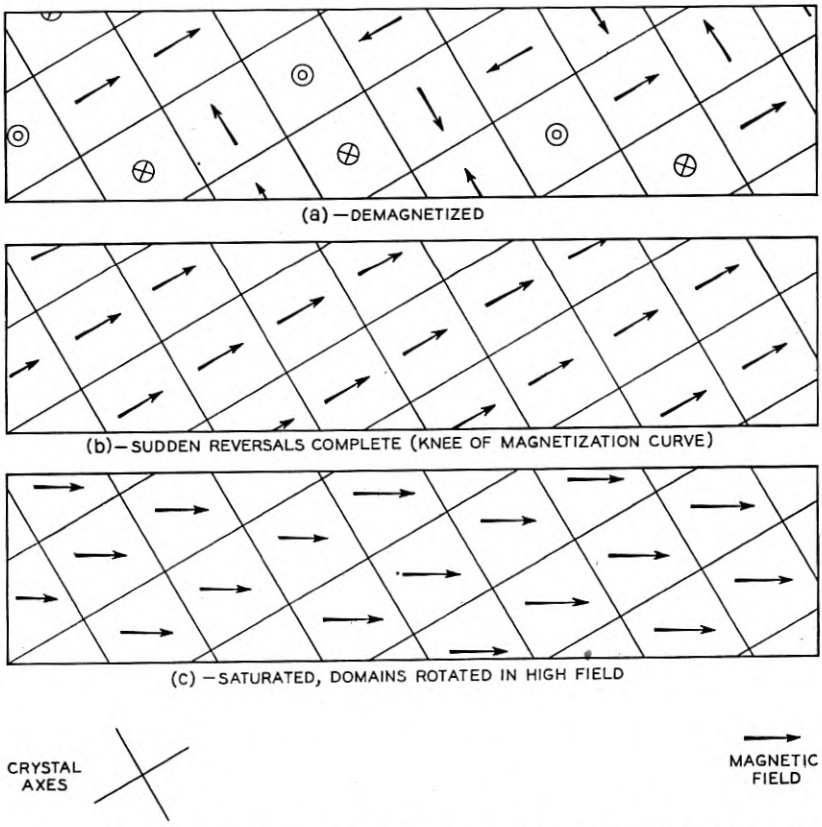


Fig. 3—Domains in a single crystal of iron. As the magnetic field increases in strength the magnetic moments first change suddenly (*a* to *b*) by displacement of the boundaries between them, then rotate smoothly (*b* to *c*).

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.

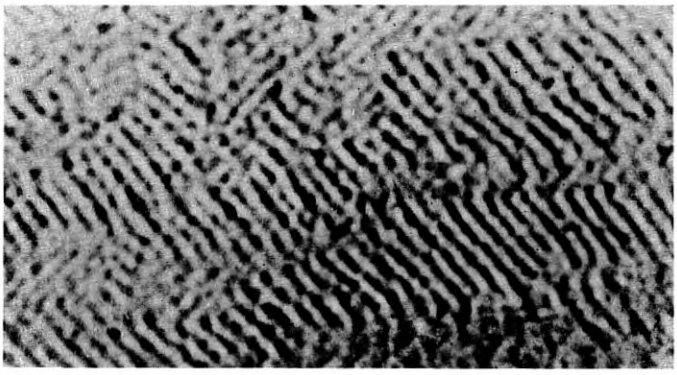


Fig. 4—The powder pattern produced by colloidal iron oxide on the surface of a demagnetized silicon-iron crystal, showing the presence of inhomogeneous magnetic fields. Magnification about 1000.

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.

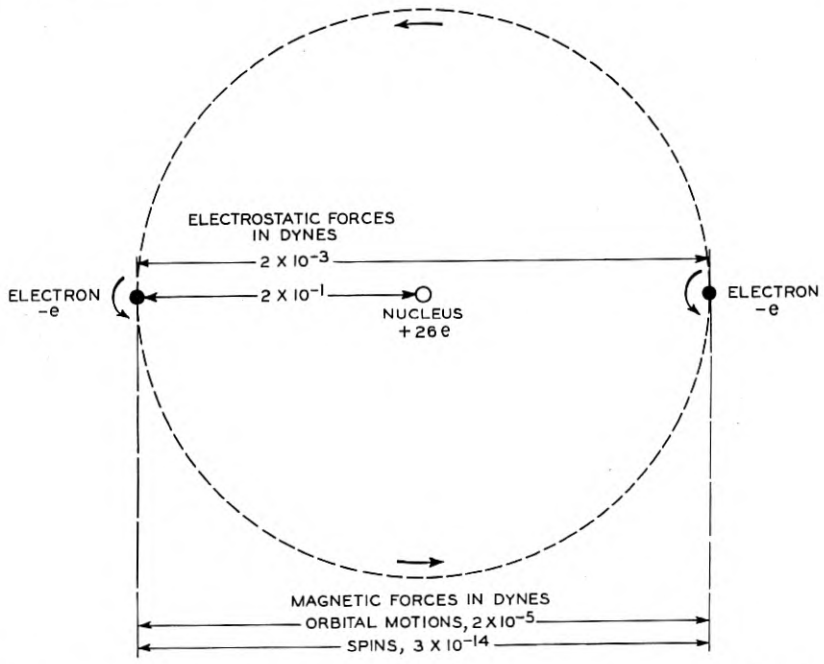


Fig. 5—The magnitudes of the forces in a hypothetical iron-like atom, showing that electrostatic forces are more powerful than magnetic forces.

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

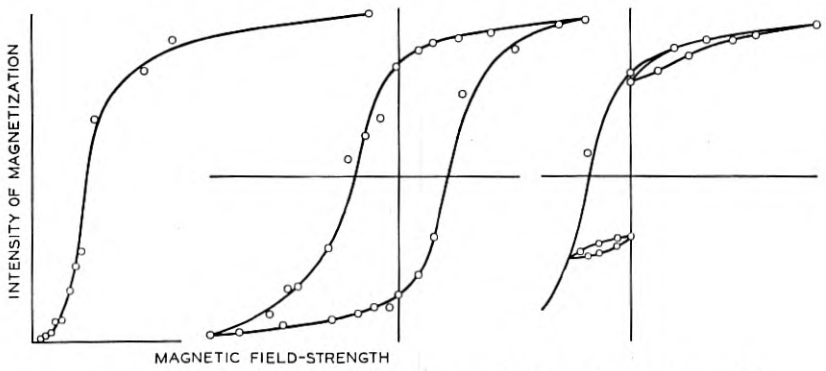


Fig. 6—A magnetization curve and hysteresis loops of a Ewing model of 130 pivoted magnets in square array.

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

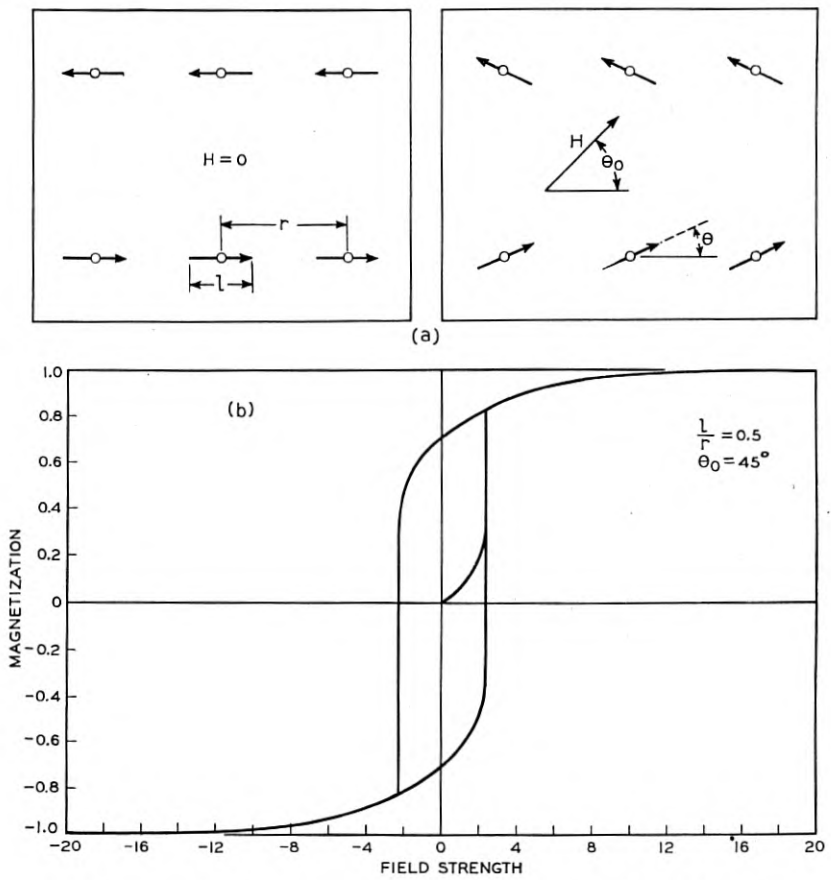


Fig. 7—A magnetization curve and hysteresis loop for an infinite line of equally spaced magnets originally "demagnetized."

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

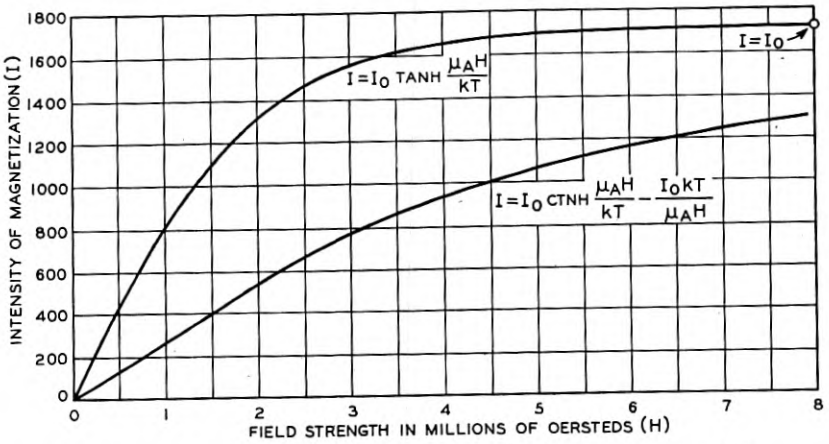


Fig. 8—With no helpful mutual action between atoms, enormous fields would be necessary to saturate a magnetic material.

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

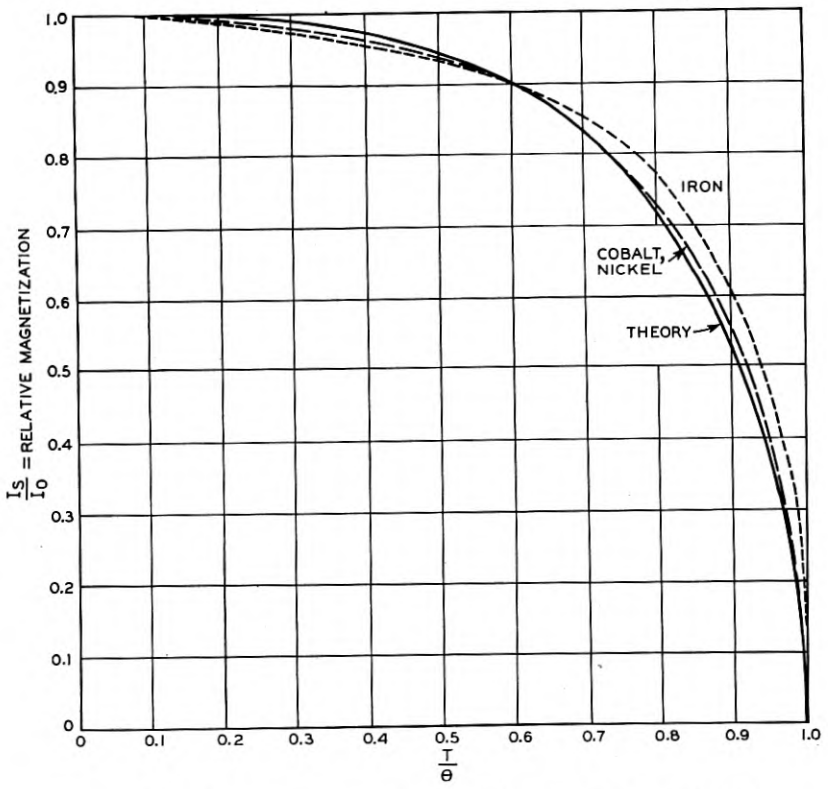


Fig. 9—Dependence on the temperature of the saturation magnetization of iron, cobalt and nickel, as compared with theory.

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).

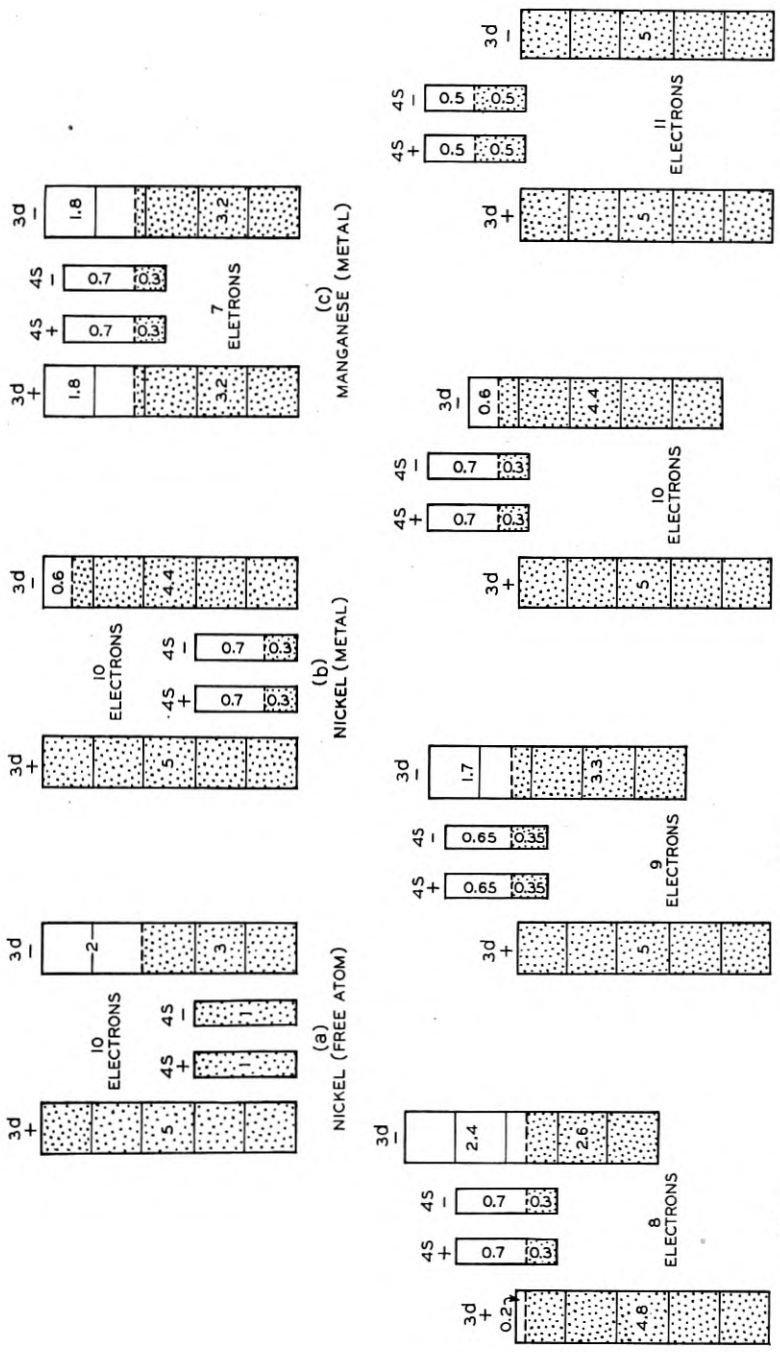


Fig. 10—The distribution of electrons among the possible electron positions in a free atom of nickel, and in manganese, iron, cobalt, nickel and copper atoms that form part of a metal.

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 3*d*+ level is represented as lower in energy than the 3*d*− since one of these bands is preferred.

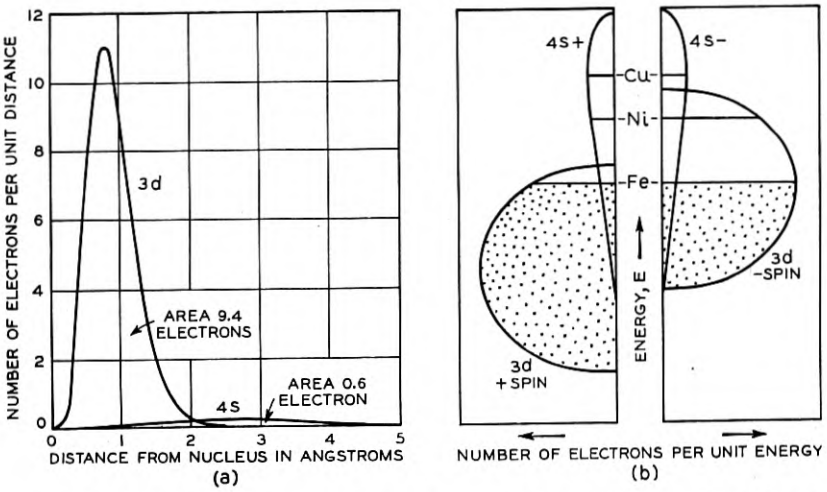


Fig. 11—The filling of electron positions in iron, and some elements near it in the periodic table. Electron positions for closed shells, containing 18 electrons, are not shown.

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 3*d*+ nor 3*d*− 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 3*d*+ but not 3*d*−, 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 3*d* bands, losing their more loosely bound 4*s* electrons when there are available places of lower energy. The data for palladium indicate that this element has the same number of outer

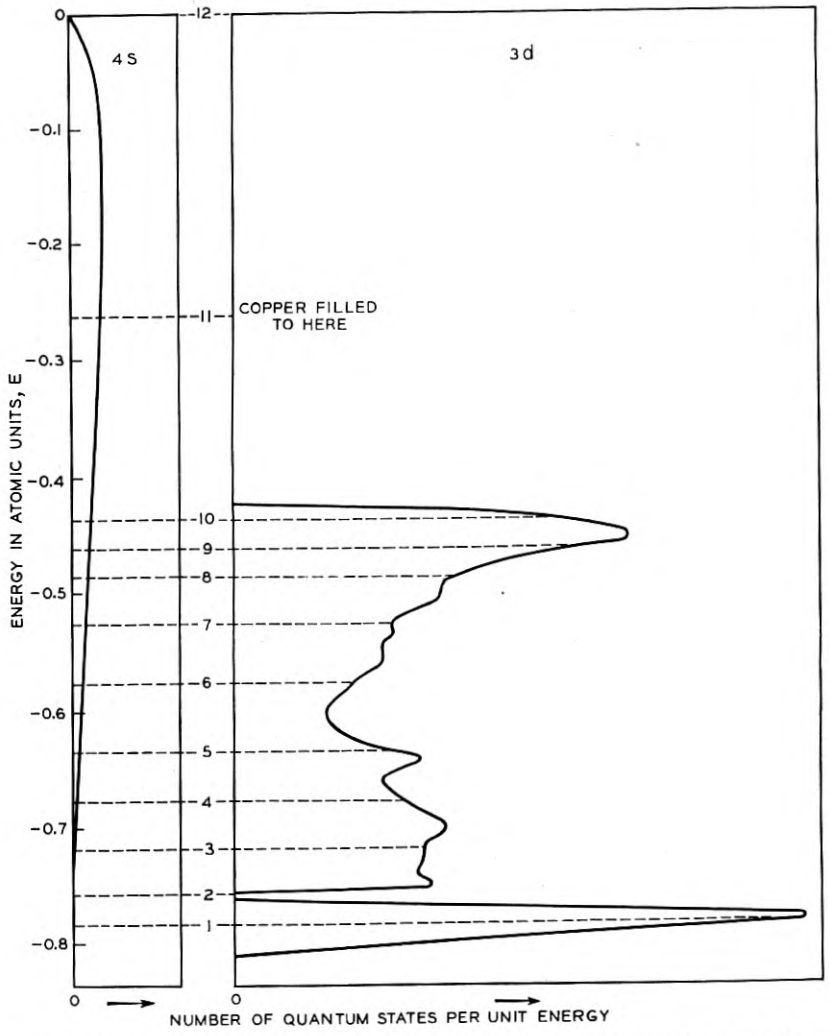


Fig. 12—Energy levels in the 3*d* and 4*s* shells in copper, according to Slater. Similar levels are believed to exist in nickel and cobalt with the levels filled to "10" and "9" respectively.

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

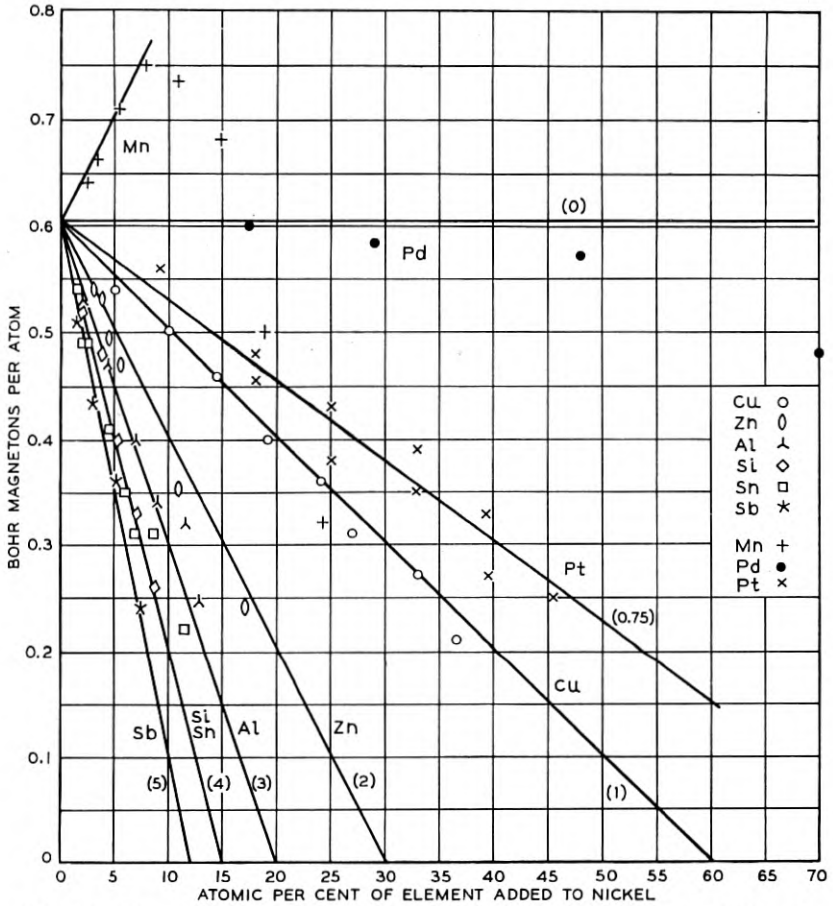


Fig. 13—The saturation magnetization of nickel decreases upon the addition of other elements having 1, 2, 3, . . . electrons in the outermost shell.

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).

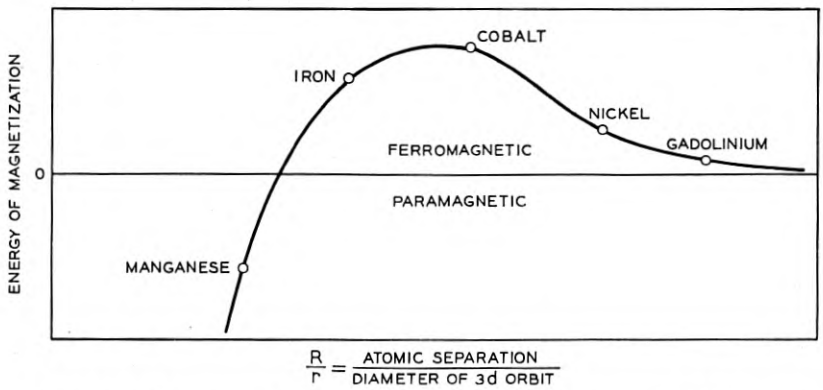


Fig. 14—Bethe's curve relating the energy of magnetization to the distance between atom-centers, with a fixed diameter of the unfilled inner shell that has the magnetic moment.

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),

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

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

$$J = \mu_A NI_0/z = k\theta/z.$$

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

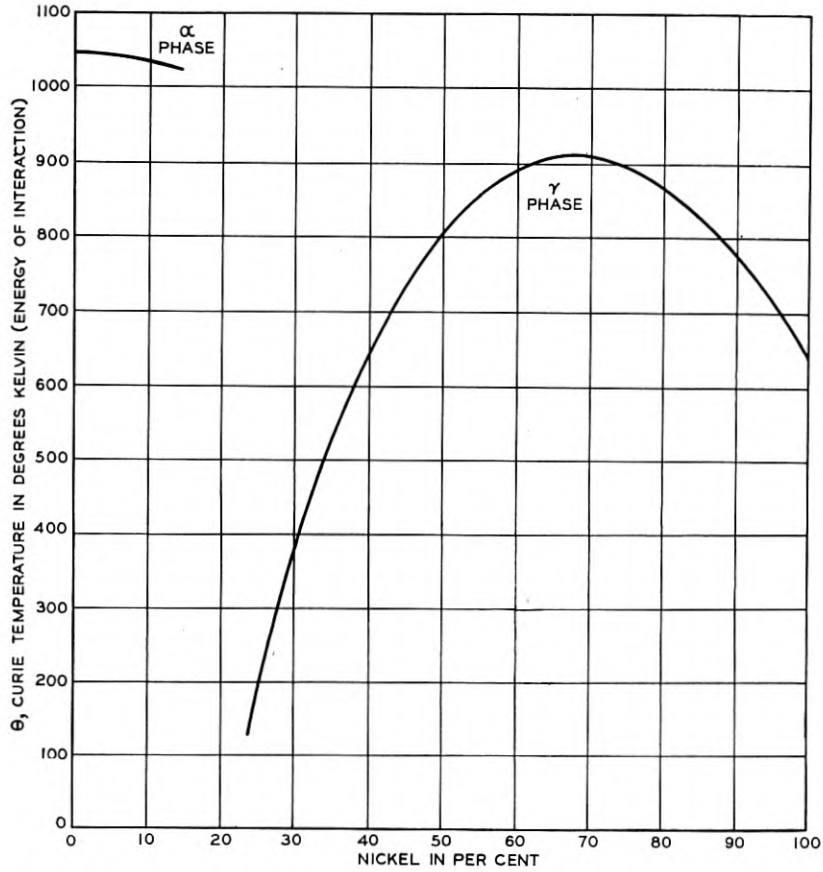


Fig. 15—The Curie temperatures for iron-nickel alloys, showing a maximum corresponding to the maximum of Bethe's curve of Fig. 14.

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

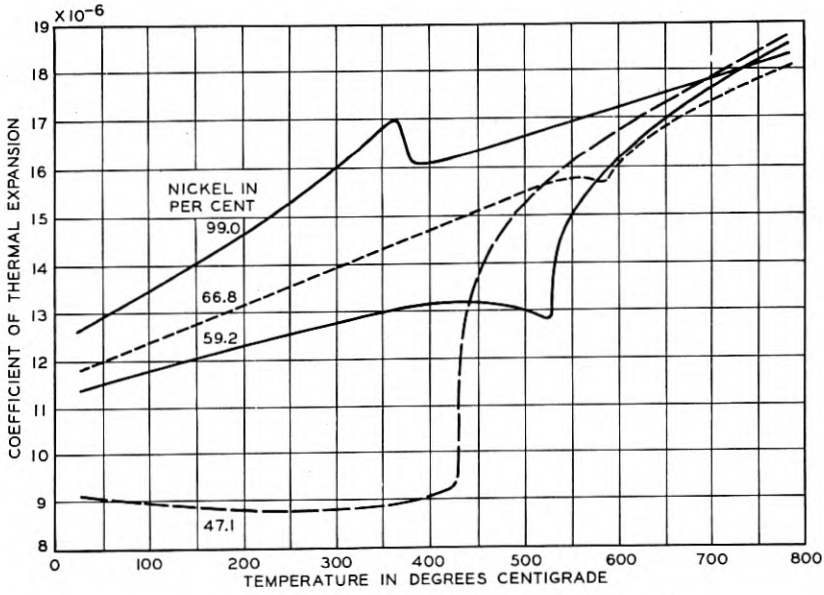


Fig. 16—The expansion coefficient of some iron-nickel alloys, showing the magnetic anomoly and its change in sign at about 70 percent nickel.

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

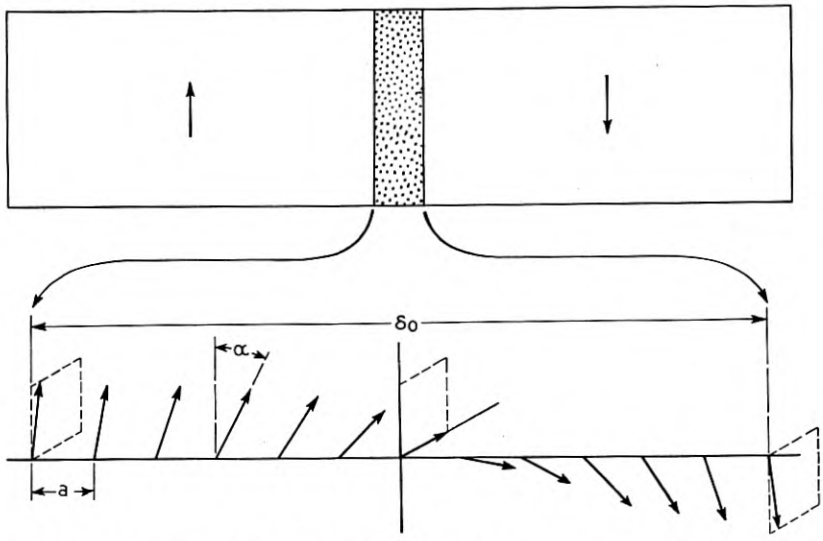


Fig. 17—The nature of the domain boundary. The transition region between two domains is believed to be about 1000 atom diameters thick.

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

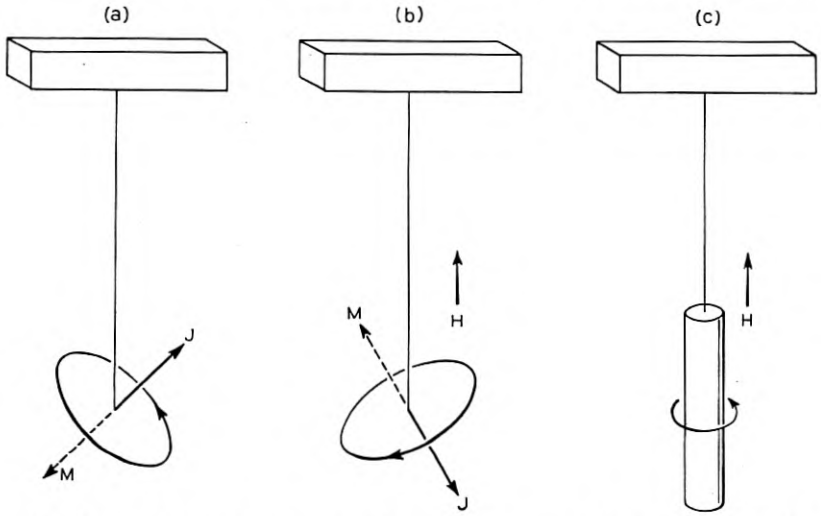


Fig. 18—The magnetic moment, M, and the moment of momentum, J, of an electron in its orbit about the nucleus. A change in one moment entails a change in the other, the (gyromagnetic) ratio remaining constant.

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

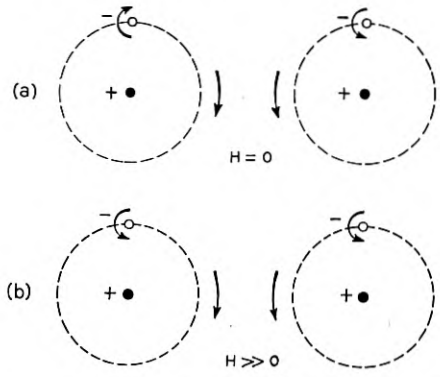


Fig. 19—In the common ferromagnetic materials a change in magnetization is effected by a change in the direction of electron spin, not in the direction of motion of the electron in its orbit.

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.

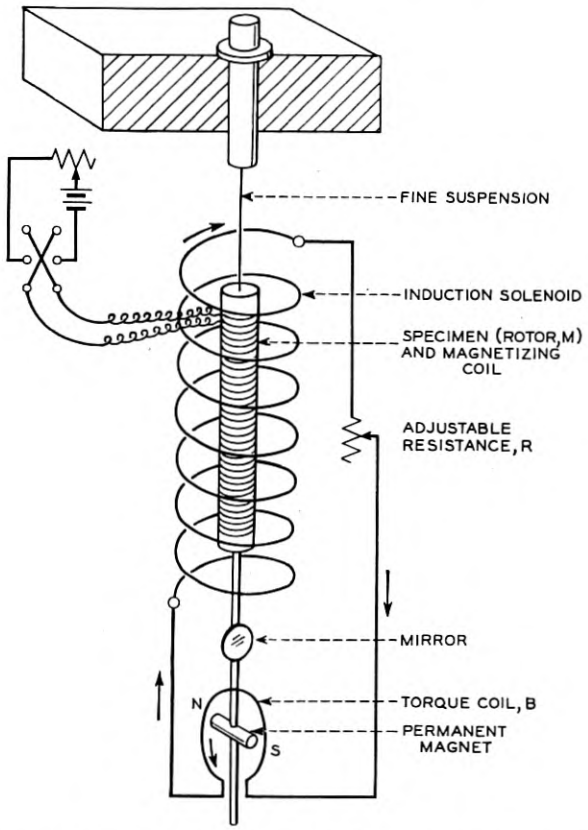


Fig. 20—Schematic diagram of the method of determining the gyromagnetic ratio.

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-

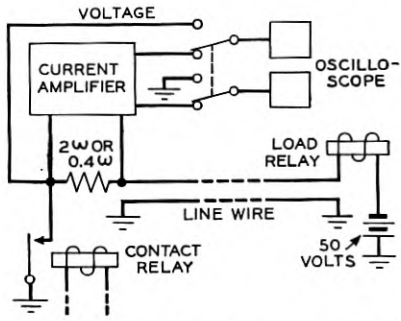


Fig. 1—Typical relay and contact circuit.

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-

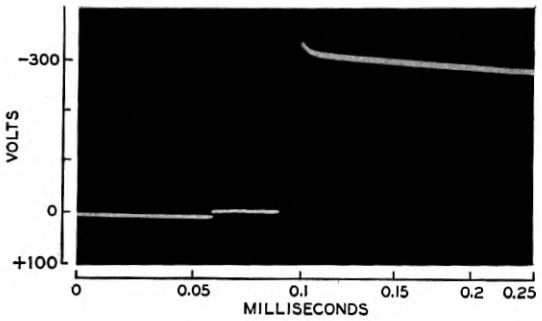


Fig. 2—"A" transient starting with metallic arc (voltage).

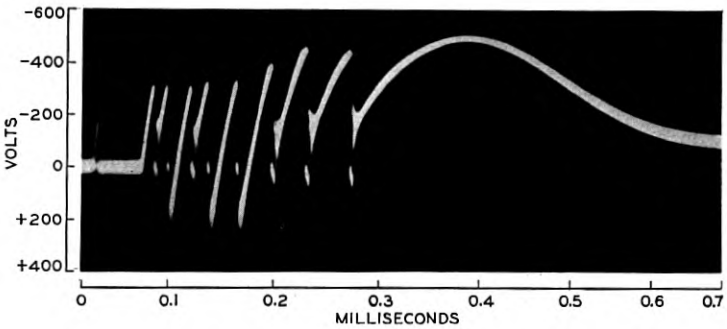


Fig. 3—Entire "B" transient (voltage).

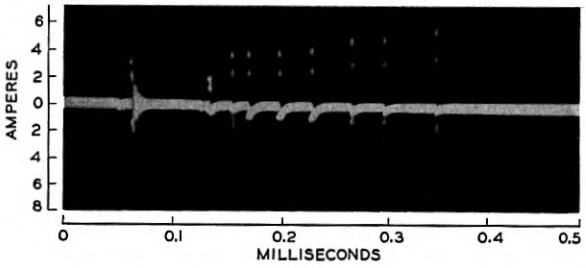


Fig. 4—Entire "B" transient (current).

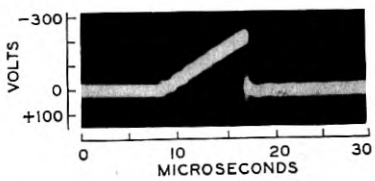


Fig. 5—Initial opening and reclosure—"B" transient (voltage).

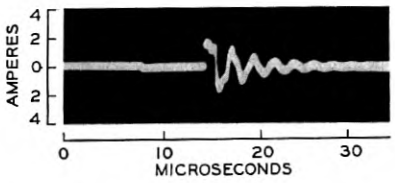


Fig. 6—Initial opening and reclosure—"B" transient (current).

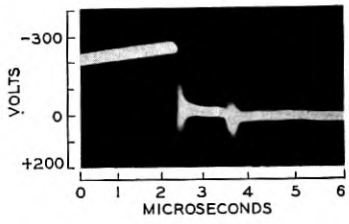


Fig. 7—Initial opening and reclosure—"B" transient (voltage) rapid sweep.

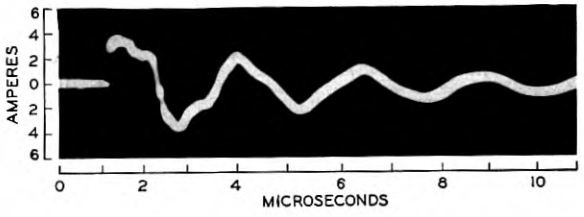


Fig. 8—Initial opening and reclosure—"B" transient (current) rapid sweep.

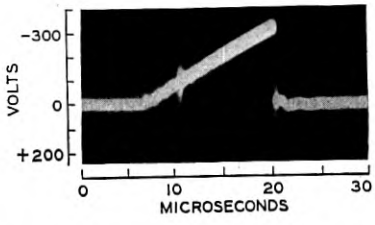


Fig. 9—Evidence of point discharge, voltage.

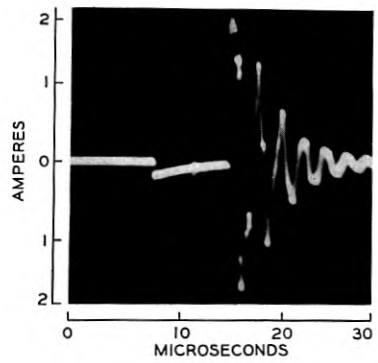


Fig. 10—Evidence of point discharge, current.

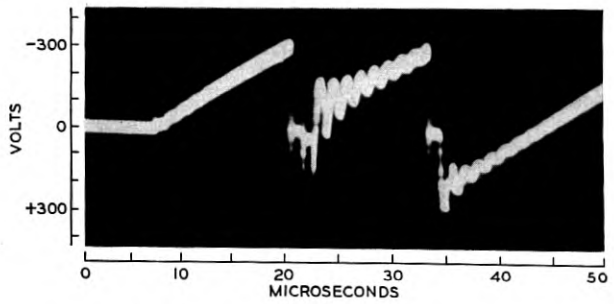


Fig. 11—Early part of "B" transient, voltage.

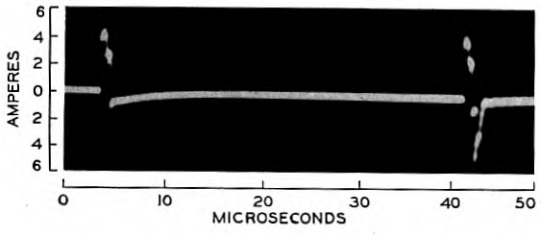


Fig. 12—Early part of "B" transient, current.

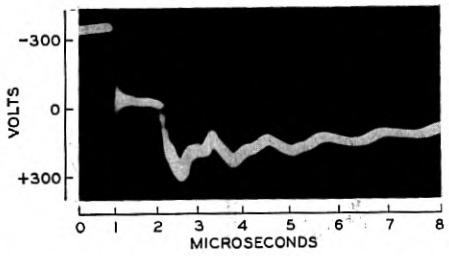


Fig. 13—Single sparkover of "B" transient, with single arc (voltage).

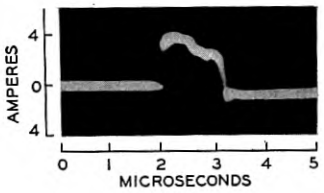


Fig. 14—Single sparkover of "B" transient, with single arc (current).

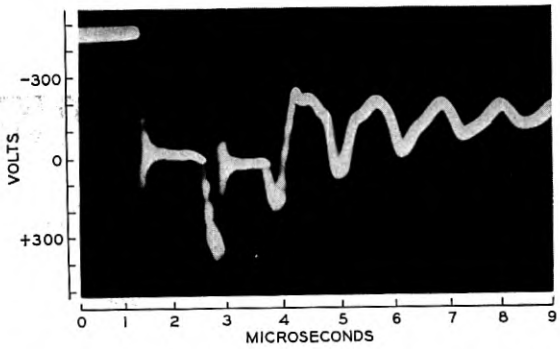


Fig. 15—Single sparkover of "B" transient, with double arc (voltage).

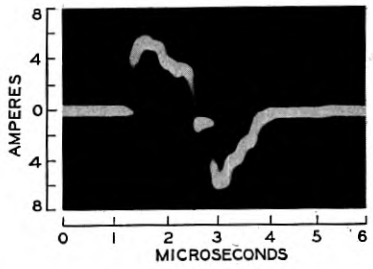


Fig. 16—Single sparkover of "B" transient, with double arc (current).

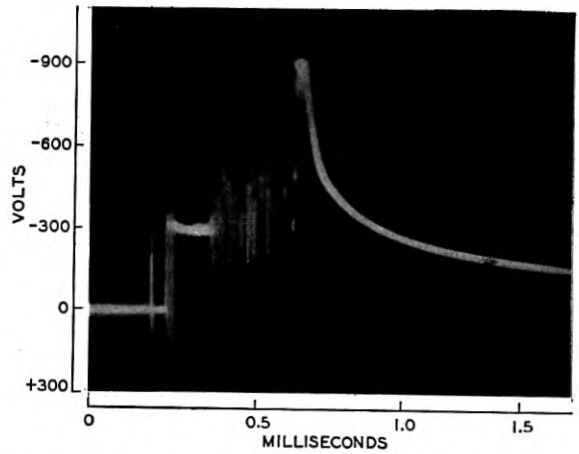


Fig. 17—Typical "mixed A and B" transient (voltage).

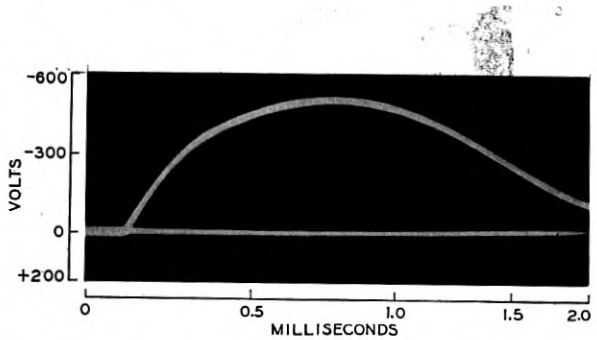


Fig. 18—Effect on voltage transient of changing wire line length—1100 ft.

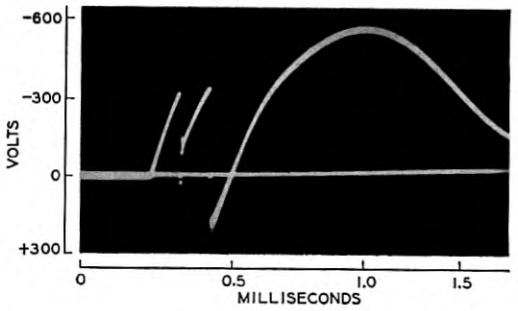


Fig. 19—Effect on voltage transient of changing wire line length—600 ft.

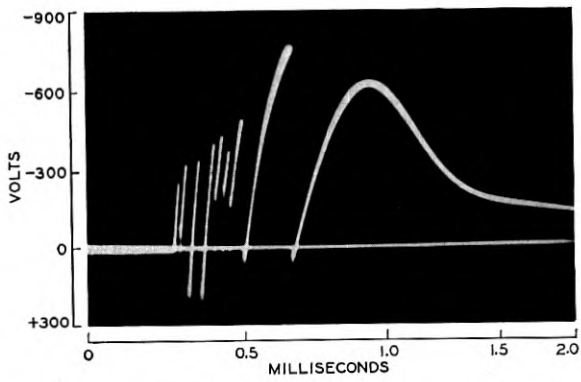


Fig. 20—Effect on voltage transient of changing wire line length—150 ft.

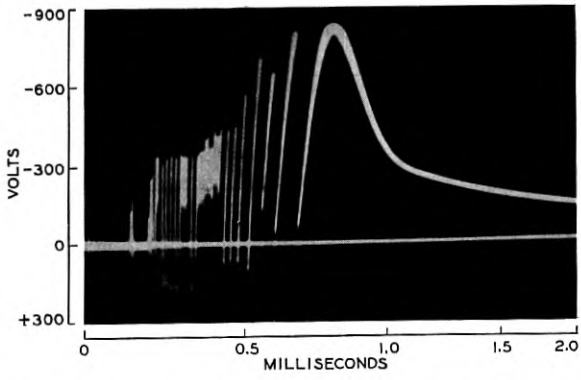


Fig. 21—Effect on voltage transient of changing wire line length—50 ft.

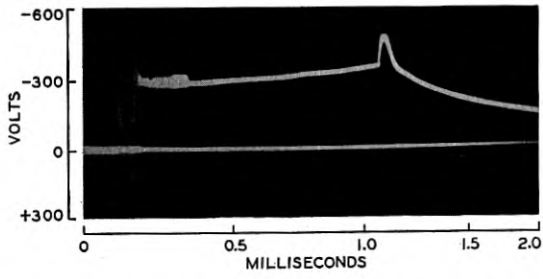


Fig. 22—Effect on voltage transient of changing wire line length—10 ft.

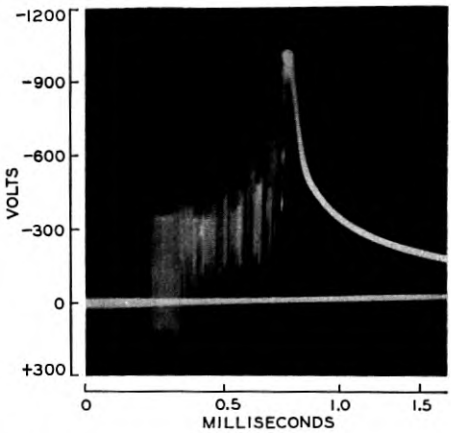


Fig. 23—Effect on voltage transient of changing wire line length—10 ft.

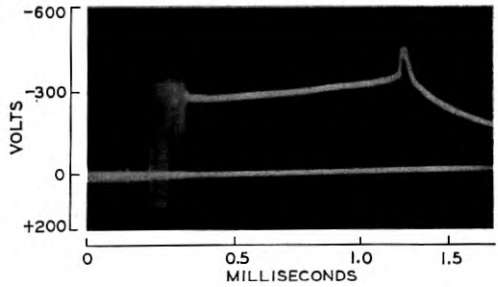


Fig. 24—Effect on voltage transient of changing wire line length—10 ft.

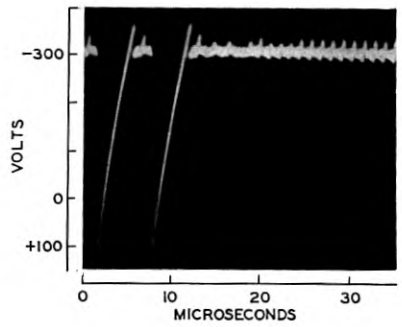


Fig. 25—Oscillation on glow discharge of "A" transient (voltage).

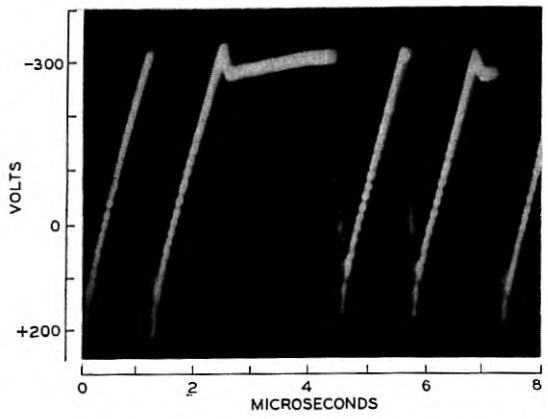


Fig. 26—Start of "A" transient (voltage).

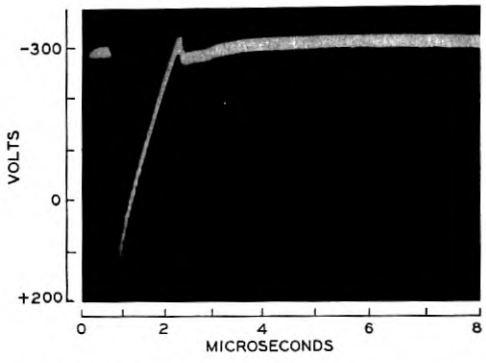


Fig. 27—Start of stable glow discharge of "A" transient (voltage).

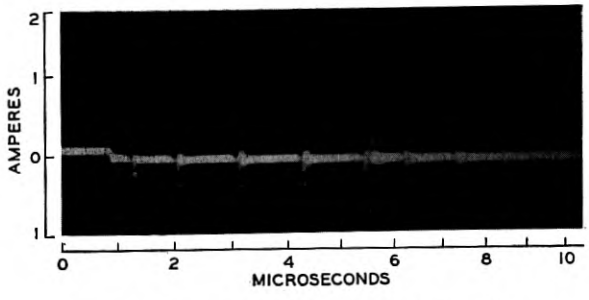


Fig. 28—Start of "A" transient (current).

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.

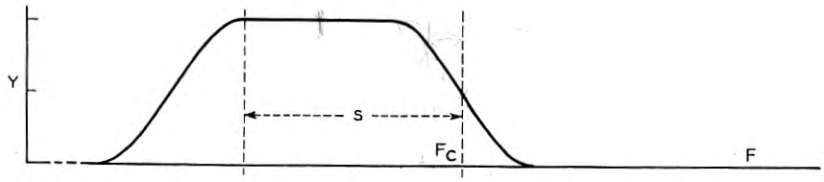


Fig. 1—Idealized transfer admittance characteristic. (Band pass system with no delay distortion; F_c is the carrier and s the fundamental dotting frequency.)

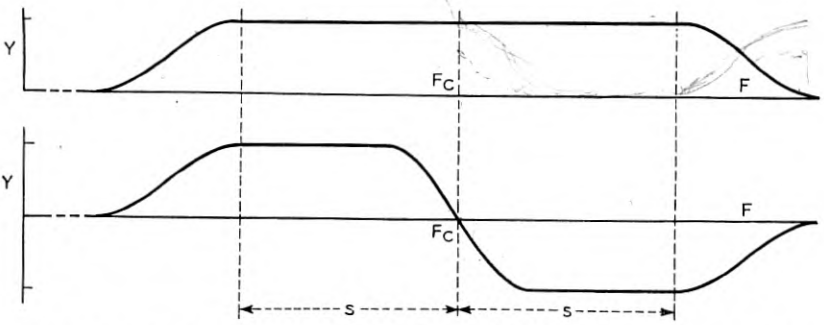


Fig. 2—Graphical analysis of transmission characteristic. (The sum of these characteristics equals that in Fig. 1. The upper gives received signals with carrier in phase, and the lower, in quadrature with the sent wave.)

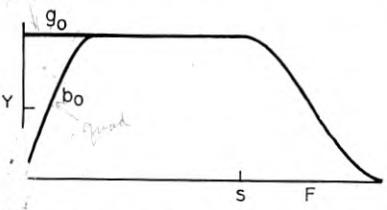


Fig. 3—Equivalent low-pass filter characteristic. (Used in computing envelopes of received signals for single sideband transmission.)

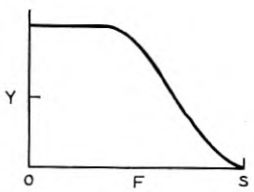


Fig. 4—Equivalent low-pass filter characteristic. (Used in computing envelopes of received signals with mid-band carrier.)

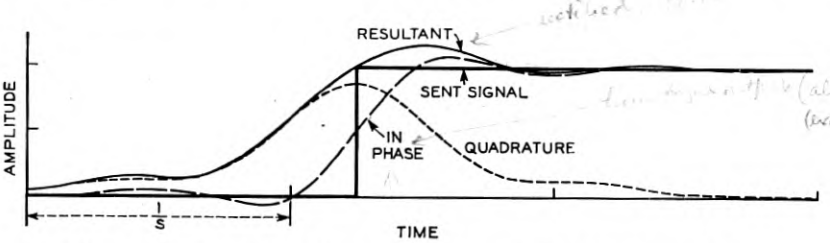


Fig. 5—Envelopes of received wave components for single transition, characteristic as shown in Fig. 1.

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

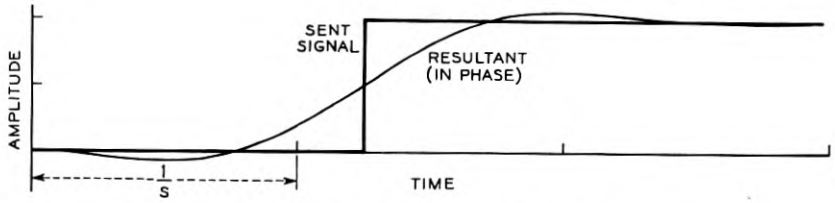


Fig. 6—Transmission of single reversal, carrier at mid-band.

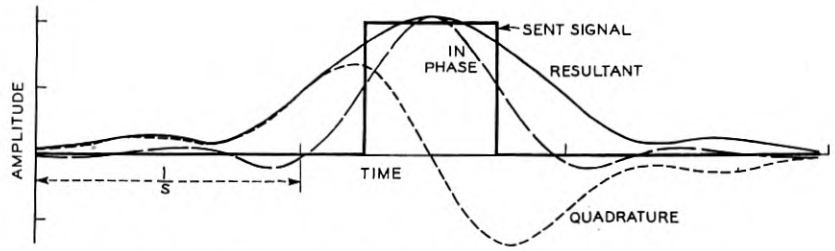


Fig. 7—Received signal for single dot, characteristic as in Fig. 1.

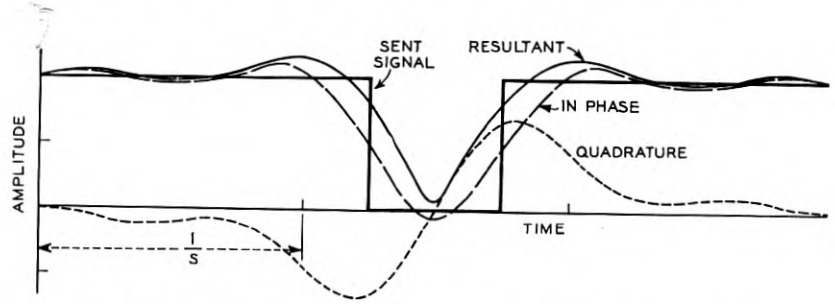


Fig. 8—Received signal for single space, characteristic as in Fig. 1.

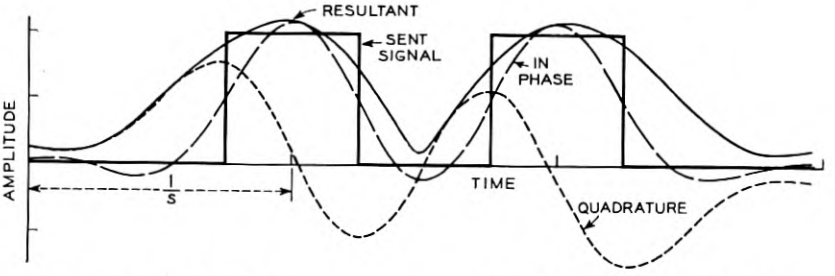


Fig. 9—Received signal for two dots in succession, characteristic as in Fig. 1.

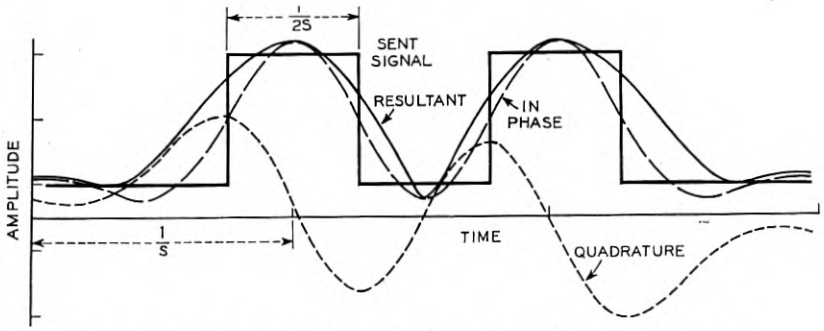


Fig. 10—Effect of transmitting dark current 14 db below maximum. (Compare with Fig. 9.)

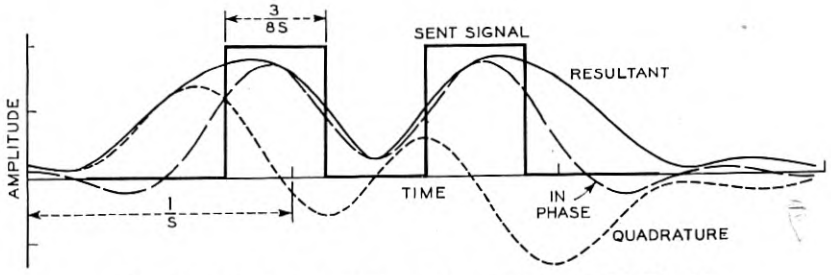


Fig. 11—Effect of shortening dots. (Compare with Fig. 9.)

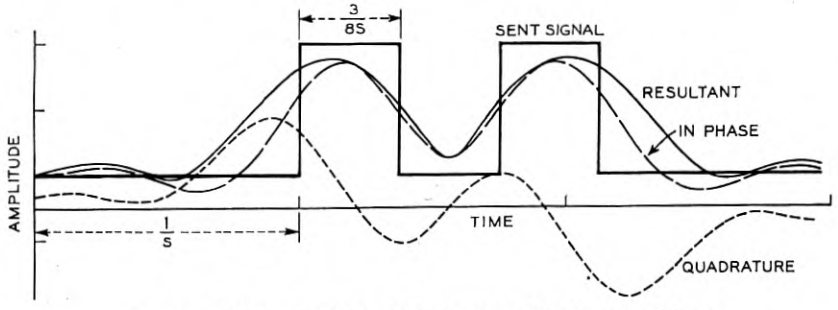


Fig. 12—Effect of adding dark current with shortened dots. (Compare with Figs. 11 and 10.)

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

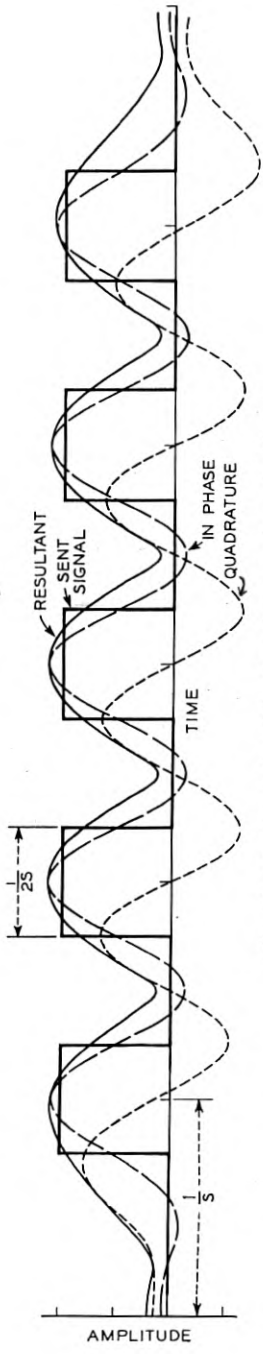


Fig. 13—Received signal for five dots, carrier as in Fig. 1.

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.

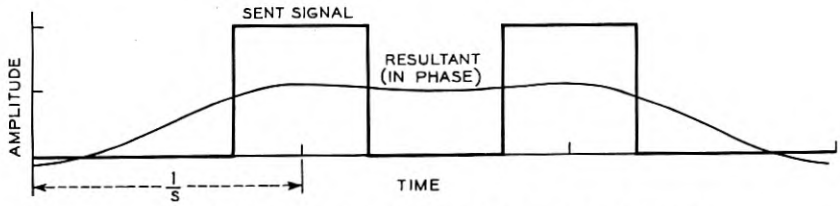


Fig. 14—Received signal for two dots, mid-band carrier.

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.

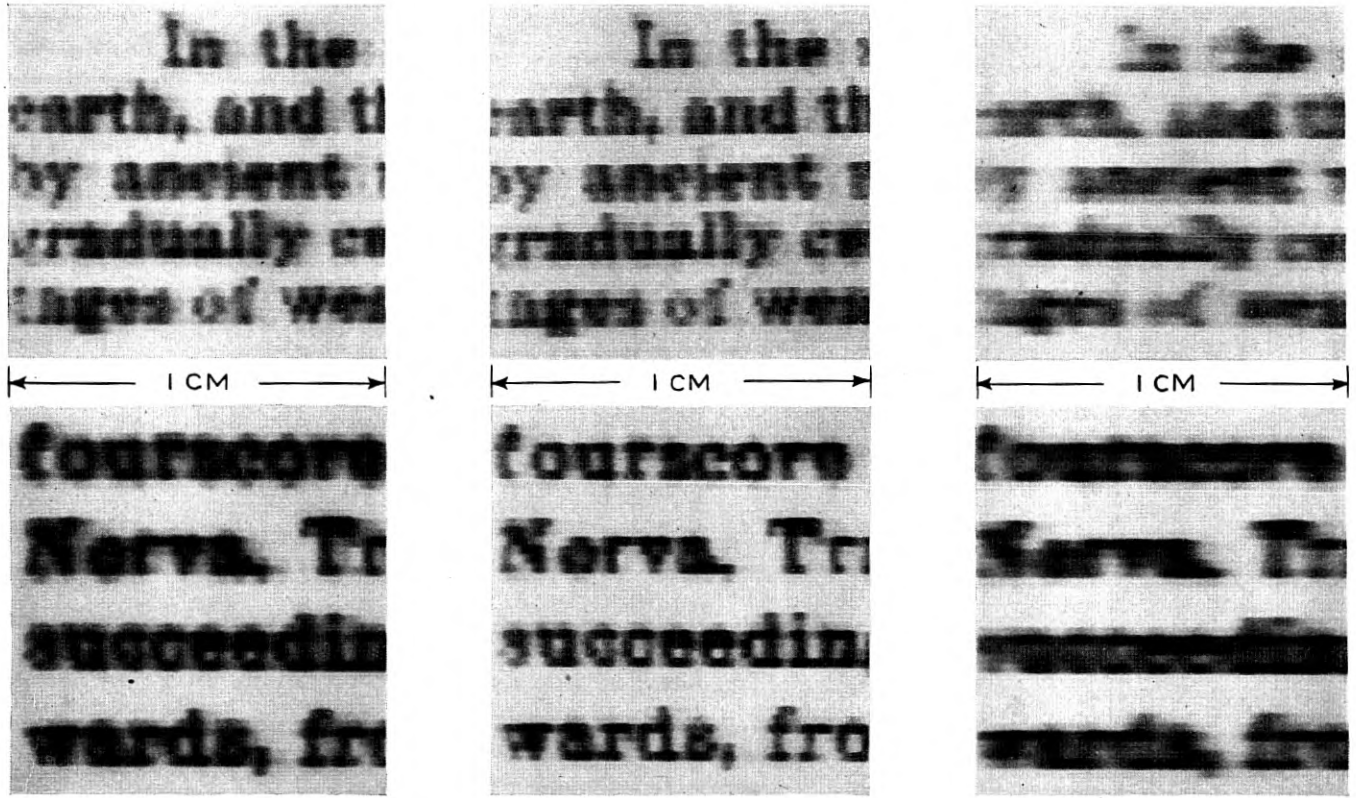


Fig. 15—Enlargements of transmitted printing. (*a*) Carrier at edge of band, dark current negligible. (*b*) Carrier at edge of band, dark current 14 db below maximum. (*c*) Carrier at center of band, dark current negligible. [The same transducer was used for (*a*), (*b*) and (*c*).]

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

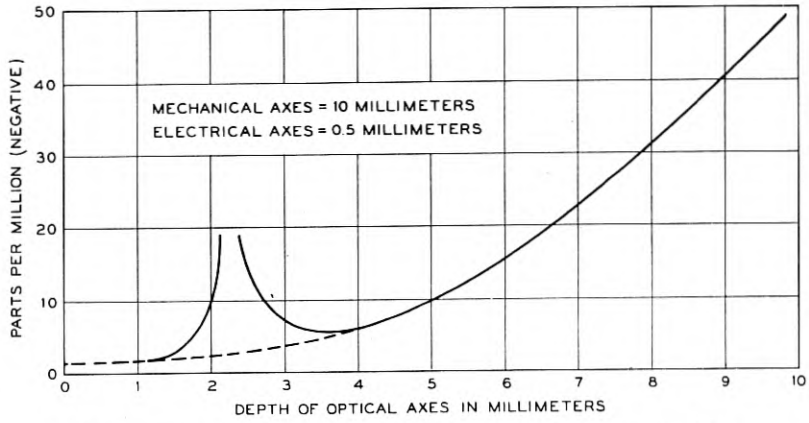


Fig. 1—Temperature coefficient of a perpendicularly cut crystal for varying ratios of width to length.

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 ± 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.

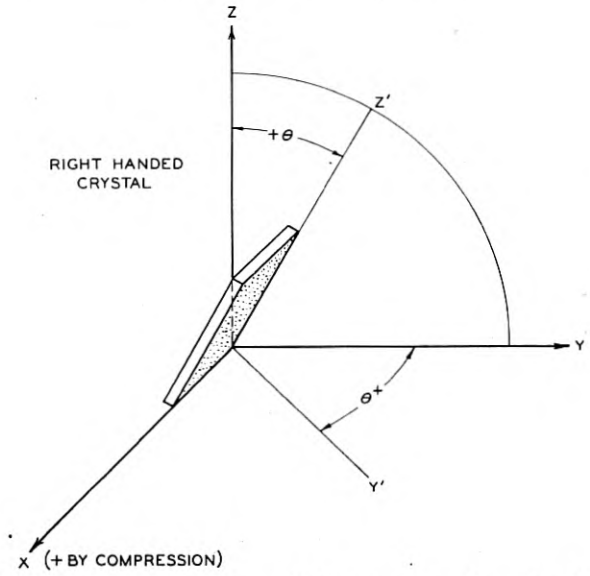


Fig. 2—Diagram illustrating angles used in expressing orientation of AT and BT plates within the natural crystal.

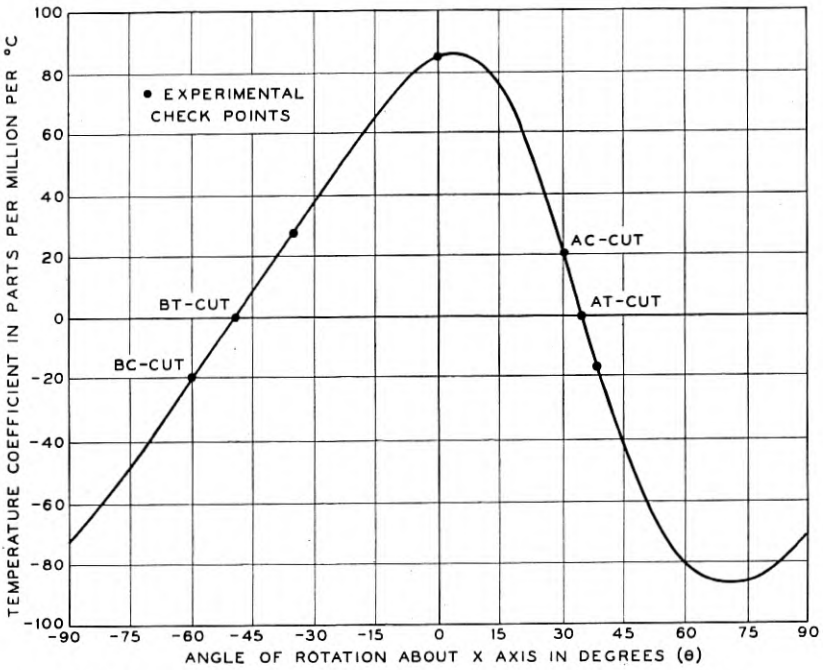


Fig. 3—Temperature coefficient for thin plates plotted as a function of the angle of cut.

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

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

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

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

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

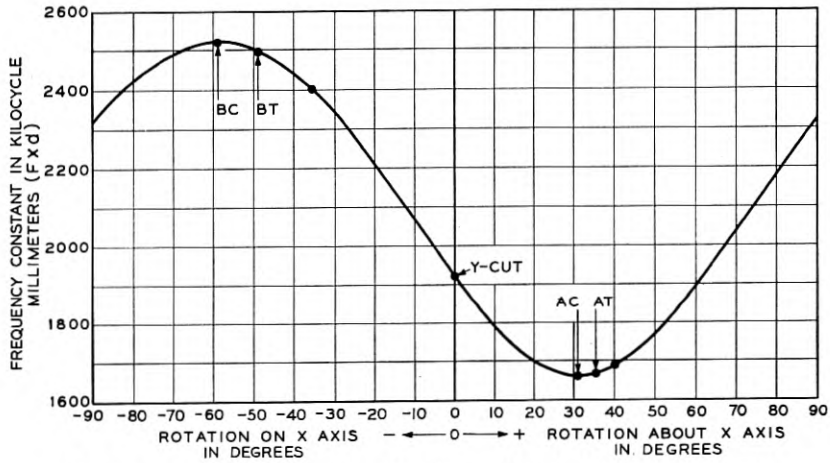


Fig. 4—Frequency constant for thin plates plotted against angle of cut.

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

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

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.

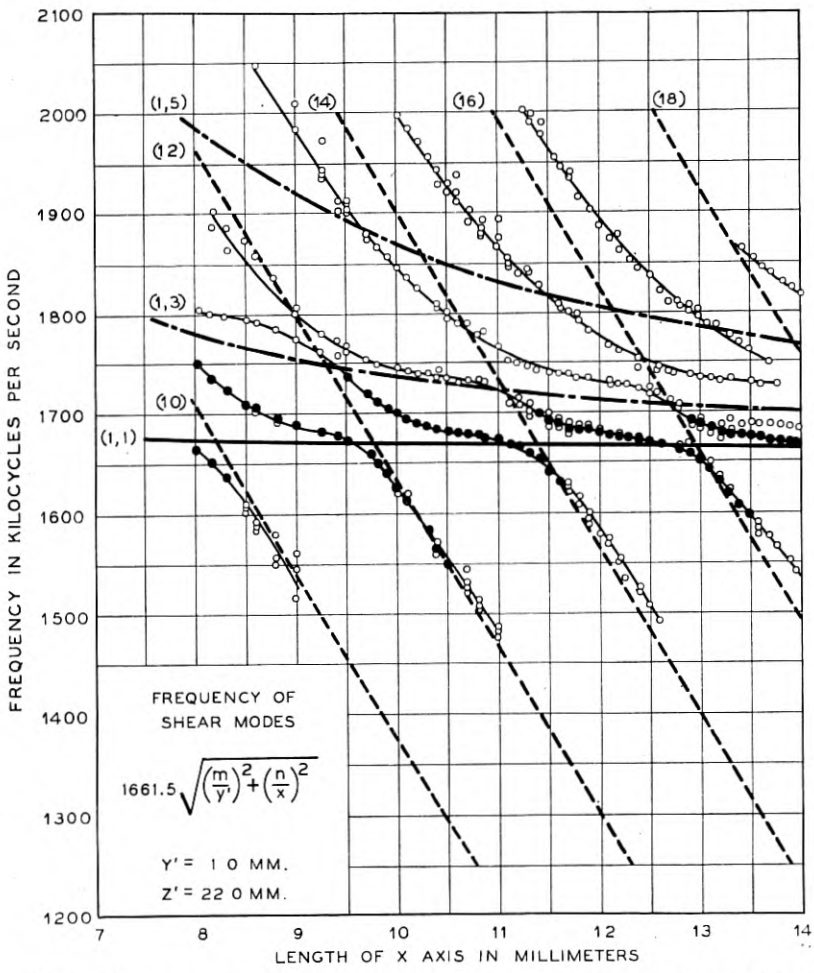


Fig. 5—Frequency spectrum (dots) for an AT plate as a function of ratio of length to thickness. The dashed lines represent calculated flexural vibrations. The whole line is the principal shear mode. The dot dash lines are other shear modes.

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

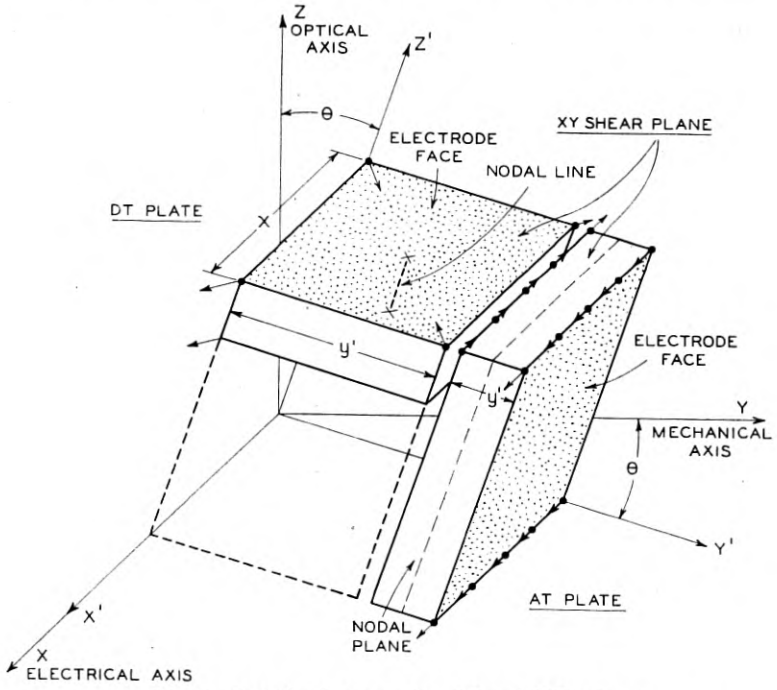


Fig. 6—Relation between the AT and DT cuts.

[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

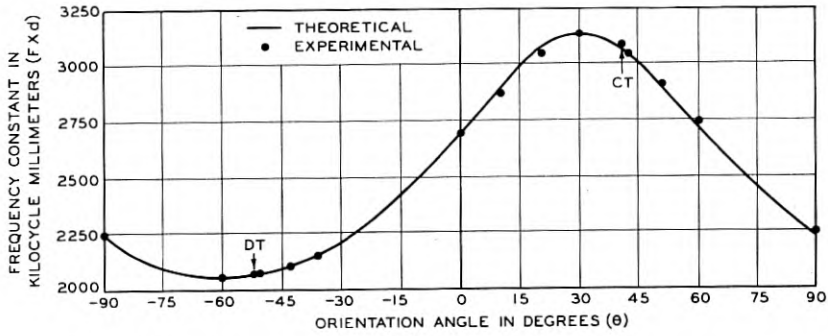


Fig. 7—Frequency constant for low-frequency shear crystal plotted against angle of cut.

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

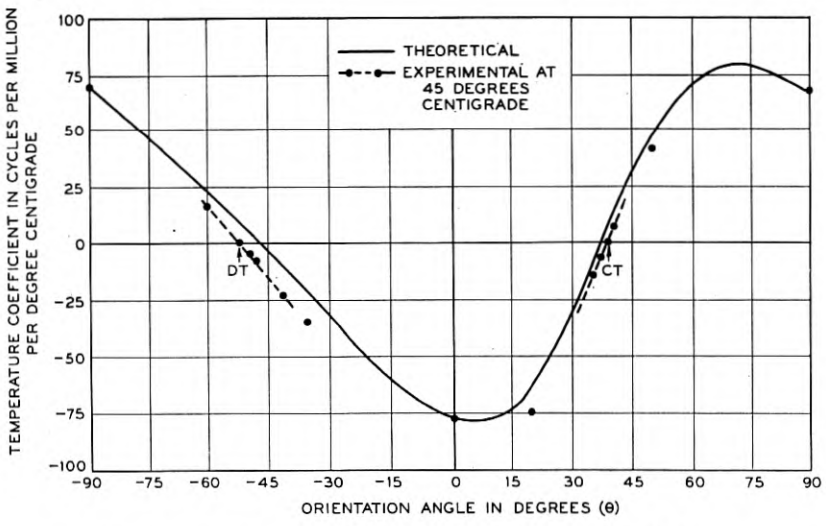


Fig. 8—Calculated and measured temperature coefficients for low-frequency shear crystals.

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.

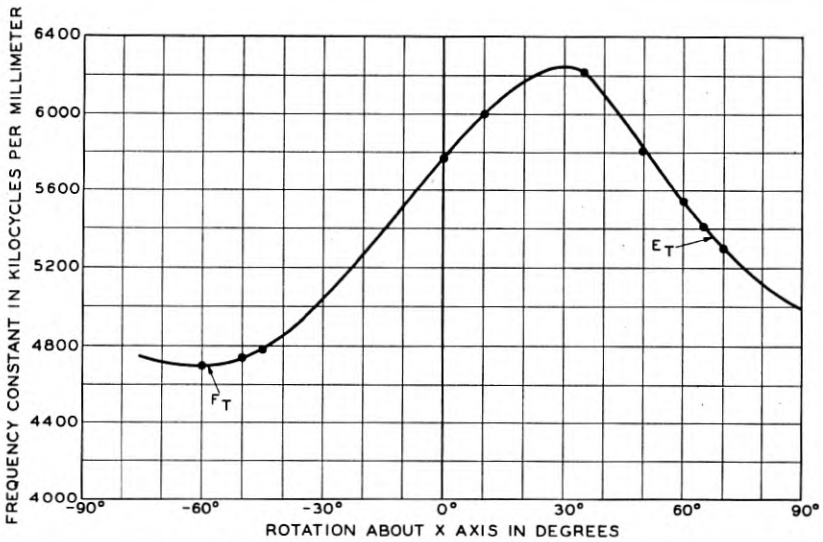


Fig. 9—Frequency constant for *E* and *F* type vibrations.

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

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

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

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

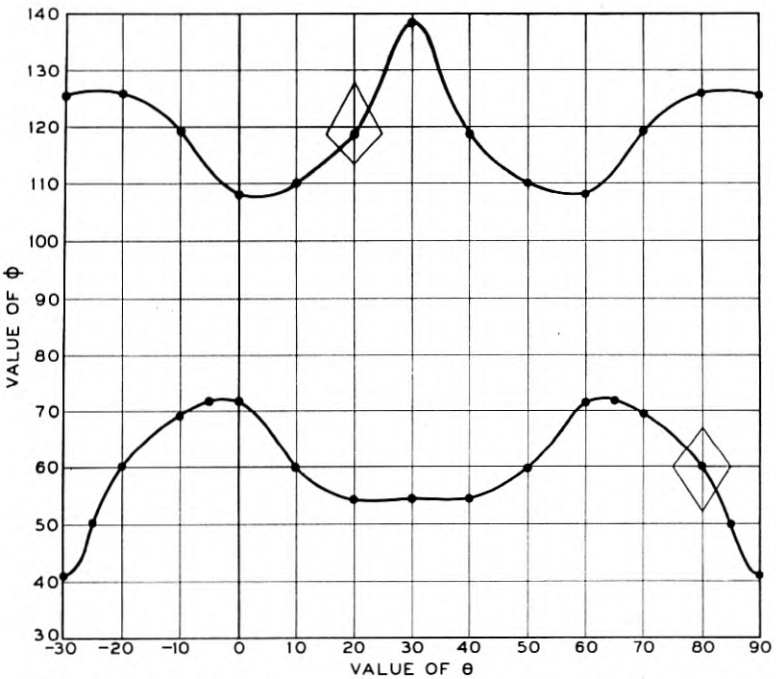


Fig. 10—Angles of cut for zero temperature coefficient high-frequency shear crystals for two rotations.

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.

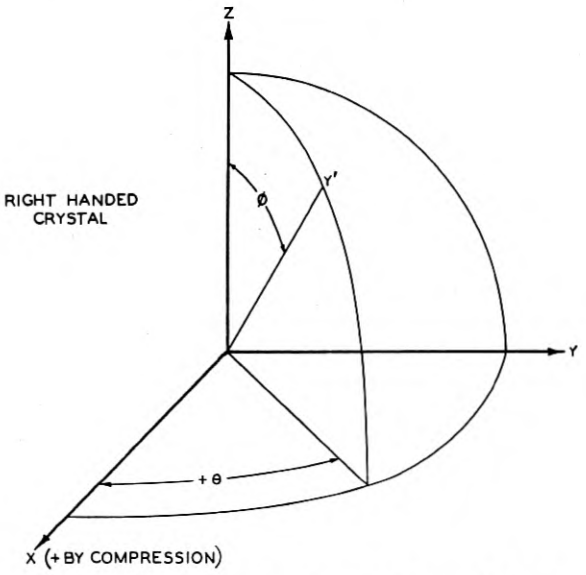


Fig. 11—Angular system for locating the axis of shear of high-frequency crystals with two rotations.

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

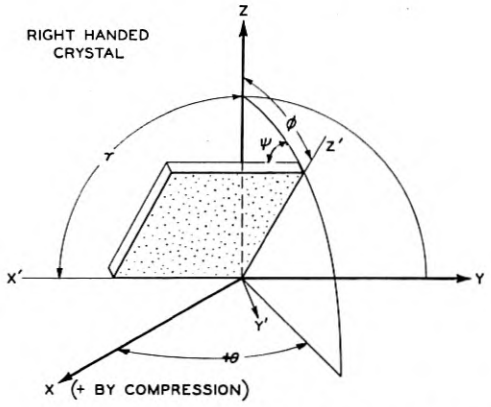


Fig. 12—Angular system for locating low-frequency shear crystals with three rotations.

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

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

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

$$T_f = \left[4.5 + 2.9 (\sin^2 \varphi \cos^2 \psi + \cos^2 \varphi) + \left[\frac{-5877.5 \cos^2 \psi \sin^2 2\varphi}{195 \cos^2 \psi \sin^2 2\varphi} \right. \right.$$
$$\frac{+ 15790 \sin^2 \varphi \sin^2 \psi + 10340 \sin \varphi [\sin 3\theta \cos \varphi (\cos^2 \psi \cos 2\varphi - \sin^2 \psi)}{+ 292.8 \sin^2 \varphi \sin^2 \psi - 172.4 \sin \varphi [\sin 3\theta \cos \varphi (\cos^2 \psi \cos 2\varphi - \sin^2 \psi)}$$
$$\frac{+ \cos 3\theta \sin \psi \cos \psi (\cos 2\varphi + \cos^2 \varphi)]}{+ \cos 3\theta \sin \psi \cos \psi (\cos 2\varphi + \cos^2 \varphi)]}$$
$$\left. \left. \frac{- 19{,}525 (\cos^2 \psi \cos^2 2\varphi + \sin^2 \psi \cos^2 \varphi)}{+ 200.5 \quad (\cos^2 \psi \cos^2 2\varphi + \sin^2 \psi \cos^2 \varphi)} \right] \right]. \quad (21)$$

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-

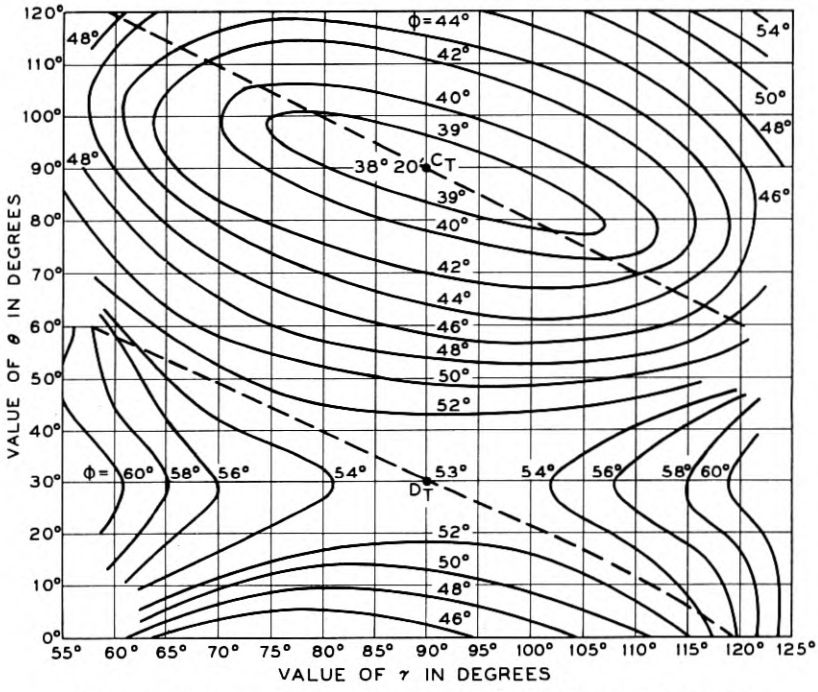


Fig. 13—Contour map of zero temperature coefficient low-frequency shear crystals with three rotations.

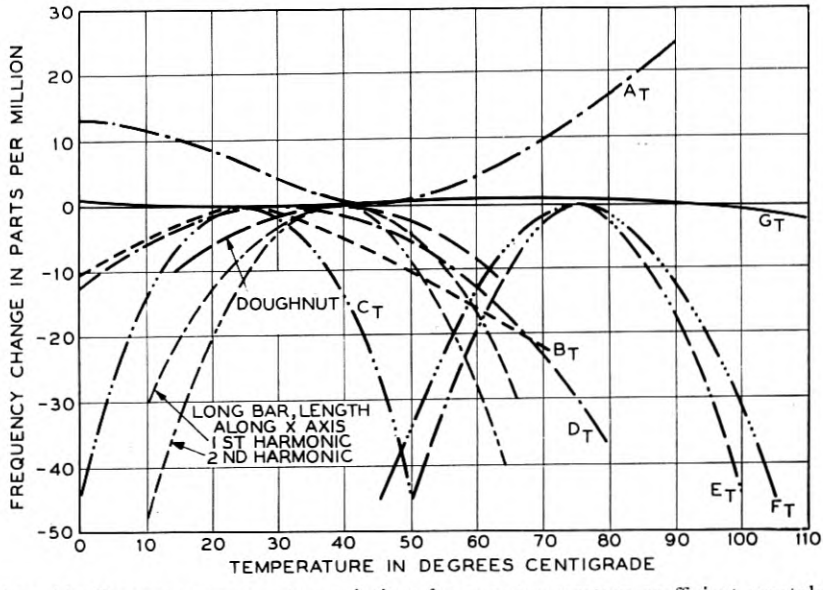


Fig. 14—Frequency temperature relations for zero temperature coefficient crystals.

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

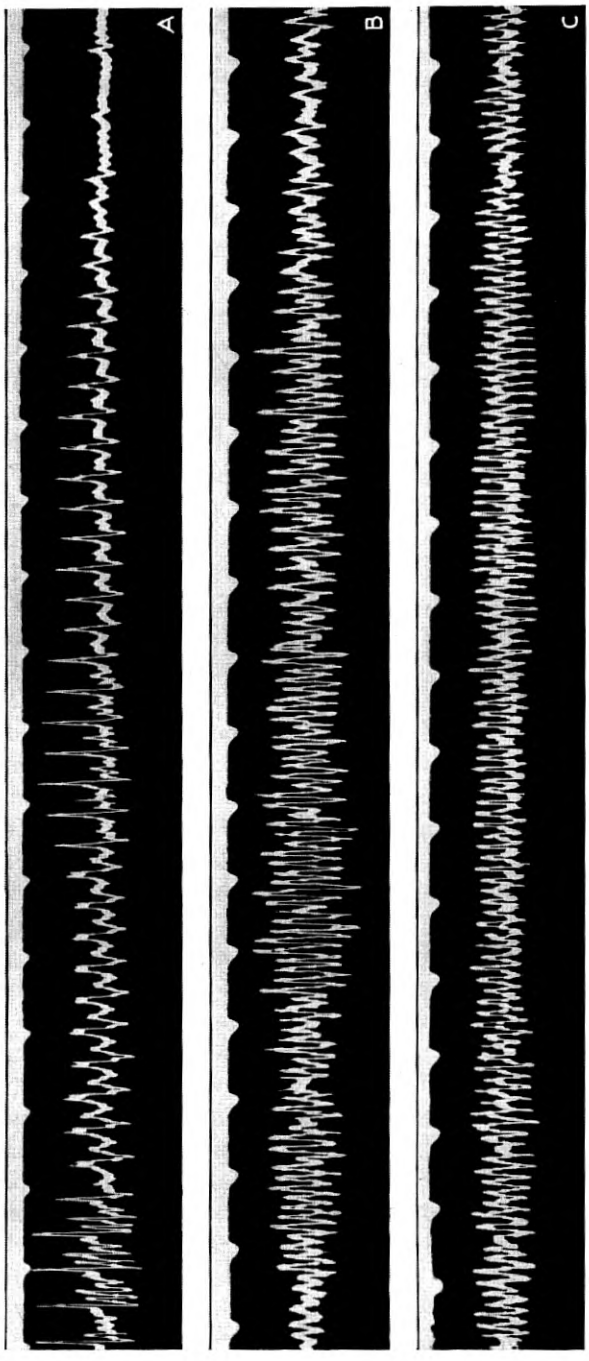


Fig. 1—Examples of program wave forms.

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.

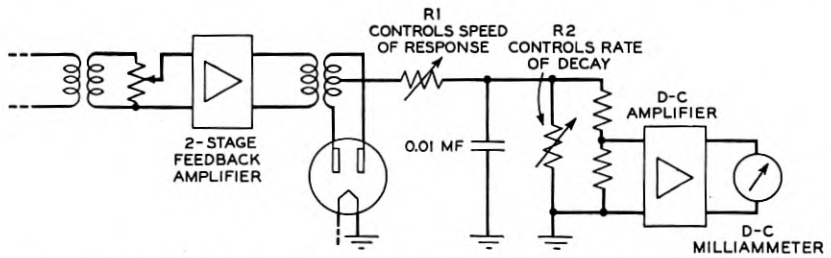


Fig. 2—Schematic diagram of experimental peak reading volume indicator.

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

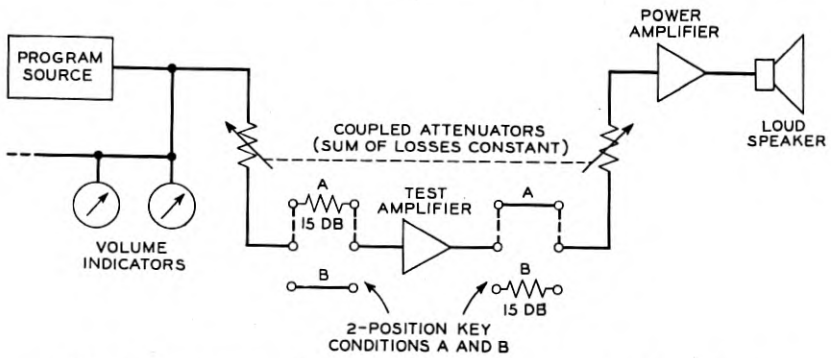


Fig. 3—Arrangements for determining volume level at which overload of amplifiers is audible.

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

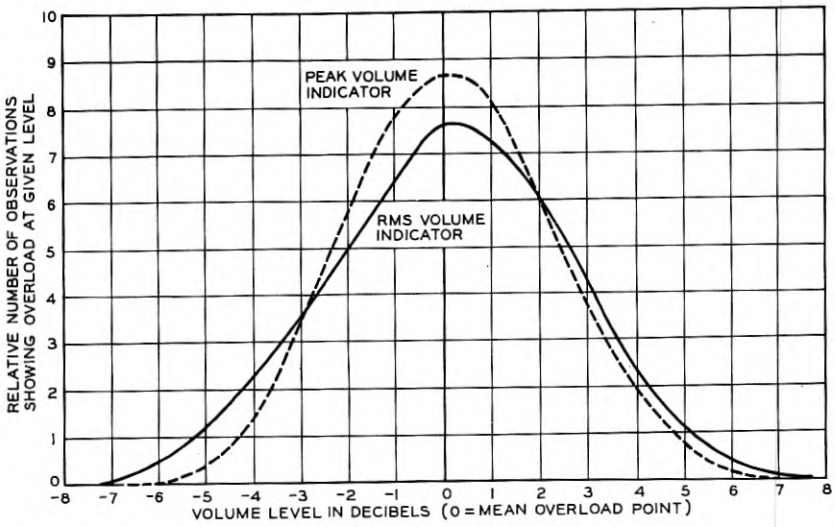


Fig. 4—Distribution of overload points.

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

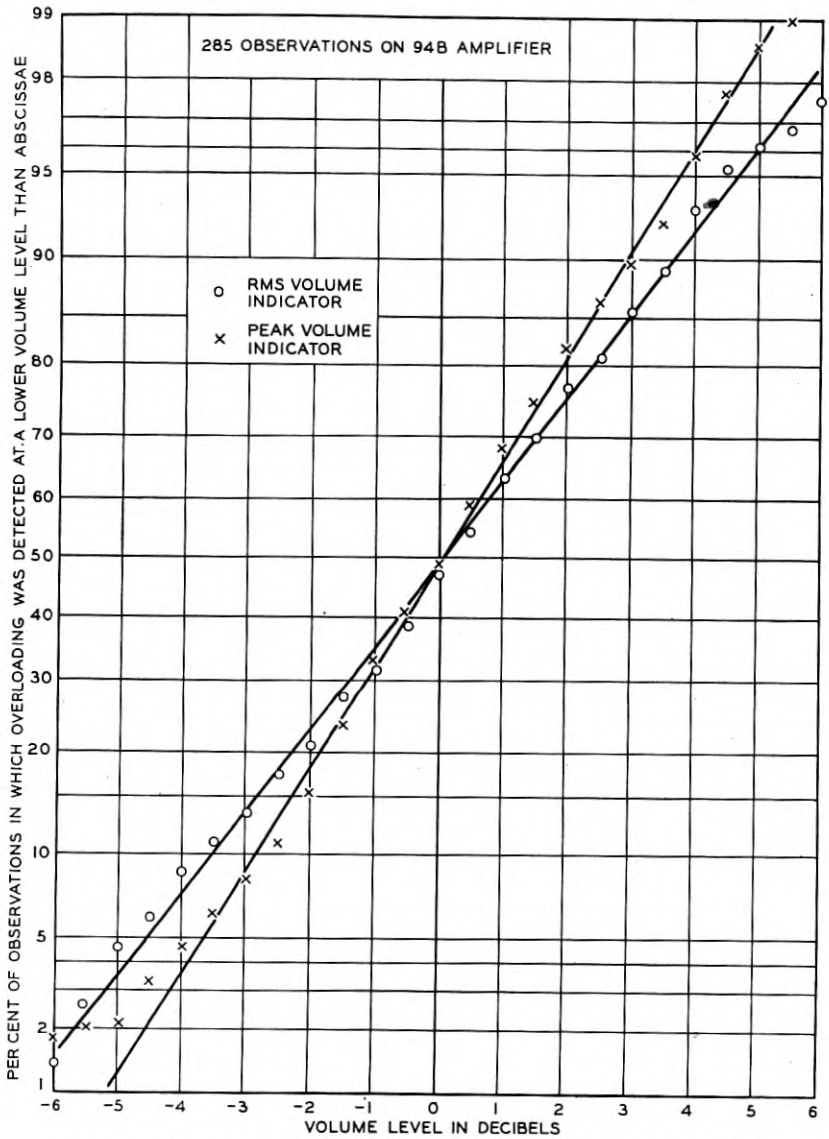


Fig. 5—Comparison of peak vs. r-m-s volume indicators as overload indicators (using W.E. 94B amplifier).

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

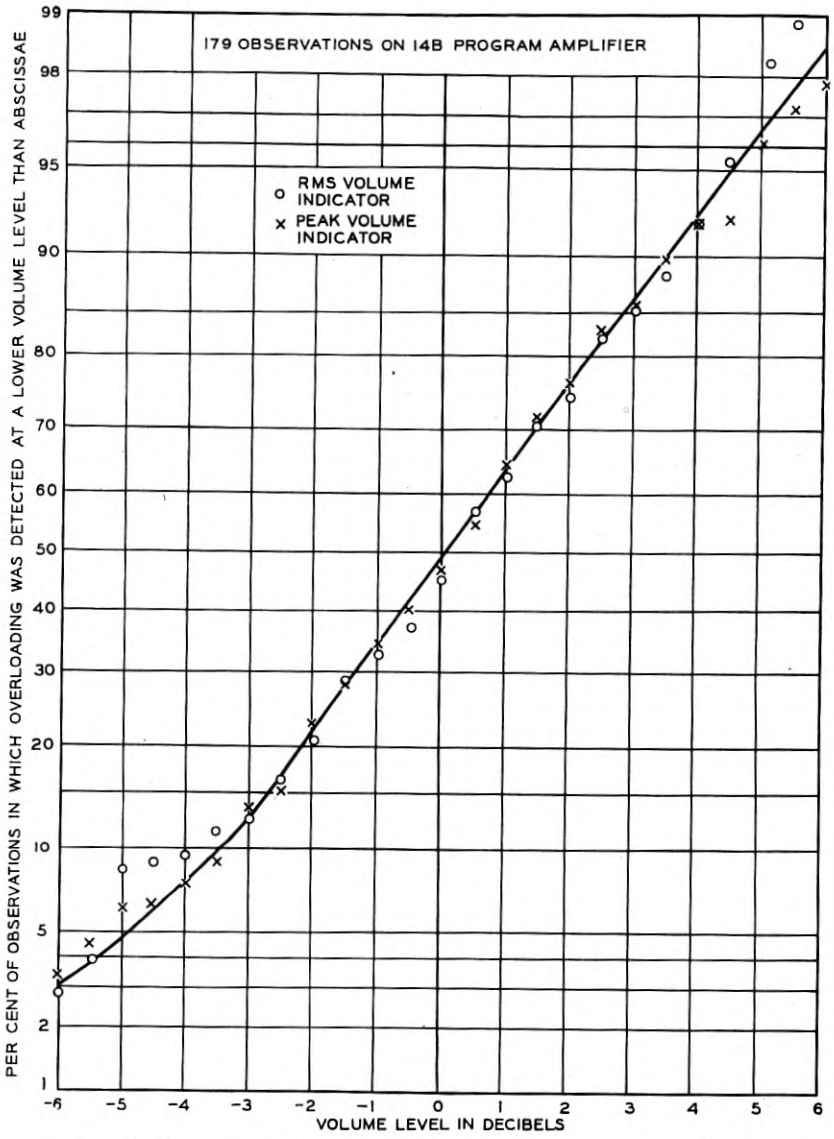


Fig. 6—Comparison of peak vs. r-m-s volume indicators as overload indicators (using W.E. 14B program amplifier).

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.

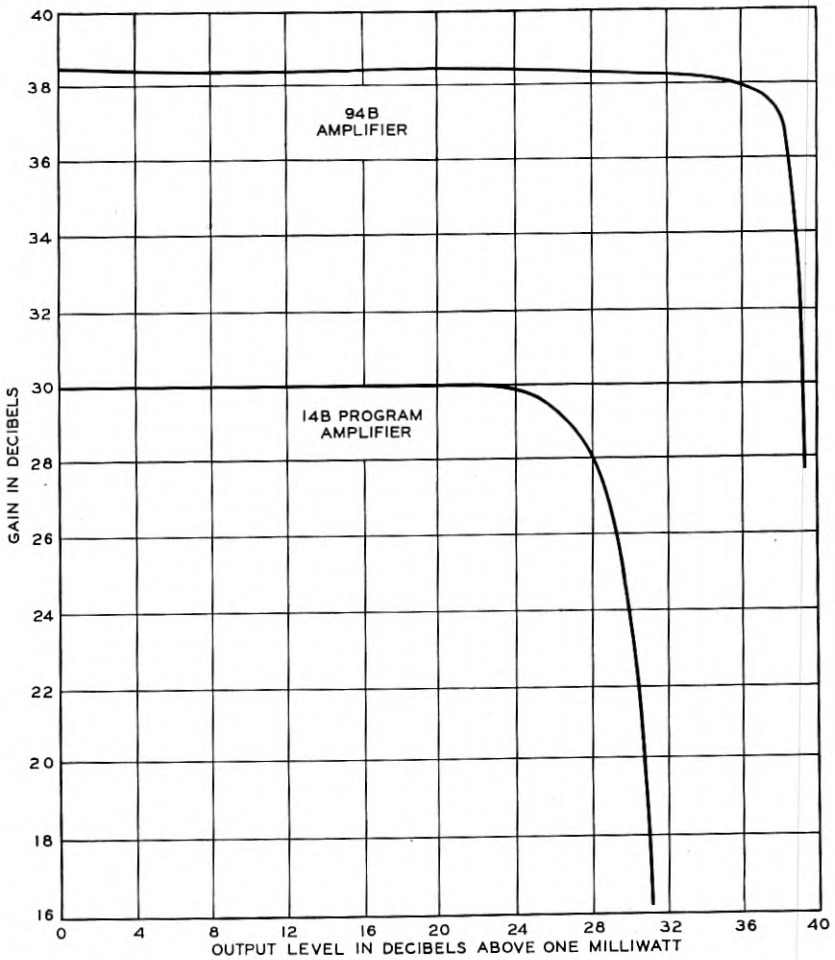


Fig. 7—Gain vs. load characteristics of amplifiers.

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

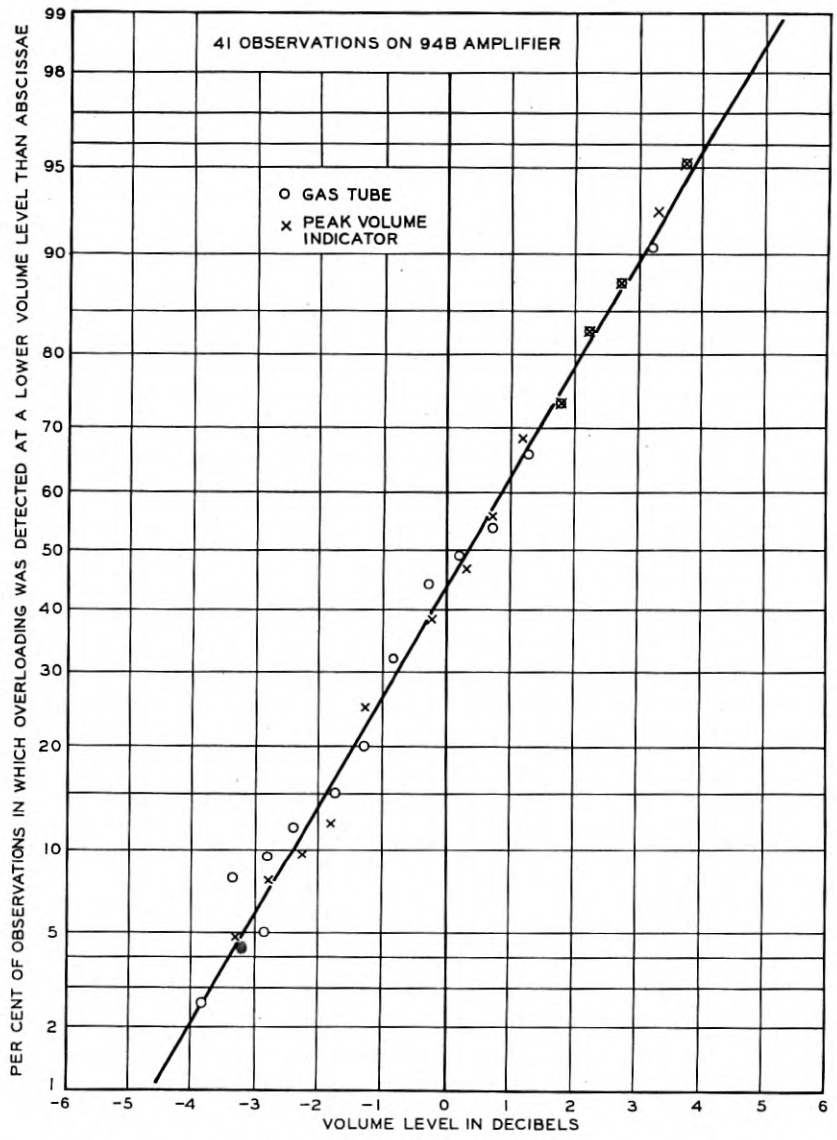


Fig. 8—Comparison of experimental peak volume indicator with gas tube trigger device as overload indicators.

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

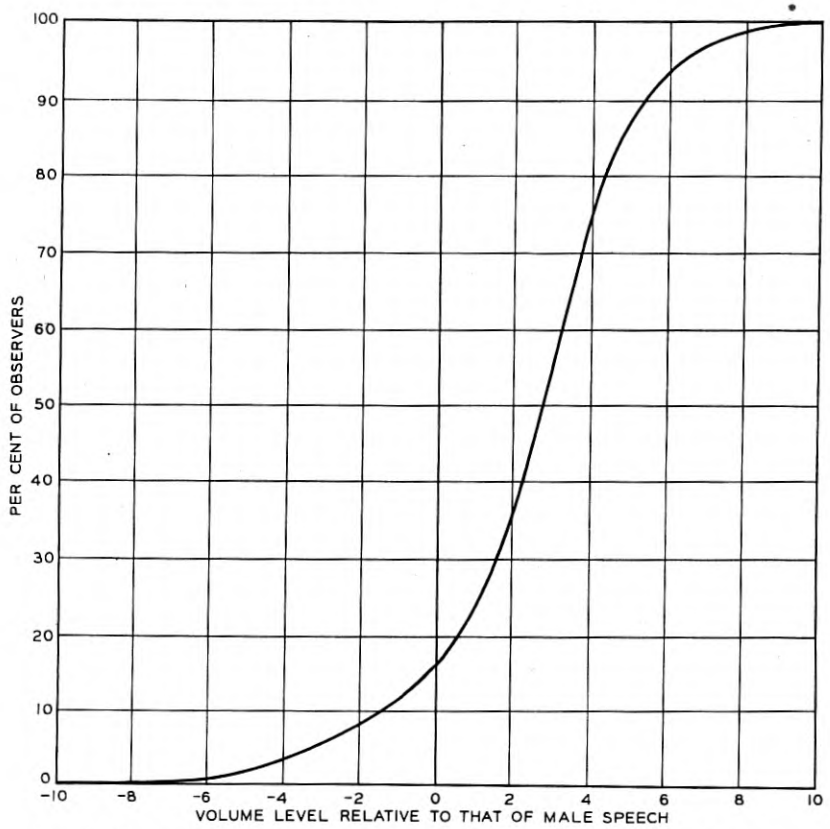


Fig. 9—Per cent of observers choosing symphony music at indicated volume levels to be louder than the male speech reference.

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

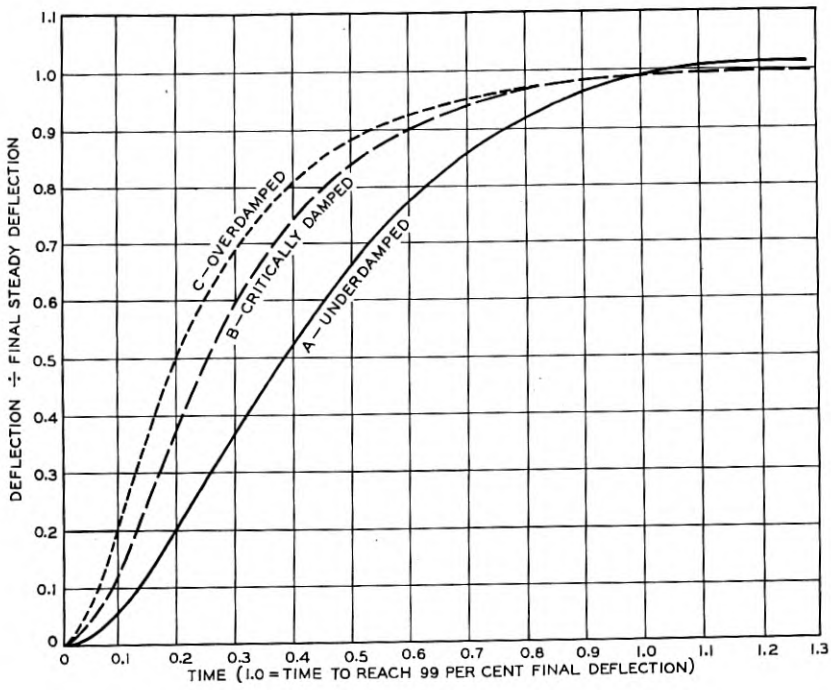


Fig. 10—Effect of damping on instrument characteristics.

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.

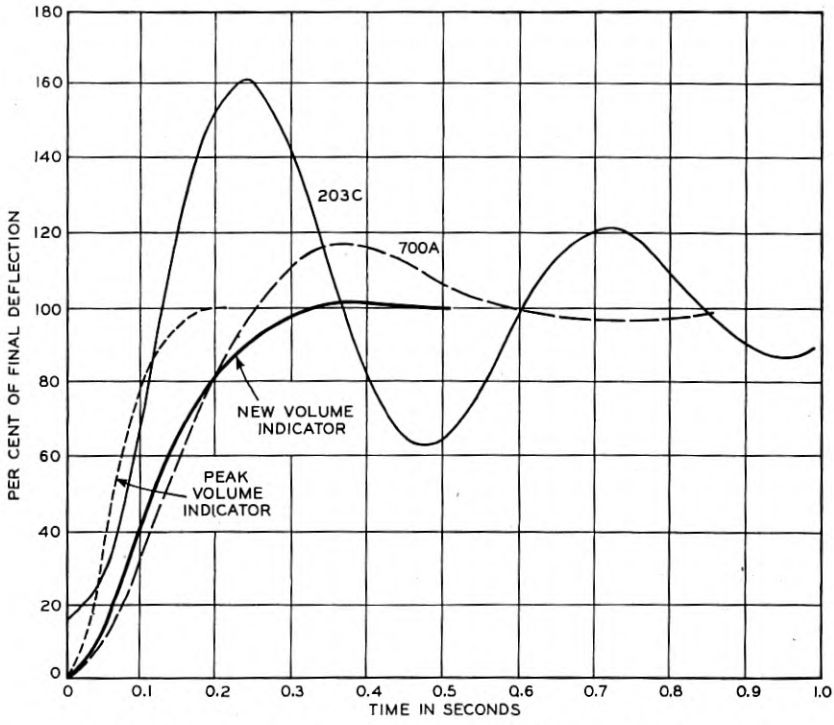


Fig. 11—Deflection of volume indicators to suddenly applied sine-wave.

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.

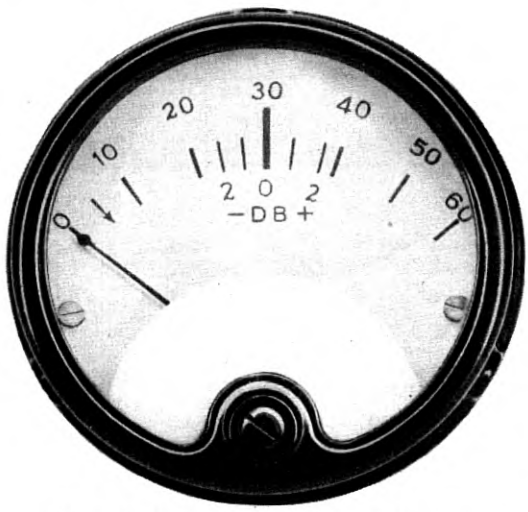


Fig. 12—Scale on 203C volume indicator.

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

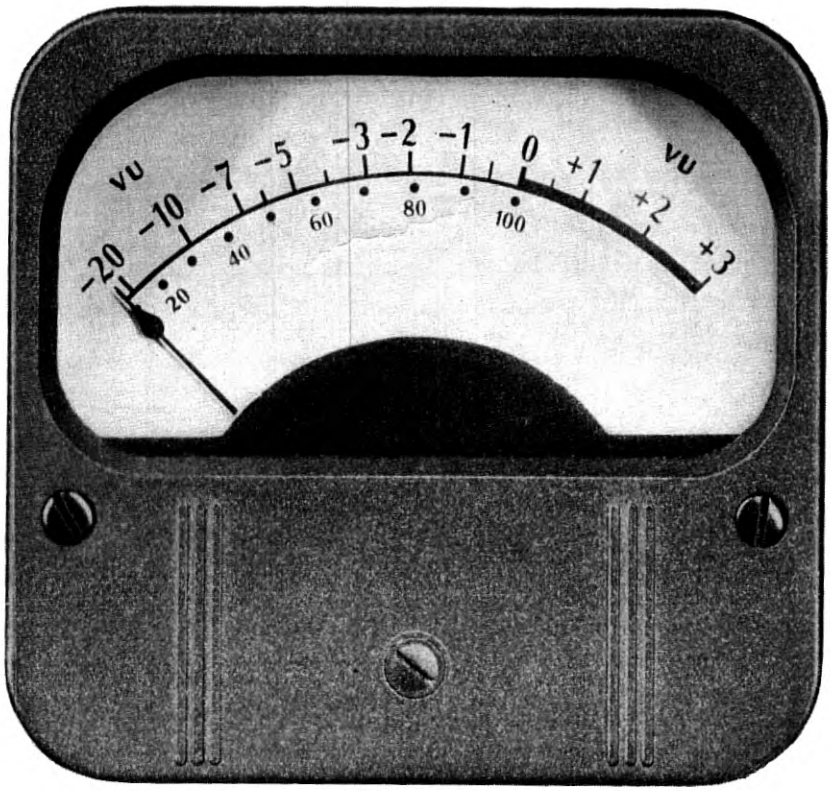


Fig. 16—New volume indicator—A scale.

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.

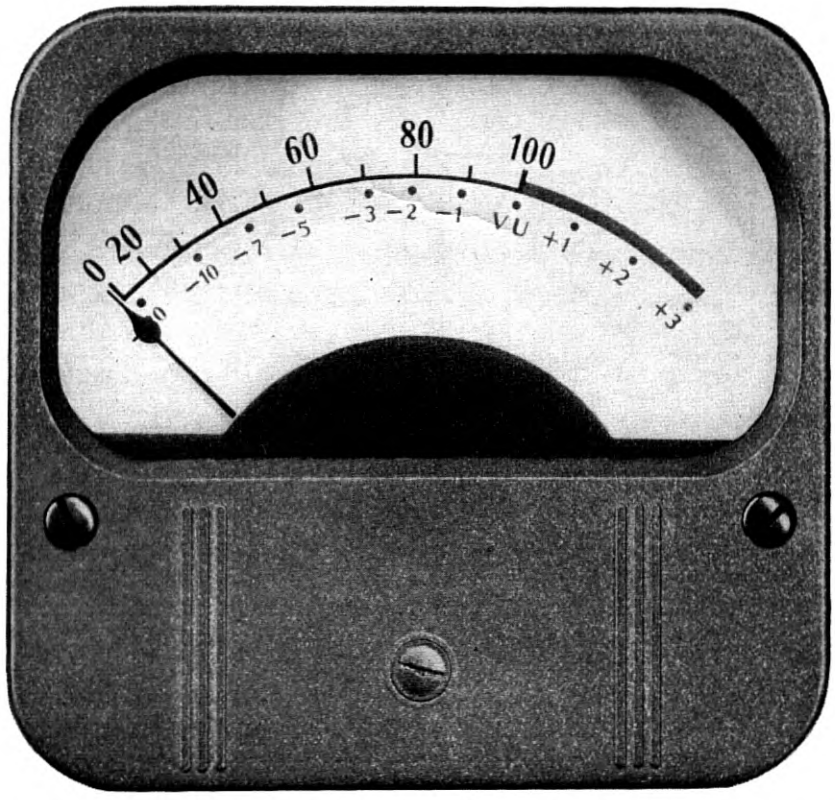


Fig. 17—New volume indicator—B scale.

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.

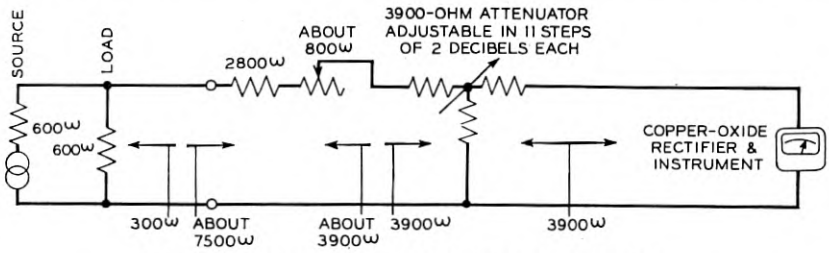
[8] Munsell Book of Color, Munsell Color Company, Baltimore, Maryland, 1929.

8. Temperature Effects

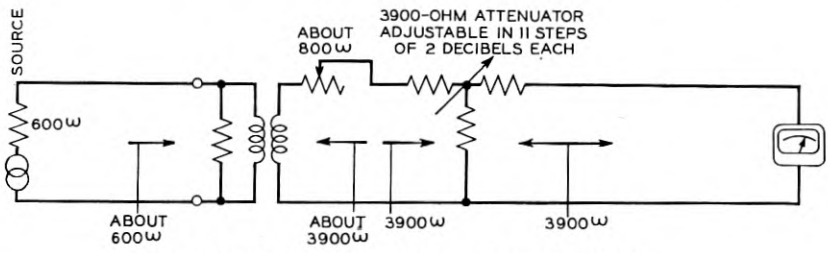
In the instruments now available, the deviation of the sensitivity with temperature is less than 0.1 decibel for temperatures between 50° F. and 120° F., and is less than 0.5 decibel for temperatures as low as 32° F.

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.



A. HIGH-IMPEDANCE ARRANGEMENT. RANGE +4 TO +26 VU (FOR A DEFLECTION TO THE 0 VU OR 100 MARK)



B. LOW-IMPEDANCE ARRANGEMENT. RANGE -6 TO +16 VU (FOR A DEFLECTION TO THE 0 VU OR 100 MARK)

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

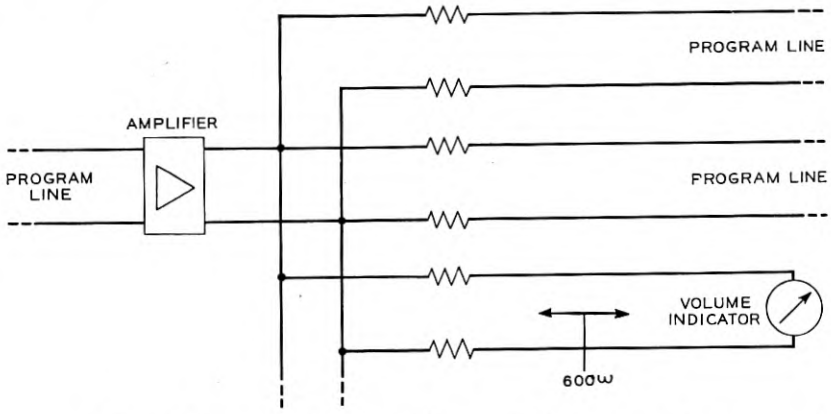


Fig. 19—Program bridge for feeding several lines from one line.

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 crcuit, 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.

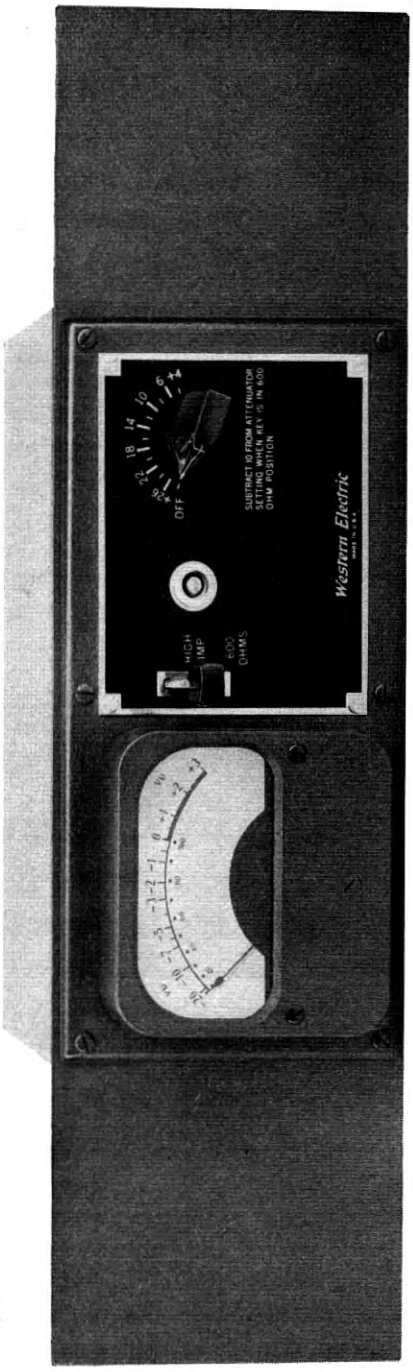


Fig. 20—754B volume indicator equipped with new standard instrument having "A" (Bell System) scale.

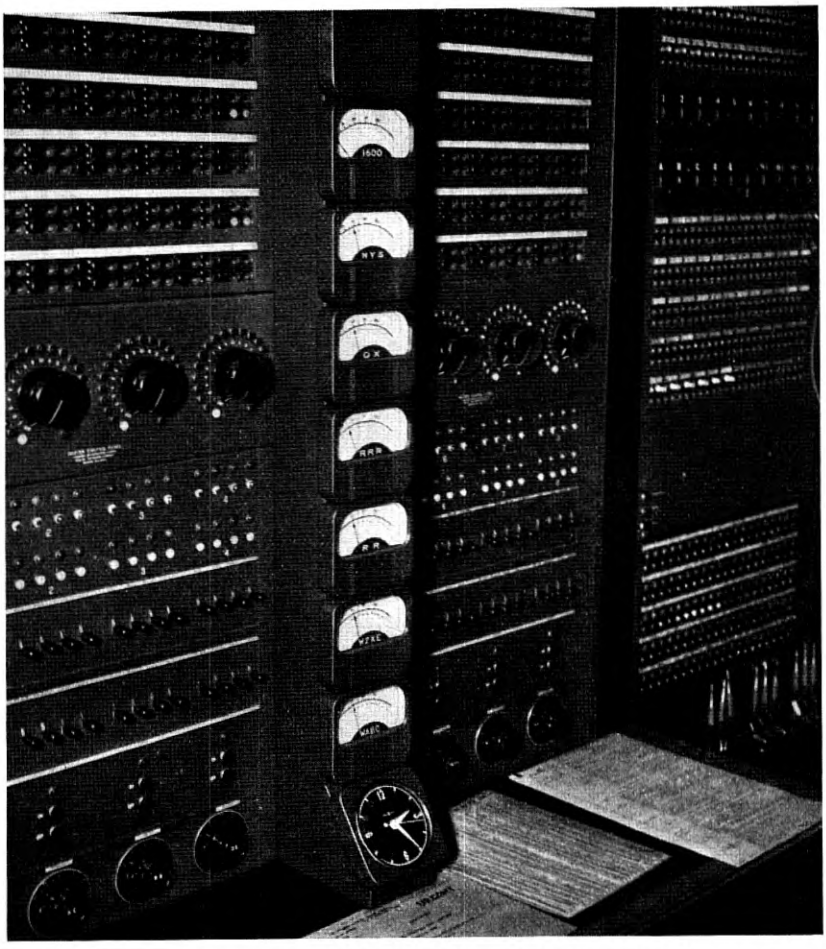


Fig. 21—New standard volume indicators installed at a network key station.

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.

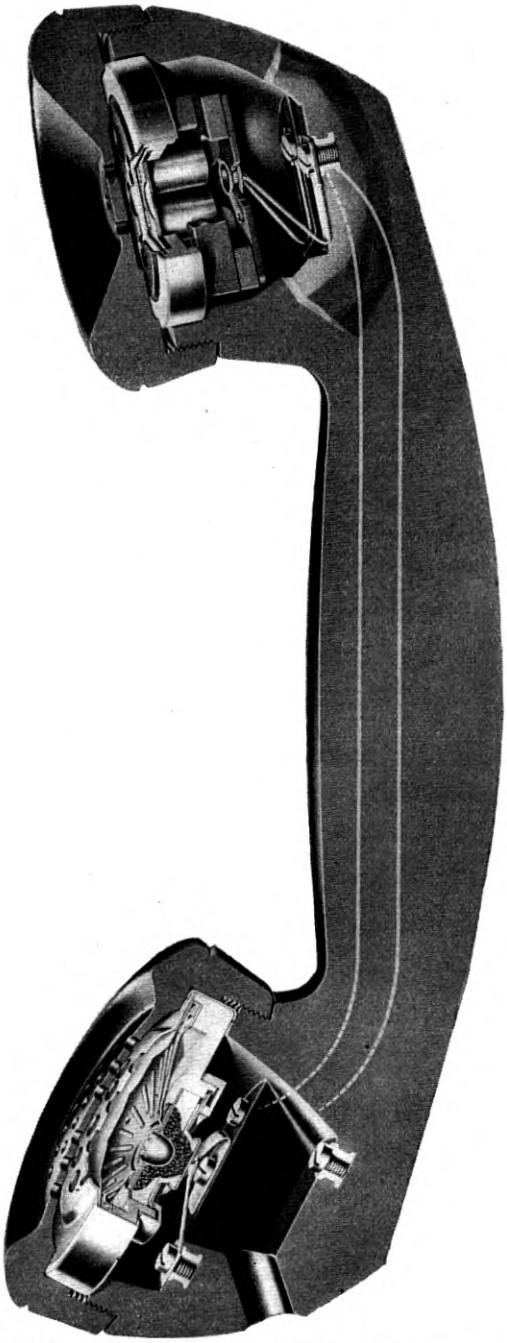


Fig. 1—Schematic cross-section of handset showing utilization of metallic materials.

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

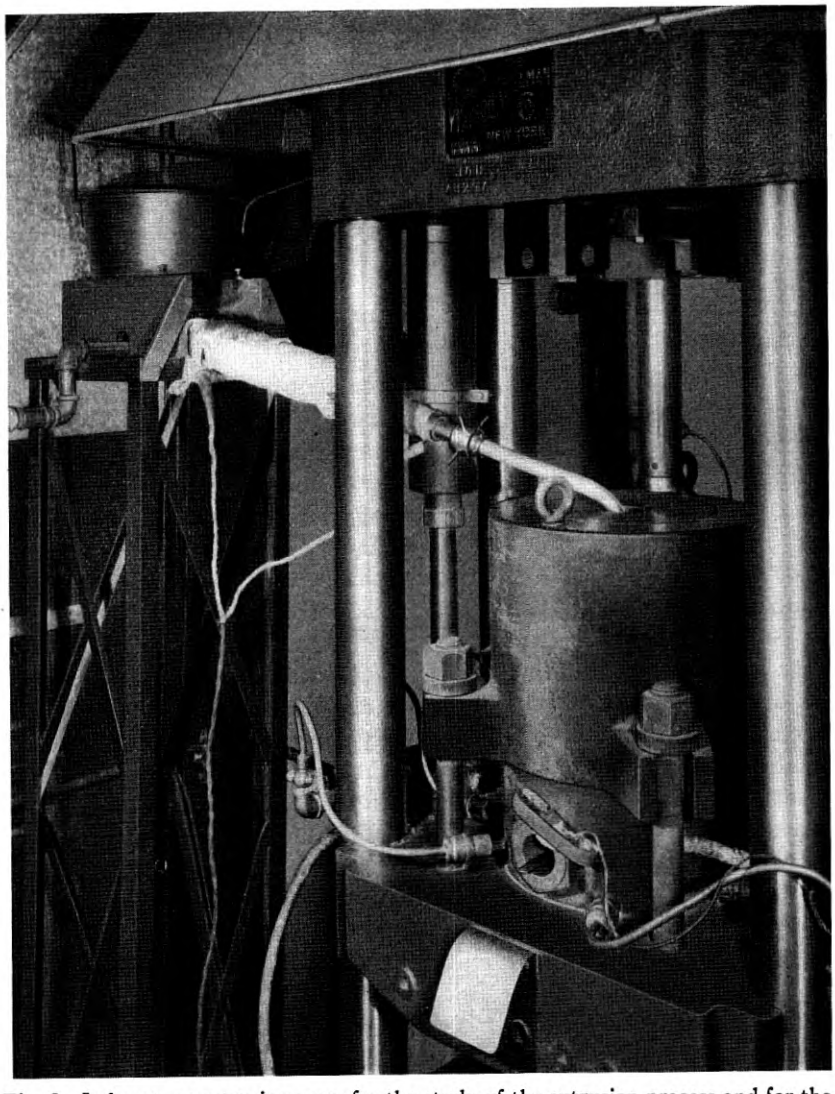


Fig. 3—Laboratory extrusion press for the study of the extrusion process and for the production of experimental cable sheathing alloys.

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

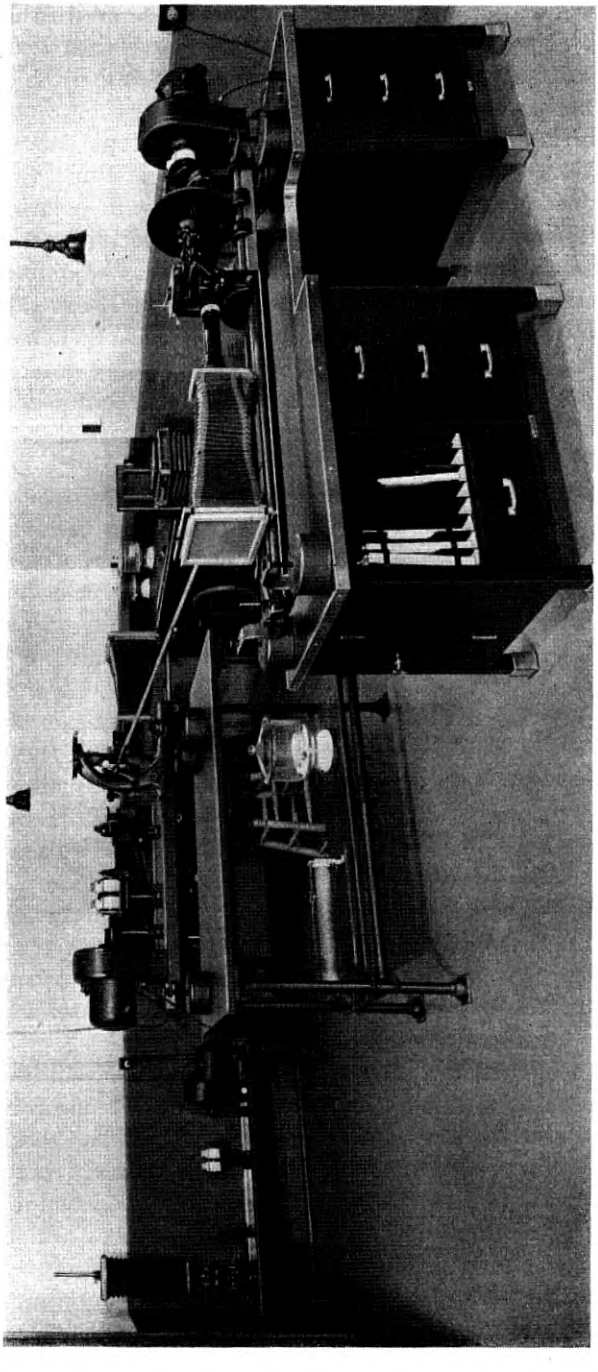


Fig. 4—Metallographic equipment used in the study of metal problems.

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.

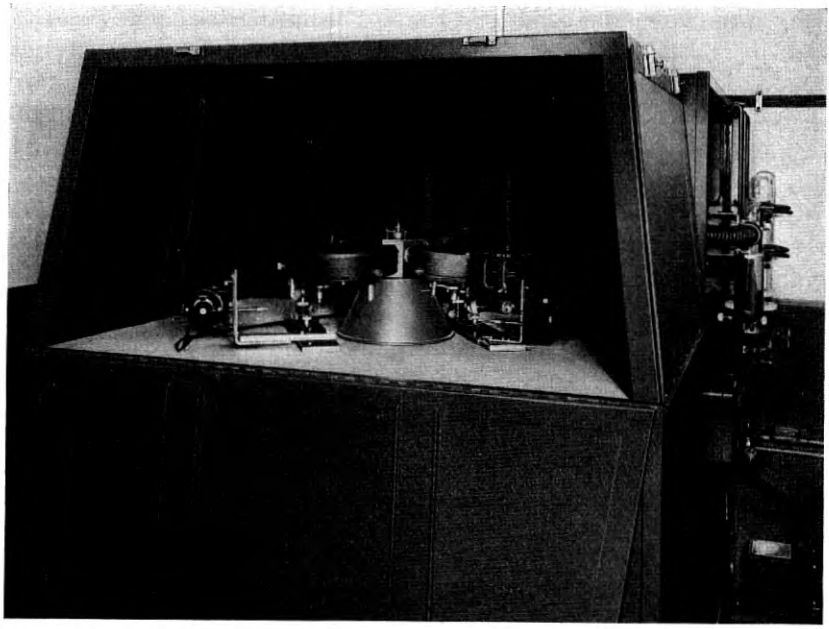


Fig. 5—X-ray diffraction apparatus showing cameras mounted for identification of structure and precision measurement of lattice constant.

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.

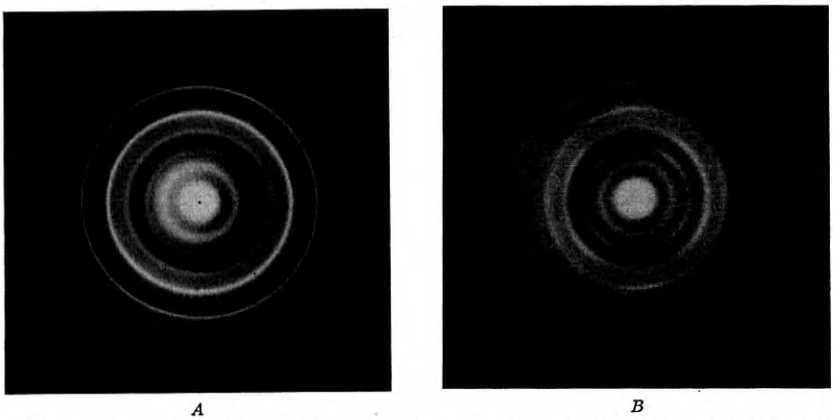
Silica films for these studies were prepared in the following manner. A glass microscope slide was first covered by gold vaporized in high vacuum from a V-shaped tungsten ribbon; then immediately in the same apparatus silica was vaporized upon the gold from a second tung-

* 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 $\times$ 10^{-7} cm) in directions parallel to the surface, and thin (2 $\times$ 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

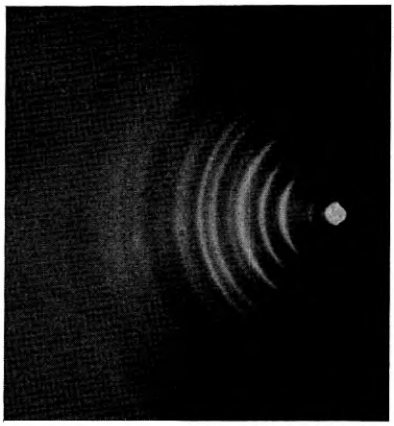


Fig. 2—Electron diffraction pattern obtained by the reflection method from finely pulverized aluminum alpha-monohydrate ($Al_2O_3 \cdot H_2O$).

as thick as 5 $\times$ 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.

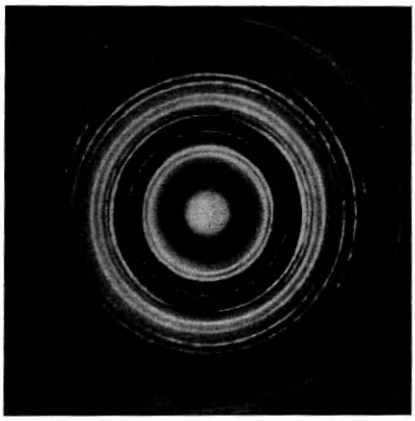


Fig. 3—Electron diffraction pattern produced by a layer of oriented crystals of an aluminum soap, which had been formed as scum upon a water surface as a result of the reaction of powdered aluminum, water and traces of organic acid present as an impurity.

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.