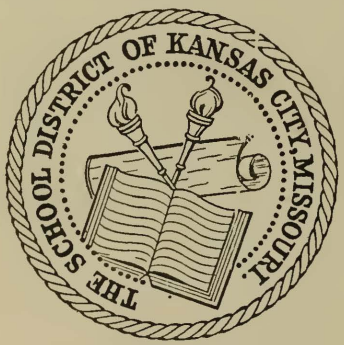
Bound
Periodical
590204

Kansas City Public Library



This Volume is for

REFERENCE USE ONLY

7 27—6m—P

From the collection of the



San Francisco, California
2008

VIRGINIA
TECH

PUBLIC LIBRARY

THE BELL SYSTEM TECHNICAL JOURNAL

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

EDITORIAL BOARD

J. J. CARTY BANCROFT GHERARDI F. B. JEWETT
E. B. CRAFT L. F. MOREHOUSE O. B. BLACKWELL
H. P. CHARLESWORTH E. H. COLPITTS H. D. ARNOLD
R. W. KING—*Editor* J. O. PERRINE—*Asst. Editor*

VOLUME VI
1927

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PUBLIC LIBRARY
KANSAS CITY
MO

Bound
Periodical

The Bell System Technical Journal

January, 1927

Electromagnetic Theory and the Foundations of Electric Circuit Theory [1]

By JOHN R. CARSON

SYNOPSIS: The familiar equations which are used to solve for the currents and charges in linear networks summarize the inductive analysis of countless observations made upon such networks. Having been arrived at by inductive methods, these familiar equations of Ohm, Faraday and Kirchhoff are substantially independent of the more general electromagnetic theory of Maxwell and Lorentz. The present paper examines the foundation of electric circuit theory from the standpoint of the fundamental equations of electromagnetic theory and a derivation of the former from the latter is made, in the course of which the assumptions, approximations and restrictions tacitly involved in the equations of circuit theory are explicitly stated. The treatment is sufficiently extended as to show how the familiar equation for the simple oscillating circuit and the so-called telegraph equation can be deduced from the Maxwell-Lorentz statement of electromagnetic theory.

ELECTRIC circuit theory, as the term is employed in the present paper, is that branch of electromagnetic theory which deals with electrical oscillations in linear networks; more precisely stated, with the distribution of currents and charges in the free oscillations of the network, or under the action of impressed electromotive forces. The network is a connected set of closed circuits or meshes each of which is regarded as made up of inductances, resistances and condensers, a simplifying assumption which is fundamental to circuit theory.

The great importance of electric circuit theory in electro-technics does not require emphasis; it is not too much to say that it is responsible in no small measure for the rapid development of electrical engineering and is an absolutely essential guide in the complicated technical problems there encountered.

The equations of electric circuit theory in their present form are essentially a generalization of the observations of Ohm, Faraday, Henry, Kirchhoff and others and their development preceded the electromagnetic theory of Maxwell and Lorentz. Naturally, in view of its early development, circuit theory embodies approximations, the precision of which cannot be determined from the observations on which it is based. For example, circuit theory explicitly ignores the finite velocity of propagation of electromagnetic disturbances, and

[1] In its original form this paper was read before the National Academy of Sciences, April 1925. Subsequently it was amplified and revised and included in a lecture course delivered at the Massachusetts Institute of Technology, April 1926.

hence the phenomena of radiation. Again it involves the assumption that the network can be represented by a finite number of coordinates and thus that it constitutes a rigid dynamic system. The rigorous equations of electromagnetic theory formulate the relations between current and charge *densities* and the accompanying fields. Circuit theory, on the other hand, expresses approximate relations between total currents and charges and impressed electromotive forces.

With the rapid development of electro-technics an increasing number of problems is being encountered where the application of classical electric circuit theory is of doubtful validity, or where the conclusions derived from it must be interpreted with great care. Such problems are the result not only of the use of very high frequency in radio-transmission but arise also in connection with the need of a more precise theory of wire transmission.

In view of the foregoing it seems desirable to examine the foundations of circuit theory. This is the problem dealt with in the present paper:—a derivation of the classical circuit theory equations from the standpoint of electromagnetic theory, in the course of which the approximations involved are pointed out.

A second reason, pedagogic in character, is believed to justify the present study. This is, that, as circuit theory is usually taught to technical students no picture of its true relation to electromagnetic theory is given, and the student comes to regard inductance, resistance, capacity, voltage, etc., as fundamental concepts.

To start with our problem in a general form, consider a conducting system of any form whatsoever, in which the *charge density* at any point x, y, z at any time t is denoted by

$$\rho(x, y, z, t) = \rho,$$

and the *vector current density* by

$$\boldsymbol{u}(x, y, z, t) = \boldsymbol{u},$$

the functional notation indicating that the charge and the current density are functions of space and time. At any point in the system let

$$\boldsymbol{E}(x, y, z, t) = \boldsymbol{E}$$

denote the *vector electric intensity*. This we shall suppose to be composed of two parts; thus

$$\boldsymbol{E} = \boldsymbol{E}^{\circ} + \boldsymbol{E}'. \tag{1}$$

In this equation $\boldsymbol{E}^{\circ}$ is the *impressed electric intensity* and $\boldsymbol{E}'$ the *electric intensity due to the reaction of the currents and charges in the system*. Thus $\boldsymbol{E}^{\circ}$ may be the electric intensity due to a distant system, as in radio transmission, or that due to a generator, battery or other

source of energy. In the following we shall suppose that E° is specified and we shall keep carefully in mind the fact that E° denotes the electric intensity *not* due to the reaction of the system itself. This distinction is extremely important.

We have now to take up the problem of specifying the electric intensity E' in terms of the currents and charges of the system. The necessary relation is furnished by the *Lorentz or retarded potentials*

$$\Phi = \int \frac{\rho(t - r/c)}{r}\, dv, \qquad \text{(scalar)} \tag{2}$$

$$A = \int \frac{u(t - r/c)}{r}\, dv, \qquad \text{(vector).} \tag{3}$$

Interpreting equation (2), Φ is equal to the volume integral of the *retarded* charge density divided by the distance between the point at which Φ is evaluated and the location of the charge. The *retarded* charge density means that at time t we take the value of the charge at the earlier time $t - r/c$, where c is the velocity of light. It is to be understood that ρ and u are the true charge and current density, and displacement currents are not included. Their effect appears in the retardation only. c also is the true velocity of propagation in vacuo. The potential Φ is therefore a generalization of the electrostatic potential into which it degenerates in an unvarying system.

Similarly the *vector potential* A of equation (3) is gotten by a volume integral of the retarded vector current density divided by distance r. As the name indicates it is a vector quantity and in Cartesian coordinates has three components A_x, A_y, A_z.

By means of the equation

$$E' = -\operatorname{grad} \Phi - \frac{1}{c}\frac{\partial}{\partial t} A, \tag{4}$$

the electric intensity due to the reaction of the system is expressed in terms of the charge and current densities.

Equations (2), (3), (4) and the additional equations

$$B' = \operatorname{curl} A, \tag{5}$$

$$\operatorname{div} u = -\frac{1}{c}\frac{\partial}{\partial t}\rho, \tag{6}$$

(where B' is the magnetic induction due to the currents in the system) are the complete equivalent of Maxwell's equations from which they are immediately derivable.

Aside from the fact that the physical significance of the foregoing equations is deducible by direct inspection, they represent a great step because they are *integral equations* whereas Maxwell's equations are *partial differential equations*. A second advantage is that only *true* currents and charges are involved, the displacement currents of Maxwell being replaced by *retarded action at a distance*. Whatever may be said for or against the physical point of view, this effects a substantial mathematical simplification. The formulation of the fundamental field equations in terms of the retarded potentials is due to Lorentz.

In order to complete the specification of the system we have to formulate the relation between the current density $\boldsymbol{u}$ and the electric intensity $\boldsymbol{E}$. In doing so we shall exclude magnetic matter and shall assume that the conductors obey Ohm's law. This restriction is not necessary but effects a great simplification in both the physical picture and the mathematical formulas.[2] We therefore assume that the conducting system is specified completely by its conductivity

$$g = g(x, y, z),$$

and that

$$\frac{1}{g}\boldsymbol{u} = \boldsymbol{E}. \tag{7}$$

Combining with (1) and (4), we have

$$\frac{1}{g}\boldsymbol{u} = \boldsymbol{E}^\circ - \text{grad } \Phi - \frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}, \tag{8}$$

which is our fundamental equation.[3] The preceding set of equations, if g and $\boldsymbol{E}^\circ$ are everywhere specified, enable us, theoretically at least, to completely solve the problem of the distribution of currents and charges in the system.

Before taking up this problem we shall first derive the energy theorem and then investigate the properties of the field by aid of the retarded potentials.

Starting with equation (8), multiply throughout by $\boldsymbol{u}$, getting

$$\frac{1}{g}\boldsymbol{u}^2 = (\boldsymbol{E}^\circ \cdot \boldsymbol{u}) - (\boldsymbol{u} \cdot \text{grad } \Phi) - \left(\boldsymbol{u} \cdot \frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}\right),$$

and integrate over the system, getting

$$\int \frac{1}{g}\boldsymbol{u}^2 dv = \int (\boldsymbol{E}^\circ \cdot \boldsymbol{u}) dv - \int (\boldsymbol{u} \cdot \text{grad } \Phi) dv - \int \left(\boldsymbol{u} \cdot \frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}\right) dv.$$

[2] See Appendix for the general formulas.

[3] See Appendix for the vector notation employed in this paper.

Remembering that $\boldsymbol{u}$ is expressed in *elm.* units, this becomes

$$\frac{1}{c}D = \frac{1}{c}W - \int (\boldsymbol{u}\cdot\text{grad }\Phi)dv - \int\left(\boldsymbol{u}\cdot\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}\right)dv$$

or

$$W = D + c\int (\boldsymbol{u}\cdot\text{grad }\Phi)dv + c\int\left(\boldsymbol{u}\cdot\frac{1}{c}\frac{\partial}{\partial t}\boldsymbol{A}\right)dv,$$

where W is the work done per unit time by the impressed electric field, and D is the *dissipation* per unit time in the system; i.e., the rate at which electrical energy is converted into heat. By means of general theorems in vector analysis, the integrals can be transformed and the equation reduced to the form

$$W = D + \frac{\partial}{\partial t}\frac{1}{8\pi}\int (\boldsymbol{E}^2 + \boldsymbol{H}^2)dv + \frac{c}{4\pi}\int [\boldsymbol{E}\cdot\boldsymbol{H}]_n dS,$$

the last integral being taken over any closed surface which includes the system. Translating this equation into words, it states that:—

The work done per unit time by the impressed forces is equal to the rate of dissipation per unit time plus the rate of increase of the field energy plus the rate at which energy is *radiated from* the system. The vector $(c/4\pi)[\boldsymbol{E}\cdot\boldsymbol{H}]$ is the *radiation vector* and gives the density and direction of energy flow per unit time;[4] it will be denoted by $\boldsymbol{S}$.

We now shall briefly consider the field due to the currents and charges in the system.

If the current density $\boldsymbol{u}$ and charge density ρ are everywhere specified, the retarded potentials are uniquely and completely determined by the formulas

$$\boldsymbol{A} = \int \frac{\boldsymbol{u}(t - r/c)}{r}dv, \qquad \text{(vector)}$$

$$\Phi = \int \frac{\rho(t - r/c)}{r}dv. \qquad \text{(scalar).}$$

The functional notation $\boldsymbol{u}(t - r/c)$ and $\rho(t - r/c)$ indicating that $\boldsymbol{u}$ and ρ are to be evaluated at time $t - r/c$ may profitably be replaced by $\boldsymbol{u}e^{-(p/c)r}$ and $\rho e^{-(p/c)r}$, so that

$$\boldsymbol{A} = \int \frac{\boldsymbol{u}e^{-(p/c)r}}{r}dv,$$

$$\Phi = \int \frac{\rho e^{-(p/c)r}}{r}dv.$$

[4] This is known as Poynting's theorem.

These expressions may be interpreted in either of two ways. (1) If $p = i\omega$ where $\omega = 2\pi f$ and $i = \sqrt{-1}$, then the formulas are the usual complex steady state expressions. On the other hand if p is regarded as d/dt, they are *operational formulas*. It is worth while to explain the latter briefly on account of its own interest and its bearing on the operational calculus.

The differential equations of A and Φ are

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) A = 4\pi u,$$

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) \Phi = 4\pi\rho.$$

where

$$\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2.$$

Now it will be recalled that the differential equation of the electrostatic potential V is

$$\nabla^2 V = 4\pi\rho$$

and that its solution is

$$V = \int \frac{\rho}{r} dv.$$

Operationally

$$\left(\nabla^2 - \frac{p^2}{c^2}\right) \Phi = 4\pi\rho$$

and the corresponding solution is

$$\Phi = \int \frac{\rho e^{-(p/c)}}{r} dv.$$

Now this is an operational equation in which ρ is an arbitrary time function. Its solution depends on the following general operational theorem.[5]

If x is defined by the operational equation

$$x = f(t)e^{-\lambda p},$$

then

$$x = f(t - \lambda).$$

Consequently, the solution of the operational equation for Φ is[6]

$$\Phi = \int \frac{\rho(t - r/c)}{r} dv.$$

[5] See "The Heaviside Operational Calculus," *Bull. Amer. Math. Soc.*, Jan., 1926.

[6] A proof of this theorem by operational methods was privately communicated to the author several years ago by Stuart Ballantine.

Let us now examine the field of the currents and charges by aid of the formulas

$$\boldsymbol{E} = -\operatorname{grad}\Phi - \frac{p}{c}\boldsymbol{A},$$

$$\boldsymbol{H} = \operatorname{curl}\boldsymbol{A}.$$

Performing the indicated operations,

$$\operatorname{curl}\boldsymbol{A} = -\int e^{-(p/c)r}[\boldsymbol{n}\cdot\boldsymbol{u}]\left(\frac{1}{r^2}+\frac{p}{c}\frac{1}{r}\right)dv,$$

$$\operatorname{grad}\Phi = -\int e^{-(p/c)r}\rho\cdot\boldsymbol{n}\left(\frac{1}{r^2}+\frac{p}{c}\frac{1}{r}\right)dv,$$

where $\boldsymbol{n}$ is a unit vector, parallel to $\boldsymbol{r}$, drawn through the contributing element.

We see from these formulas that the magnetic field due to the currents, and the electric field due to the charges, consist each of two components; one varying inversely as the square of the distance from the contributing element and the other inversely as the distance. Writing $p = i\omega = i\cdot 2\pi f$, the orders of magnitude of the two components are $1/r^2$ and ω/c^2 and their ratio is $2\pi(r/\lambda)$, where λ is the wave length.

The first component is the *induction field*, and involves the frequency only through the exponential term; the second is the *radiation* field and involves the frequency linearly.

If we are considering points in the system itself, and if the dimensions of the system are so small that $2\pi(r/\lambda)$ is small compared with unity, the expressions reduce to

$$\operatorname{curl}\boldsymbol{A} = -\int\frac{[\boldsymbol{n}\cdot\boldsymbol{u}]}{r^2}dv,$$

$$\operatorname{grad}\Phi = -\int\boldsymbol{n}\frac{\rho}{r^2}dv.$$

If therefore the dimensions of the system are sufficiently small with respect to the wave length, these expressions can be employed in calculating the distribution of the currents and charges in the system. This is usually the case in circuit theory, even at radio frequencies.

At a great distance from the system, however, the case is quite different. For no matter how large the wave length, λ, if we consider points outside the system such that $2\pi(r/\lambda)$ is everywhere large compared with unity, the second or radiation field will predominate. This leads to the important conclusion that the field which determines the distribution of currents and charges in the system is quite different

from the field which determines the radiation, and explains the fact that radiation may usually be neglected in calculating the distribution in the network.[7]

To examine the radiation field, consider a point P at such a distance from the system that $2\pi(r/\lambda)$ is very large. Choose any point in the system as the origin and write r_0 as the distance from the origin to the point P, and r' the distance from the contributing element P' to the origin. Then

$$r = r_0 - (r' \cdot n),$$

where n is the unit vector parallel to r_0 and

$$A = \frac{e^{-i\omega r_0}}{r_0} \int u \cdot e^{i\omega(r' \cdot n)} dv = \frac{e^{-i\omega r_0}}{r_0} J,$$

$$\text{curl } A = -i\omega \int \frac{e^{-i\omega r_0}}{r_0} [n \cdot J],$$

which determines H.

Instead of calculating E from the formula

$$-\text{ grad } \Phi - \frac{i\omega}{c} A,$$

we make use of the fact that in the dielectric

$$i\omega E = \text{curl } H,$$

whence

$$E = -[H \cdot n].$$

The interpretation of these equations is that in the radiation field E and H are equal, are in phase and are perpendicular to each other and to the vector r_0. Consequently the radiation vector S is given by

$$S = \frac{c}{4\pi} H^2$$

$$= \frac{c\omega^2}{4\pi} \frac{|J|^2}{r^2},$$

and the radiation is everywhere outward.

These formulas can be used to calculate the radiation in terms of the current distribution alone, and the charge distribution does not appear explicitly.

[7] Conversely the field in the immediate neighborhood of the system is no criterion of the radiation field or the radiating properties of the system. This fact is not always kept in mind by radio-engineers.

Derivation of the Familiar Circuit Theory Relations

In the foregoing we have tacitly assumed that the distribution of currents and charges in the systems is known. We now take up the more difficult problem of determining the distribution in terms of the impressed field and the geometry and electrical constants of the system. This will introduce us to circuit theory and the enormous complexity of the general rigorous expressions will show its important role in physics and engineering. In fact without the beautiful simplifications of circuit theory very few problems of this type could be solved.

In taking up this problem there are two possible modes of approach. In accordance with one we start with Maxwell's differential equations and try to find a solution which satisfies the geometry of the system and the boundary conditions. For conducting systems of simple geometrical shapes solutions in this way are possible. For complicated networks, however, this mode of approach is quite hopeless.

The other mode of approach is to start with the equation

$$\begin{aligned} \frac{1}{g}\boldsymbol{u} &= \boldsymbol{E}^{\circ} - \operatorname{grad}\Phi - i\omega\boldsymbol{A} \\ &= \boldsymbol{E}^{\circ} - \operatorname{grad}\int\frac{\rho(t-r/c)}{r}\,dv - i\omega\int\frac{\boldsymbol{u}(t-r/c)}{r}\,dv, \end{aligned} \tag{8}$$

which, together with the relation

$$i\omega\rho = -\operatorname{div}\boldsymbol{u},$$

is an integral equation which completely determines the distribution of currents and charges in the system provided g and $\boldsymbol{E}^{\circ}$ are specified.

For general purposes of calculation it is quite hopeless as it stands. It has, however, several advantages. First, it is a direct and complete statement of the physical relations which obtain everywhere. Second, it uniquely determines the distribution and does not, like the differential equations, involve the determination of integration constants from the boundary conditions. Third, it leads, through appropriate approximations, to the philosophy and equations of circuit theory.

To start with a simple case, the solution of which can be extended without difficulty to the general network, consider a conductor forming a closed circuit. We suppose that it is exposed at every point to an impressed electric force $\boldsymbol{E}^{\circ}$, and we suppose that the surrounding dielectric is perfectly non-conducting. It is now our problem to derive, for this simple circuit, the circuit equations, in terms of *total currents* and *charges*, from the rigorous integral equation for the current and charge *densities*.

In the interior of the conductor let us assume a curve s defined as parallel, at every point, to the direction of the resultant current. We do not know precisely the path of this curve but we do know that such a curve can be drawn. In the case of wires of uniform cross section it will be approximately parallel to the axis of the wire. Let the cross section of the conductor normal to s be denoted by S. The total current I_s, parallel to s, is then given by

$$I_s = I = \int u_s dS.$$

Now corresponding to the surface S and its element dS, let us define a hypothetical surface $\sum$ and its element $d\sigma$ by the equation

$$u_s dS = I d\sigma,$$

whence

$$\int u_s dS = I = I \int d\sigma = I \cdot \sum,$$

so that $\sum$ is always unity. Now multiply the equation

$$\frac{1}{g} u_s = E_s^{\circ} - \frac{\partial}{\partial s} \Phi - i\omega A_s \tag{9}$$

by $d\sigma$ and integrate over the cross section $\sum$; we get

$$\int \frac{u_s}{g} d\sigma = \int E d\sigma - i\omega \int A_s d\sigma - \frac{\partial}{\partial s} \int \Phi d\sigma.$$

This can be written as

$$r(s) I(s) = \overline{E}(s) - i\omega \overline{A}_s(s) - \frac{\partial}{\partial s} \overline{\Phi}(s),$$

or simply

$$rI = \overline{E} - i\omega \overline{A} - \frac{\partial}{\partial s} \overline{\Phi}; \tag{10}$$

r is simply the resistance per unit length of the conductor, since

$$rI^2 = \int \frac{u_s^2}{g} dS = \text{dissipation per unit length due to current } I_s,$$

while $\overline{E}$ is the mean impressed electric force, parallel to s, averaged over the surface $\sum$.

Now consider the term $i\omega \overline{A}$; we have

$$\overline{A} = \int A_s d\sigma = \int d\sigma \int \frac{u'_s}{r} dv, \qquad u' = u(t - r/c)$$

or, neglecting the retardation,

$$\bar{A} = \int d\sigma \int \frac{u_s}{r} dv.$$

We now assume that the " charging " current normal to s is negligibly small in its contribution to the vector potential, whence

$$\bar{A} = \int ds' I(s') \cdot \cos(s, s') \int d\sigma \int d\sigma' \frac{1}{r}$$

$$= \int I(s') \frac{\cos(s, s')}{r} \lambda(s, s') ds',$$

where

$$\lambda(s, s') = \int d\sigma \int \frac{1}{r} d\sigma'.$$

The term $\bar{\Phi} = \int \Phi d\sigma$ of (10) is next to be considered. Writing

$$\rho dS = Q d\tau,$$

where Q is the total charge per unit length, it becomes

$$\int ds' Q(s') \int d\sigma \int \frac{1}{r} d\tau' = \int Q(s') \mu(s, s') ds',$$

and we get finally

$$rI = \bar{E} - i\omega \int I \cdot \cos(s, s') \lambda(s, s') ds' - \frac{\partial}{\partial s} \int Q \cdot \mu(s, s') ds'. \quad (11)$$

This, together with the further relation

$$i\omega Q = -\frac{\partial}{\partial s} I, \quad (12)$$

constitutes an integral equation in the total current $I = I_s$. That is to say, we have succeeded in passing from the rigorous integral equation in the point function *densities* to an approximate integral equation in terms of the total current and charge per unit length of the conductor. The functions λ and μ of this equation, however, while theoretically determinable from the rigorous equation, are not solvable from the approximate integral equation. Indeed they are, strictly speaking, functions of the mode of distribution of the impressed field E°. This fact in most cases, however, is of purely academic interest and λ and μ can be approximately evaluated from the geometry of the conductor by assuming a certain distribution of current density over the cross section. With this problem, however, we have no concern here, we are merely concerned to deduce the form of the canonical equations of circuit theory.

Now let us integrate with respect to s, around the closed curve; we get

$$\begin{aligned} \int rIds &= \int \bar{E}ds - i\omega \int Ids \int \cos(s, s')\lambda(s, s')ds' \\ &= V - i\omega \int lIds, \end{aligned} \tag{13}$$

thus defining the *impressed voltage* V, and the inductance per unit length l. Finally, if we assume that this current variation along the conductor is negligibly small, we get

$$I \int rds = V - i\omega I \int lds,$$

which may be written as

$$RI + i\omega LI = V, \tag{14}$$

which is the usual form of the equation of circuit theory for a closed loop.

In deducing (14) from (10) there is one important point which should be noticed. The assumption that the variation in the current I along the conductor is sufficiently small to justify passing from (13) to (14) does not by any means imply that the effect of the distributed charge, which is absent in (14), is negligible. The term $(\partial/\partial s)\Phi$ vanishes in passing from (12) to (13) because the integration is carried around a closed path. Actually comparing the terms $i\omega\bar{A}$ and $(\partial/\partial s)\bar{\Phi}$, we see that their ratio involves the factor $(\omega/c)^2$ which is an exceedingly small quantity even at very high frequencies. Consequently extremely small variations in the current are sufficient to establish charges which can and do profoundly modify the resultant electric field. These, in the case of a closed circuit, are eliminated from explicit consideration by integrating around a closed curve.

This may be illustrated by brief consideration of a second case where the conductor is not closed but is terminated in the plates of a condenser at $s = s_1$ and $s = s_2$ respectively. Making the same assumption as above, after integrating (11) from $s = s_1$ to $s = s_2$, we get

$$RI + i\omega LI + \Phi_2 - \Phi_1 = V, \tag{15}$$

where $\Phi_2 - \Phi_1$ is the difference in Φ between the condenser plates. Assuming these very close together, $\Phi_2 - \Phi_1$ is approximately proportional to the charge on the condenser, that is, to

$$\int Idt = \frac{1}{i\omega} I,$$

and may be written as $I/\omega C$, whence

$$RI + i\omega LI + \frac{1}{i\omega C} I = V, \tag{16}$$

which is the usual circuit equation for series resistance, inductance and capacity.

Extension of the foregoing to networks containing a plurality of circuits or meshes is straightforward and involves no conceptual or physical difficulties, although branch points may be analytically troublesome. These questions will not be taken up, however, as the foregoing is sufficient to show the connection between general electromagnetic theory and circuit theory and to show how circuit equations may be rigorously derived and their limitations explicitly recognized.

The Telegraph Equation

A particularly interesting and instructive application of the preceding is to the problem of transmission along parallel wires and the assumptions underlying the engineering theory of transmission.[8]

Consider two equal and parallel straight wires so related to the impressed field that equal and opposite currents flow in the wires. Here, corresponding to equation (11), we have

$$\begin{aligned} rI = \bar{E} - i\omega \int I\{\lambda(s, s') - \lambda'(s, s')\}ds' \\ - \frac{\partial}{\partial s} \int Q\{\mu(s, s') - \mu'(s, s')\}ds'. \end{aligned} \tag{17}$$

In this equation $\lambda(s, s')$ is the "mutual inductance" between points s, s' in the same wire while $\lambda'(s, s')$ is the corresponding mutual inductance between point s in one wire and point s' in the other. μ and μ' have a corresponding significance as "mutual potential coefficients."

Now $\lambda(s, s') - \lambda'(s, s')$ is a rapidly decreasing monotonic function of $|s - s'|$ and the same statement holds for $\mu - \mu'$. In view of this property and further assuming the variation of I and Q with respect to s as small, (17) to a first approximation may be replaced by

$$rI = \bar{E} - i\omega I \int (\lambda - \lambda')ds' - \frac{\partial}{\partial s} Q \int (\mu - \mu')ds'. \tag{18}$$

At a sufficient distance from the physical terminals of the wires the

[8] For an entirely different treatment of this problem, reference may be made to "The Guided and Radiated Energy in Wire Transmission," *Trans. A. I. E. E.*, 1924.

integrals become independent of s and approach the limits

$$\int_{-\infty}^{\infty} (\lambda - \lambda')ds' = l,$$

$$\int_{-\infty}^{\infty} (\mu - \mu')ds' = \frac{1}{c},$$

whence

$$rI + i\omega lI + \frac{1}{i\omega c} I = \bar{E}.$$

Finally assuming that the impressed electric intensity $\bar{E} = 0$, and introducing the relation

$$i\omega Q = -\frac{\partial}{\partial s} I,$$

we get

$$\left(r + i\omega l + \frac{1}{i\omega c} \frac{\partial^2}{\partial s^2} \right) I = 0,$$

which is the *telegraph equation*.

Besides its formal theoretical interest the foregoing derivation of the telegraph equation admits of some deductions of practical importance. These deductions, which are rather obvious consequences of the analysis, may be listed as follows.

1. The telegraph equation, as derived above, applies with accuracy only at points at some distance from the physical terminals of the line.
2. The accuracy of the telegraph equation in formulating the physical phenomena decreases in general with increasing frequency.
3. The telegraph equation is the first approximate solution of an integral equation. The first approximate solution decreases in accuracy with decreasing wave length of the propagated current.
4. *While the telegraph equation indicates a finite velocity of propagation of the current along the line, it is based on the assumption that the fields of the currents and charges* (as derived from the potential functions Φ and $\boldsymbol{A}$) *are propagated with infinite velocity.*
5. As a consequence of (4), the telegraph equation does not take into account the phenomena of *radiation*, and in fact indicates implicitly the absence of radiation.

The Coil Antenna

An important example of the type of problem to which the foregoing analysis is applicable is the coil antenna. To this problem

equations (11) and (12) immediately apply but, at least at high frequencies, the approximations introduced above to derive the telegraph equation are not legitimate. This is due to the geometry of the conductor, and also to the fact that the impressed field is not approximately concentrated but is distributed over the entire length of the coil. It is intended to apply these equations to a detailed study of this problem. In the meantime, however, it may be noted that the current depends *not only on the line integral of the impressed electric intensity but also on its mode of distribution along the length of the coil.* This fact may possibly have practical significance in the design of coil antenna and their calibration at very short wave lengths.

Appendix

In the beginning of this paper, it was stated that the analysis applied only to the case of conductors of unit permeability and specific inductive capacity which obey Ohm's Law. The reason for this restriction and the formal extension of the analysis to the more general case will now be briefly discussed.[9]

Suppose that the conductor, instead of having the restricted properties noted above, obeys Ohm's Law but has a permeability μ and specific inductive capacity k which may differ from unity.

The equation (1),

$$\boldsymbol{E} = \boldsymbol{E}^{\circ} - \text{grad } \Phi - i\omega \boldsymbol{A}, \tag{1}$$

still holds, as do also the potential formulas (2) and (3) and the formulas for the electric and magnetic intensities (4) and (5). The relation

$$-iw\rho = \text{div } \boldsymbol{u}$$

is also valid.

The equation $\boldsymbol{u} = g\boldsymbol{E}$ must, however, be modified in the following manner. If we write

$$\boldsymbol{P} = \frac{k-1}{4\pi}\boldsymbol{E},$$

$$\boldsymbol{M} = \frac{\mu - 1}{4\pi\mu}\boldsymbol{H},$$

then the foregoing equations are correct, provided we substitute for the equation $\boldsymbol{u} = g\boldsymbol{E}$ the more general expression

$$\boldsymbol{u} = g\boldsymbol{E} + iw\boldsymbol{P} + \text{curl } \boldsymbol{M}.$$

[9] For a previous discussion, see "A Generalization of the Reciprocal Theorem," *B. S. T. J.*, July, 1924.

By aid of these relations, the problem involves the solution of the simultaneous integral equations

$$\boldsymbol{E} = \boldsymbol{E}^{\circ} - \text{grad } \Phi - iw\boldsymbol{A},$$

$$\boldsymbol{H} = \boldsymbol{H}^{\circ} + \text{curl } \boldsymbol{A}.$$

These simultaneous equations can immediately be reduced to a single integral equation in $\boldsymbol{u}$, the formal solution of which is straightforward. A study of this equation, however, has not been carried far enough to justify further discussion in the present paper.

Note on Vector Analysis and Notations

In the foregoing, vectors are indicated by Clarendon, or bold-faced type. To those unfamiliar with vector analysis the following may be helpful:

grad Φ is a vector with the Cartesian components

$$\text{grad}_x \Phi = \frac{\partial}{\partial x}\Phi, \qquad \text{grad}_y \Phi = \frac{\partial}{\partial y}\Phi, \qquad \text{grad}_z \Phi = \frac{\partial}{\partial z}\Phi;$$

curl $\boldsymbol{A}$ is a vector with the Cartesian components

$$\text{curl}_x \boldsymbol{A} = \frac{\partial}{\partial y}A_z - \frac{\partial}{\partial z}A_y,$$

$$\text{curl}_y \boldsymbol{A} = \frac{\partial}{\partial z}A_x - \frac{\partial}{\partial x}A_z,$$

$$\text{curl}_z \boldsymbol{A} = \frac{\partial}{\partial x}A_y - \frac{\partial}{\partial y}A_x;$$

div $\boldsymbol{u}$ is a scalar; in Cartesian notation

$$\text{div } \boldsymbol{u} = \frac{\partial}{\partial x}u_x + \frac{\partial}{\partial y}u_y + \frac{\partial}{\partial z}u_z.$$

$(\boldsymbol{E}\cdot\boldsymbol{u})$ denotes the *scalar product* of the vectors $\boldsymbol{E}$ and $\boldsymbol{u}$ and itself is a scalar. In Cartesian notation

$$(\boldsymbol{E}\cdot\boldsymbol{u}) = E_x u_x + E_y u_y + E_z u_z.$$

$[\boldsymbol{E}\cdot\boldsymbol{H}]$ denotes the *vector product* of the vectors $\boldsymbol{E}$ and $\boldsymbol{H}$. It is

itself a vector with the Cartesian components

$$[\boldsymbol{E}\cdot\boldsymbol{H}]_x = E_yH_z - E_zH_y,$$
$$[\boldsymbol{E}\cdot\boldsymbol{H}]_y = E_zH_x - E_xH_z,$$
$$[\boldsymbol{E}\cdot\boldsymbol{H}]_z = E_xH_y - E_yH_x.$$

The symbol ∇^2 denotes, in Cartesian coordinates, the operator

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Toll Switchboard No. 3

By J. DAVIDSON

IN the early days of telephony the toll signaling apparatus consisted of a magnetic drop in the line and a drop or ringer in the cord. With the advent of common battery signaling in the local plant, relays and lamps replaced the old type drops and the subscriber was given means for calling the toll operator on a toll connection by operating the switchhook instead of ringing. Up to this time the toll operators were located at the local switchboard and had direct access to the subscriber's line, but with the growth of toll and local traffic, it was no longer economical to place the toll operators at the local board. This led to the development of a separate toll switchboard called the No. 1 board, which had access to the subscriber's line over switching trunks between the toll and local boards. For many years the No. 1 switchboard filled the needs of the time but with the expansion of the toll service and the growth of machine switching local service, it became evident that new arrangements were desirable. The No. 3 toll switchboard was developed to meet the new requirements and it has the following advantages as new installations are required.

(*a*) Reduction in apparatus, resulting in equipment economies.
(*b*) Improved maintenance arrangements.
(*c*) More readily adapted to modifications required by new operating methods.

In discussing the features of the No. 3 board, frequent comparisons will be made with the No. 1 switchboard to set forth the changes which have been made in the design of the new circuits.

Main Features

Cord Simplified by Locating Supervisory Relays in Line and Trunk Circuits

The cord circuit of the No. 1 switchboard is equipped with two supervisory relays. One of these relays responds to 20-cycle current and gives the toll operator a ringing signal, indicating that the distant operator is calling. The second relay responds to direct current received from the switching trunk and gives the operator switchhook supervision of the subscriber. Associated with these two relays are other relays which prevent false signals, and permit the operator to

make a busy test or use the cord for a terminating or a through connection. This cord is shown in Fig. 1.

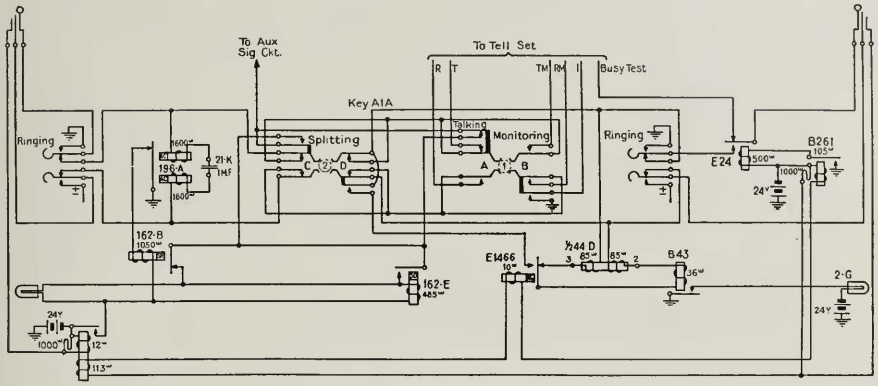


Fig. 1—High impedance toll cord for toll switchboard No. 1

In the No. 3 switchboard the ringing relay and the direct-current supervisory relay, which were formerly connected across the tip and ring conductors of the cord circuit, have been moved from the cord to the line and switching trunk, respectively, and the cord circuit has been simplified as is illustrated in Fig. 2. In this board the line and trunk

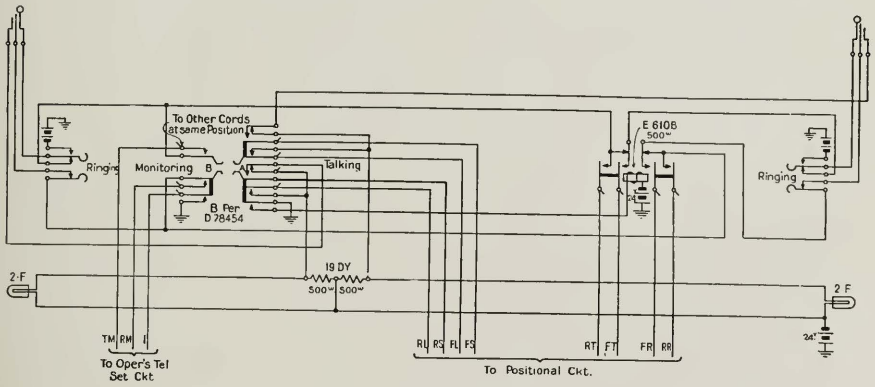


Fig. 2—Toll cord for toll switchboard No. 3

signals are transferred to the cord over the sleeve circuits. This is accomplished by using a nominal sleeve resistance of 1,800 ohms for the line and trunk circuits and connecting the lamps in the sleeves of the cord. Under these conditions there is not sufficient current flowing in the sleeve to light the lamp, but when a ringing signal is received over a line and a cord is associated with that line or when a receiver-on-the-

hook signal is received over a switching trunk, the sleeve resistance of the line or trunk is changed from 1,800 ohms to 80 ohms, which increases the current in the sleeve of the cord sufficiently to light the lamp.

Line Relay Functions in Twofold Capacity

The majority of toll lines in the plant today are of the ringdown type and the operator at one end calls the operator at the distant end by ringing over the line. To receive this ringing signal in the No. 1 board, the lines are equipped with relays which respond to the ringing current received from the distant end of the line and give a line signal. After the operator answers this signal by connecting a cord to the line, the line relay is disconnected and replaced by the ringing relay in the cord which responds to further rering signals over the line. This arrangement of the line and cord, as well as the switching trunk for the No. 1 board, is shown schematically in Fig. 3.

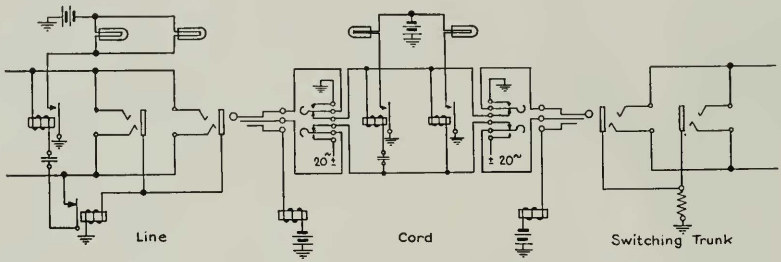


Fig. 3—Schematic: Toll switchboard No. 1 circuits

By transferring the ringing relay from the cord to the line in the No. 3 toll board, this relay is made to function in a twofold capacity, that is, to give the line signal as well as the cord rering signal. When a call is received from a distant point, the apparatus in the line functions to light the line signal and this remains lighted until a toll cord is inserted in the line jack. Further signals over the line cause the apparatus in the line to light the lamp in the cord. This is obtained by changing the sleeve resistance of the line from 1,800 ohms to 80 ohms and is illustrated in schematic form in Fig. 4. As in the past, the line signal is multipled before several operators and appears as a steady illuminated lamp which is extinguished by an operator answering the call. The cord signal appears before one operator and has been changed from a steady lamp signal to a flashing signal for the purpose of obtaining prompt attention on the part of the operator. The cord signal is extinguished when the operator connects to the circuit by the

operation of the talking key. This connects an additional 600 ohms in the sleeve circuit, which releases relays which are held operated in the line circuit and control the lamp.

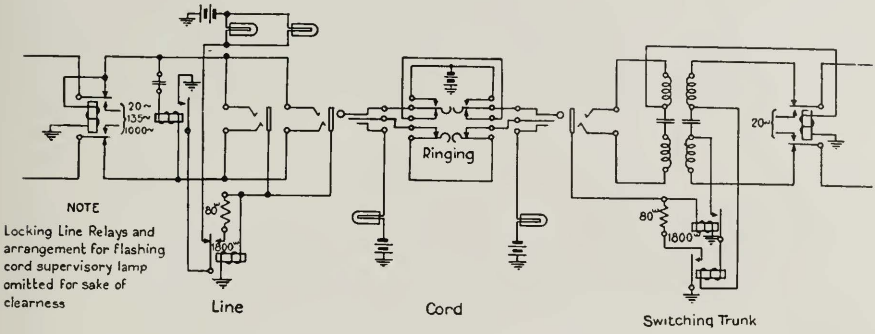


Fig. 4—Schematic: Toll switchboard No. 3 circuits; monitoring and positional circuit keys are not shown

Composite Ringer Simplified

In order that the toll lines may be used for telegraph as well as telephone service, composite sets are often connected into the line circuit at each end. These composite sets are electrical filters which separate the telephone and telegraph currents and direct the telephone currents to the switchboard and the telegraph currents to the telegraph equipment.

When composite sets are connected in the lines terminating in a No. 1 switchboard, it is also necessary to connect a composite ringer in the circuit between the composite set and the switchboard. This is necessary because the 20-cycle current, which is used as ringing current from the switchboard, is in the telegraph range of frequencies and consequently will not pass through the telephone branch of the composite set. The composite ringer substitutes for the 20-cycle outward ringing current received from the switchboard, a higher frequency current which will pass through the telephone path of the composite set. Likewise on incoming ringing signals, the ringer substitutes for the higher frequency current which comes over the line and through the telephone path of the composite set, a 20-cycle current which will operate the ringing relays of the line or cord circuits. A schematic of the composite set and composite ringer, as used with the No. 1 board, is shown in Fig. 5.

In general, the composite ringer for the No. 3 switchboard has been greatly simplified and made a part of the terminating line equipment. This has been accomplished, as illustrated schematically in Fig. 4, by arranging the line circuit so that a relay may be cross-connected in the

line to receive the 20-cycle, or the higher frequency ringing current, and arranging this relay so that it gives the line signal or the cord supervisory signal direct without going through the step of changing ringing frequencies.

Furthermore, the practice of using 20-cycle current in the cord circuit for ringing has been discontinued and ringing is effected in the No. 3 switchboard by connecting 24-volt direct current through the

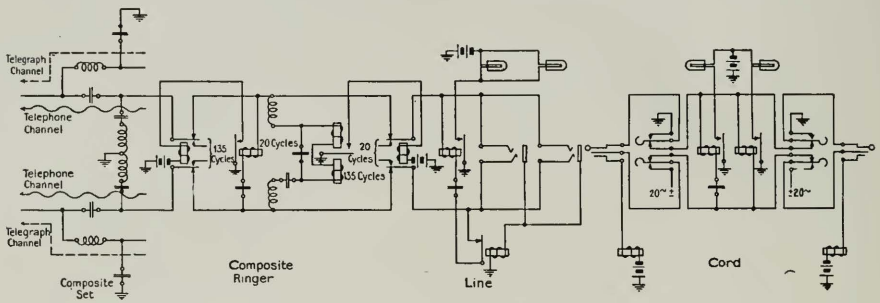


Fig. 5—Schematic: Composite ringer and composited toll line for toll switchboard No. 1

ringing key to the tip conductor of the cord. This current operates a relay of the line or trunk circuit which applies the proper frequency of ringing current to the line or trunk circuit. By this arrangement one relay in the line circuit accomplishes the same result as was accomplished by several relays in the composite ringer. As the ringing current leads to the relay in the line are brought through terminals on the frames, the line can be readily changed for any desired frequency of ringing current.

Elimination of Transfer Key from Face of Inward Switchboard

In the past the practice has been to provide one or two transfer keys per line for each multiple appearance of the line lamp at the inward toll switchboard. The function of these keys is to transfer the inward call from the inward switchboard to the outward delayed positions or to the through positions. With the No. 3 toll switchboard, the use of these transfer keys individual to the line and appearing in the face of each section of the inward switchboard has been discontinued and the transfer is effected by a transfer key in the positional circuit which may be used to transfer a call on any line. This key applies 24-volt battery either directly or through a resistance to the ring conductor of the line and operates the proper transfer relay in the toll line and causes lamps individual to that line to light at the out-

ward, or through positions. This feature not only effects a saving in equipment but saves the space in the face of the switchboard which was formerly occupied by the transfer keys.

Use of Positional Circuit

Another circuit feature of the No. 3 switchboard which marks an improvement over switchboard No. 1 is the use of a so-called positional circuit in which is located much equipment such as splitting keys, dialing keys, etc., which heretofore were individual to each cord. Under normal conditions the tip and ring conductors of the front cord are connected to the tip and ring conductors of the corresponding back cord with no shunts across the circuit. This is illustrated in Fig. 6. By the operation of the talking key associated with each cord

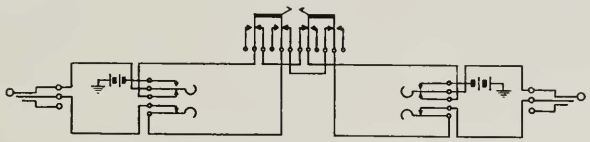


Fig. 6—Schematic: Toll cord talking circuit; talking key normal for toll switchboard No. 3

circuit, the positional circuit is connected between the front and back cords and the operator's telephone set is connected across the circuit as illustrated in Fig. 7. With the talking key of any cord operated, the operator may

(*a*) Dial on either the front or the back cord.
(*b*) Split the talking circuit between the front and the back cords.
(*c*) Transfer an inward call from the inward to the outward or the through positions.

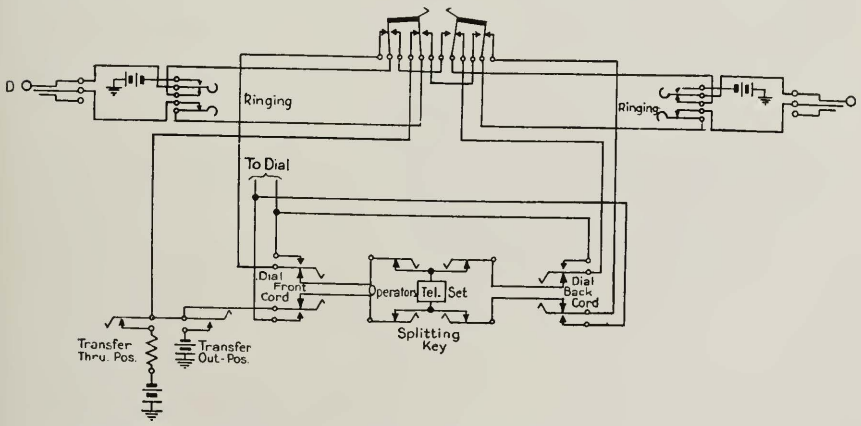


Fig. 7—Schematic: Toll cord positional circuit; talking key operated for toll switchboard No. 3

This circuit arrangement not only effects substantial economies but it is much more flexible and will lend itself to new developments without requiring changes in the cord circuit.

Monitoring and Ringing Keys Individual to Cords

The monitoring and ringing keys are, as in the past, individual to each cord.

Switching Trunk Features

In the No. 3 toll switchboard a repeating coil which has a high impedance to 20-cycle ringing current is used in the outgoing end of the switching trunk. This arrangement has equipment and signaling advantages. Also where loaded toll switching trunks are involved, the use of a repeating coil of the type referred to, but having the proper transmission characteristics, has the advantage of reducing reflection losses by providing for a uniform terminal impedance of the switching trunks.

Principal Advantages

Equipment Economies

As has been pointed out, the expansion in toll business, together with recent developments in the telephone art, have been such that with the circuit arrangements used in the past there has been a growing necessity to add equipment to the cord circuit with the result that the positions are becoming congested with apparatus. With the circuit arrangements outlined for the No. 3 toll switchboard, however, the transfer of the signaling apparatus from the cord to the line and switching trunk makes a marked simplification in the cord and incidentally reduces the congestion in the section. Also it should effect a substantial economy in equipment because of the fact that we are approaching a situation where there are approximately 60 per cent. more cords than lines and 25 per cent. more cords than switching trunks.

The use of the positional circuit and the elimination of the individual splitting key from the toll cord has simplified the switchboard keyshelf. This simplification together with the equipment savings effected by the simplification of composite ringers and the transfer of the supervisory relay equipment from the toll cord to the toll line and switching trunk circuits has effected substantial economies.

Maintenance

In addition to the saving in first cost of equipment the No. 3 switchboard facilitates maintenance. The ordinary toll cords in an office

must be suitable to work with any toll line terminating at the switchboard and consequently with the circuit arrangement used in toll switchboard No. 1, the ringing relay in all the toll cord circuits must be maintained to operate in connection with the longest as well as the shortest line circuit. In the case of the No. 3 toll switchboard, however, the ringing relay is individual to the line and consequently may be adjusted to meet the operating conditions of that line. Long lines with severe ringing conditions require the relay to have a sensitive adjustment while short lines with easy ringing conditions permit a less sensitive relay adjustment to be used which is more easily maintained.

Easily Adaptable to Machine Switching Methods

The introduction of machine switching requires provision for dialing on the trunks and may in the future require the same feature for dialing over toll lines. Such provision in the boards previously employed requires the addition of the necessary keys and relays on a "per cord" basis, whereas with the No. 3 board the equipment can be placed in the positional circuit, without any change in the cord circuit. This results in a great economy in apparatus and makes a change to a dialing basis rather simple.

Summary

It is interesting to note in conclusion that heretofore an increase in cord circuit apparatus has necessarily followed the development of new and improved switchboard systems and the extension of the area of long distance communication. For example, the magneto cord with a single drop bridged across the circuit sufficed in the early days of small magneto boards. The advent of the common battery multiple switchboard brought the necessity for extending switchhook supervision to the toll operator, and resulted in the condenser-type cord consisting of 5 relays, now largely abandoned because of the relatively large transmission loss introduced by it. The high-efficiency cord consisting of 8 relays resulted from the demand for a cord having a minimum transmission loss, and additional complications have resulted in the requirement for dialing in machine switching areas, each improvement, of course, increasing the number of relays in the cord circuit. The No. 3 system, on the other hand, makes possible by the transfer of apparatus to the line and switching trunk and by the use of common positional equipment the relatively simple toll cord shown in Fig. 2 in which the individual apparatus is limited to two keys and one

relay per cord. This provides in many cases a toll cord suitable for either inward, outward or through operation, reduces the apparatus congestion in the section and results in decreased maintenance, while being easily adapted to the future trend in toll development.

The Location of Opens in Toll Telephone Cables

By P. G. EDWARDS and H. W. HERRINGTON

Synopsis: Improved methods have recently been developed for the location of opens in toll cable conductors. The discussion of these methods is prefaced by a review of older practices.

This improved open location method and equipment are sufficiently accurate that in practically all cases a fault in a 60-mile length of cable may be located within a maximum variation of plus or minus one half the length of a cable section (a section is the length of cable between splices—about 750 feet), and therefore enables one to select, prior to the opening of the cable, one or the other of the two splices between which the fault lies. This degree of accuracy is very desirable for practical reasons.

In this development, the line characteristics are considered. The accuracies of calculated locations, assuming no errors in measurements, are compared for different lengths of lines. The impedance bridge circuit is treated to bring out the method of obtaining a balance. The effects of several frequencies of testing potential are analyzed. The probable errors and inaccuracies of measurement which would interfere with the correct location of faults are classified and methods for their correction are developed. The accuracy of the method and the sensitivity of the apparatus are given.

THE location of "opens," or breaks in the continuity of telephone conductors, has always been an important problem in the testing and maintenance of the toll cables of the telephone plant. Although the number of opens encountered in the cable circuits is relatively small as compared to aerial wire lines, their location is more difficult because of certain electrical characteristics of the cable circuits. This condition, coupled with the fact that any work on a toll cable may interfere with a large number of facilities, renders the quick and accurate location of such faults imperative.

A high resistance voltmeter was used in the first feasible attempt to locate opens by means of electrical measurements. The original method of locating opens by the use of a voltmeter consisted essentially of a series of continuity tests. The faulty conductor was connected to ground or to the other wire of the pair at succeeding test stations until the fault was isolated between two adjacent test stations. The trouble was then found, either by inspection of the entire line between test stations, or by further continuity tests with a lineman, first near the middle of the section and later at points gradually approaching the location of the fault.

The inconvenience of such a procedure, however, led to an improved use of the voltmeter. The method employed afforded a rough comparison of the capacitance to ground of the portion of the faulty wire adjacent to the measuring station with that of its good mate or another conductor (of like gauge) following the same route. This was done by allowing the wire to become charged through a volt-

meter and grounded battery, and noting the amount of momentary deflection on the good and bad conductors, respectively, when the polarity of the battery was reversed. The ratio of these deflections gave a general indication of the location of the fault. This method, while giving more accurate results than the continuity test, was still only an approximation and as such was materially affected by line conditions. An appreciable error was produced by leakage due to trees and other causes, and much depended on the judgment of the tester.

Later, the Wheatstone bridge [2] largely displaced the voltmeter. A standard capacitance was compared with the impedance between the open conductor and ground, by varying the ratio arms of the impedance bridge. In this comparison it was necessary to employ some form of alternating testing potential. At first, the simple expedient prevailed of reversing the bridge battery, as in the voltmeter test, the bridge being balanced until no transient unbalance current, or "kick," was indicated by the galvanometer when the battery was reversed. However, when the battery was reversed rapidly, the galvanometer displayed a tendency to stand still at all times. A considerable improvement in the method was effected by providing or the reversal of the galvanometer connections at the same time the battery was reversed. With this arrangement, the galvanometer always read in the same direction, and a balance could be more easily obtained.

Later a relay system was arranged for automatically reversing the connections to the galvanometer as well as for reversing the testing battery. This arrangement relieved the operator of the necessity of doing the reversing manually. Following this, a source of 20-cycle ringing voltage was used for the bridge and also for operating the galvanometer reversing relay. With this arrangement open locations, made on aerial wire lines and short lengths of cable, were fairly satisfactory. However, with the extensive installation of the long toll telephone cables, it was found that open locations made on this type of conductors, did not give a consistently accurate indication of the location of the fault.

As a preliminary step in the development of a suitable open location method and associated apparatus, an analysis was made of all errors which, in general, might enter into a open location. These errors can be classed in several groups for treatment or correction. One group

[2] References: Frank A. Laws, "Electrical Measurements," 1917, McGraw-Hill Book Co., Inc., N. Y.; page 381, "Bridge Measurements of Capacity and Inductance."

"Bridge Methods for Alternating-Current Measurements," D. I. Cone, *Transactions of A. I. E. E.*, July, 1920.

includes errors which are small in comparison to the accuracy of the testing equipment. Errors placed in this group obviously require no compensation. Another group of errors is produced by, or is characteristic of, certain designs of testing equipment. This class of errors has been reduced to negligible magnitude by a redesign of the testing equipment. One general group of errors results from faulty manipulation of the testing equipment or mistakes of computation. This group has been minimized by a convenient arrangement of testing equipment and by outline forms for use in computation.

Another class of errors is introduced by irregularities in the lines or cables on which open locations are made. Some of these are capable of compensation by constant correction factors included in formulae used for computation. Other errors of this class are found to be irregular functions of the length of line, and for their correction or compensation curves have been prepared for each type or condition of irregularity which can be used in the computation of the open location. To simplify the application of the corrections, the curves are so drawn that the correction is given as a simple multiplier.

The preparation of other types of corrections will be developed later in connection with the analysis and treatment of certain specific errors.

In the development of a more sensitive and reliable method of locating opens in telephone lines and cables, it was necessary to make an exact study of the electrical constants of the several types of conductors on which open locations are required. This involved the capacitance and leakance of the conductor to ground, or to neighboring conductors, as well as the series resistance and inductance. The general formula for the impedance of a line open at the distant end is

$$Z_l = Z_0 \coth \theta, \tag{1}$$

where Z_0 is the characteristic impedance of the line and $\theta = Pl$, where P is the propagation constant and l is the length. More fully

$$P = \sqrt{(R + j\omega L)(G + j\omega C)},$$

where R is the series resistance, L the inductance, G the leakance, and C the capacitance, all expressed in terms of the same unit length, and $\omega = 2\pi f$.

In formula (1)

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}},$$

the terms having the same significance as above.

In cable circuits, G is usually zero, and for non-loaded lines, L is

also practically zero. For loaded lines, the inductance is effectually that of the loading coils, and for the frequencies usually employed in open location tests it can be considered as being uniformly distributed. For non-loaded lines, the equation of line impedance is reduced to one of series resistance and capacitance to ground. These constants can be determined by the measurement of short lengths of cable. In making capacitance measurements, the remaining three wires of the quad should be grounded to eliminate their capacitances to ground. When the other three wires of the quad are left free, the capacitance to ground of the faulty wire increases as the length of good wire beyond the break increases. This effect is shown in Fig. 1.

The values of R and C obtained by measuring short lengths of cable are used to calculate P and Z_0. Where the lines are loaded, the nominal inductance and resistance of the loading coils are used and the characteristic constants R and C are, if possible, determined from non-loaded conductors. These constants for one particular cable are listed as follows:

TABLE 1

Constant per Mile	Grade of Loading			
	Non-Loaded	H–44–25	H–174–106	H–245–155
	19-Gauge Inner Layer Conductors			
R	42.90	44.65	47.78	50.34
L	.000+	.015	.062	.089
C	.100	.100	.100	.100
	19-Gauge Outer Layer Conductors			
R	44.00	45.75	48.88	51.44
L	.000+	.015	.062	.089
C	.110	.110	.110	.110
	16-Gauge Conductors			
R	21.00	22.75	25.88	28.44
L	.000+	.015	.062	.089
C	.100	.100	.100	.100

These constants represent values per single-wire mile, R being in ohms, L in henries and C in microfarads.

In making an open location, two impedance measurements are made. One measurement is made on the faulty wire. The other measurement is made on a good wire which follows the same route as the faulty wire. The input impedance of the open conductor divided by the input impedance of the good wire gives an indication of the location of the fault. For short cables, the impedance measured to ground may be regarded as identical with the capacitance com-

ponent of the impedance, because the resistance component of the impedance is negligibly small. However, as the length of cable increases, the input impedance can no longer be regarded as equivalent to the capacitance component of the impedance between the conductor

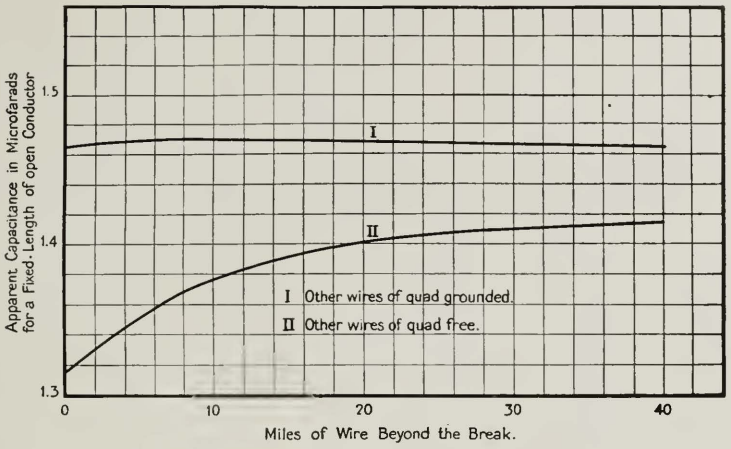


Fig. 1

and ground, because the input impedance is not proportional to the capacitance but varies as the hyperbolic cotangent (Formula 1). The significance and magnitudes of errors inherent in this relation are analyzed and discussed below.

In a homogeneous line, then, if Z_1 and Z_2 represent the input impedances of the bad and good wires respectively, the percentage distance to the fault is most accurately determined by separating the impedances into their real and imaginary parts.

Thus

$$Z_1 = a_1 - jb_1$$

and

$$Z_2 = a_2 - jb_2.$$

If the corresponding lengths are respectively l_1 and l_2, the true distance ratio is l_1/l_2. Assuming that b_2/b_1 is the ratio given by measurement (the shorter line having the higher capacitive reactance),

$$\frac{l_1}{l_2} - \frac{b_2}{b_1} = c,$$

where c is a correction which must be added to or subtracted from the ratio of capacitive reactances to secure the ratio of total capacitance, i.e., the distance ratio.

It is necessary, then, to calculate the ratio b_2/b_1 and the correction c for the different lengths of good and bad wires l_2 and l_1 in order to determine the amount of error arising from the assumption that $b_2/b_1 = l_1/l_2$. The values of capacitive reactance are determined from the fact that

$$b = Z \sin \phi,$$

where Z is the impedance and ϕ its angle as determined from formula (1).

The variation in b with variation in length of line is shown diagrammatically in Fig. 2. The total length of line is represented by Ol_2. The reciprocal of b is plotted on the vertical axis for different lengths of line. For an open at l_1 the location indicated by the ratio b_2/b_1 is at n whereas the true location is at m. The correction $mn = c$ must be subtracted from the apparent location to give the true location of the open. In the lower curve, this error is plotted against the total length of line l.

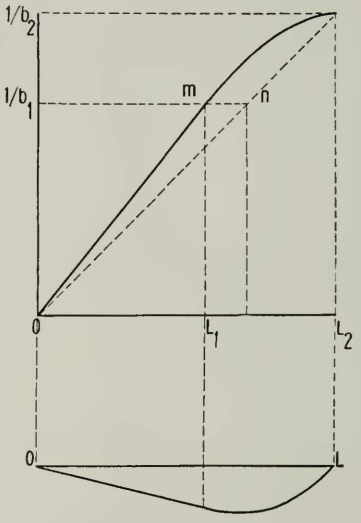


Fig. 2—Diagram showing the construction of a correction curve. Abscissas represent true linear distance; ordinates of upper curve represent measured capacitance; ordinates of lower curve represent errors of computed location.

It remains, then, to calculate b for a large number of lengths from zero to the maximum length of cable to be encountered. These values of b are used to calculate b_2/b_1 for different total lengths of line l and different fault locations l_1: that is, b is calculated for different lengths up to one hundred miles; then a set of ratios of b_2/b_1 and l_1/l_2 can be determined using the b of one hundred miles as b_2 and b for all the shorter lengths as b_1. Similarly, a set of ratios can be calculated for 95, 90, 85, 80, 75, etc., miles as total lengths. Since the interpolation of hyperbolic functions is at best a tedious calculation, even values of hyperbolics can be used in formula (1) and the corresponding odd lengths in miles calculated.

The ratio b_2/b_1 is then subtracted from the ratio l_1/l_2, in each instance this procedure resulting in a family of correction curves expressed in percentage such as that shown in Fig. 3. In this figure the correction is plotted against apparent rather than actual percentage distance in

order to facilitate the use of the curves in locating actual cases of trouble. Where the length of line being measured does not correspond to one of the curve indices, the curves can be interpolated and the desired curve drawn in.

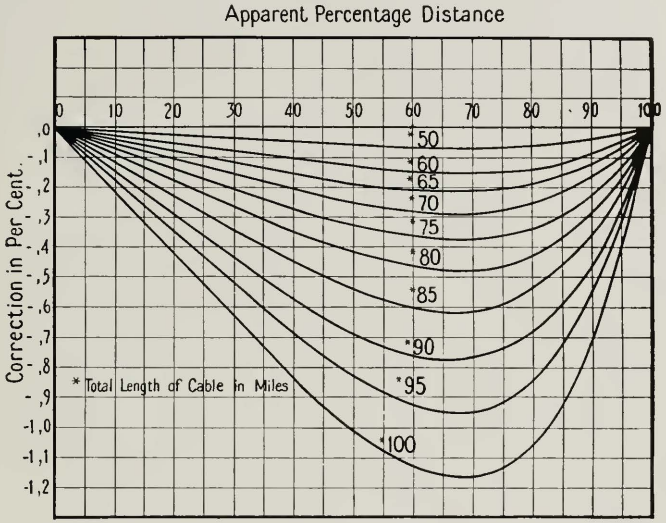


Fig. 3—Corrections to be used in finding true position of a fault from the apparent position. Apparent distance to the fault is indicated as a percentage of the total length of the cable.

The treatment becomes more involved for cables which are not homogeneous. The quads in the outer layers of any unspliced length of cable tend to differ in resistance from pairs in the inner layers, both because of greater length of turn in the outer layers and because of a possible difference in temperature between inner and outer layers. For homogeneous cables treated above these variations in resistance as well as attendant differences in capacitance to ground are equalized by mixing the quads among the various layers at each splice. In general, cables are spliced in this way, although there are several long distance cables now in service which are not homogeneous in this respect. Such cables are spliced in a special arrangement which is essentially a "transposition" system of splicing.

The "transposed" cable may readily be visualized by a consideration of the system of splicing which is employed. Instead of the 19-gauge quads being mixed promiscuously among all the layers, they are divided into an outer layer group and an inner layer group. The quads of each group are mixed among themselves (but not with

quads of the other group), at every splice but one. This particular splice, near the middle of the cable, is known as the "transposition point." At this splice the two groups are "transposed," that is, outer layer quads are spliced to inner layer quads and inner layer quads are spliced to outer layer quads. In this way the differences in resistance and capacitance to ground of outer and inner layer quads are averaged at the "transposition point" for each group. The average resistance or capacitance to ground of a conductor of the outer layer group will therefore differ appreciably from the average resistance or capacitance to ground of a conductor of the inner layer group. The constants given in Table 1 are for a cable of this type.

As in the case of the non-transposed cable, it is necessary to calculate values of b for different lengths of line. Up to the transposition point the procedure is the same as above, viz.,

$$Z_{l1} = Z_{O1} \coth P_1 l_1,$$

where the subscript denotes the first section adjacent to the measuring station. As soon as the point of open falls on the distant side of the transposition point, where the conductor changes layers, the calculation of Z_l is a composite one. That is

$$Z_{l12} = Z_{O1} \tanh (P_1 l_1 + \delta), \tag{2}$$

where Z_{l12} is the combined input impedance, Z_{O1} is the characteristic impedance of the adjacent section, P_1 and l_1 its propagation constant and length respectively, and

$$\delta = \tanh^{-1} \frac{Z_{l2}}{Z_{O1}},$$

where Z_{l2} is the input impedance of the distant section calculated from the formula

$$Z_{l2} = Z_{O2} \coth P_2 l_2.$$

However, the calculation of formula (2) involves practical difficulties, and it is best reduced as follows:

Denoting $P_1 l_1$ as θ_1 and $P_2 l_2$ as θ_2,

$$Z_{l12} = Z_{O1} \tanh \left(\theta_1 + \tanh^{-1} \frac{Z_{l2}}{Z_{O1}} \right),$$

and expanding,

$$Z_{l12} = Z_{O1} \left\{ \frac{\tanh \theta_1 + \dfrac{Z_{l2}}{Z_{O1}}}{1 + \tanh \theta_1 \left(\dfrac{Z_{l2}}{Z_{O1}} \right)} \right\}.$$

But

$$Z_{l2} = \frac{Z_{O2}}{\tanh \theta_2},$$

whence, substituting,

$$Z_{l12} = Z_{O1}\left\{\frac{\tanh \theta_1 + \frac{Z_{O2}}{Z_{O1}} \cdot \frac{1}{\tanh \theta_2}}{1 + (\tanh \theta_1)\left(\frac{Z_{O2}}{Z_{O1}}\right)\left(\frac{1}{\tanh \theta_2}\right)}\right\}$$

$$= Z_{O1}\left\{\frac{\tanh \theta_1 \tanh \theta_2 + \frac{Z_{O2}}{Z_{O1}}}{\tanh \theta_2 + \tanh \theta_1 \frac{Z_{O2}}{Z_{O1}}}\right\}$$

$$= Z/\phi.$$

Here, as in the case of the non-transposed cable, even values of hyperbolic functions are chosen and the corresponding odd values of length are calculated. Thus, the impedances of different arrange-

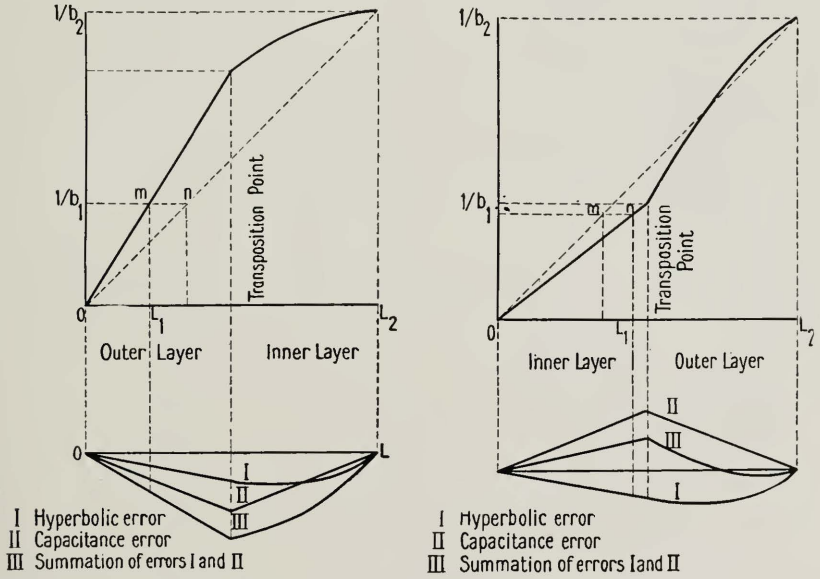


Fig. 4—Diagram showing the construction of a correction curve for a transposed cable. The fault is assumed to be at point L_1. Ordinates are the same as in Fig. 2. The lower curves I and II show the two errors which are algebraically added to give the total error.

Fig. 5—This diagram is similar to Fig. 4 except that the fault at L_1 is in that group of conductors which enters the measuring office in the inner layer of the transposed cable.

ments of conductor for a given transposed cable can be found and the calculated values of b_2 and b_1 may be used to plot a correction curve.

If, for a given transposed cable, a correction curve is calculated as outlined above, it will be found to have the general characteristics shown diagrammatically in Figs. 4 and 5. Fig. 4 is for the case where the faulty conductor enters the measuring station in the outer layer and Fig. 5 for the case where the faulty conductor enters the measuring station in the inner layer. In the lower half of each figure the total errors are separated into their component parts. In either Fig. 4 or Fig. 5 the total error curve can be assumed to be made up of two factors, a hyperbolic error similar to that shown in Fig. 2, and a straight line error due to the fact that the two halves of the cable do not have the same unit capacitance. The latter error would be present in such a cable even though its length were insufficient to cause a hyperbolic error. Such a division can be made because the constants of the inner and outer layers do not differ enough to affect the hyperbolic error appreciably. It is not possible to plot a general family of curves of the type shown in Fig. 3 due chiefly to the fact that the location of the transposition point and the difference in total length of different cables constitute a double variable. The need for a correction involving the double variable has been met by the development of open location equipment and methods which reduce this error to a negligible magnitude for the lengths and types of lines encountered in practice. The rigid treatment is, however, that outlined in formula (2).

The amount of the hyperbolic error can be calculated closely enough using the average constants of the inner and outer layers. The size of the straight line error due to the different capacitances of the inner and outer layers is found as follows:

Let C_1 and C_2 represent the adjacent and far end capacitances per unit length and D_1 and D_2 the respective lengths of these sections. The total conductor capacitance is then $D_1C_1 + D_2C_2$. If D is the location of the fault and D is less than D_1, that is, the fault is in the half of the cable adjacent to the measuring station, the bad wire capacitance is DC_1. The apparent location is

$$\frac{DC_1}{D_1C_1 + D_2C_2}$$

and the correction is

$$\frac{D}{D_1 + D_2} - \frac{DC_1}{D_1C_1 + D_2C_2}.$$

Similarly when the trouble occurs beyond the transposition point,

and D is greater than D_1, the capacitance of the bad wire is

$$\frac{D_1C_1 + (D - D_1)C_2}{D_1C_1 + D_2C_2}$$

or

$$\frac{D_1(C_1 - C_2) + DC_2}{D_1C_1 + D_2C_2}$$

and the correction is

$$\frac{D}{D_1 + D_2} - \frac{D_1(C_1 - C_2) + DC_2}{D_1C_1 + D_2C_2}.$$

The relative sizes of the capacitance and hyperbolic errors are shown in Fig. 6 where these two components and their sum are plotted as corrections against the apparent percentage distance to the fault.

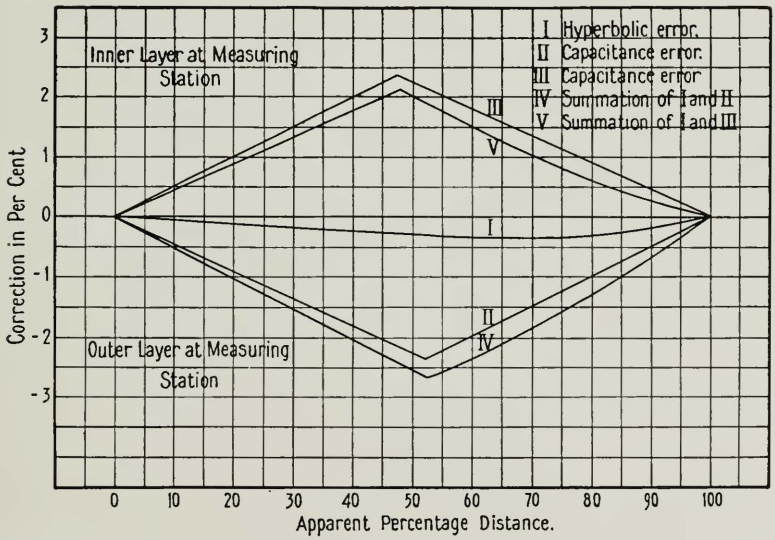


Fig. 6—Diagram for a transposed cable, showing the hyperbolic error as compared to the capacitance error,—measuring frequency four cycles.

In the development of a suitable open location method, it was necessary to select a definite frequency of testing potential for use with an impedance bridge. It was desirable that the selected frequency of testing potential should permit of a design of testing equipment which would have convenient operating characteristics. It was also desirable that the frequency of testing potential be selected to minimize errors which varied with frequency. The calculation of

hyperbolic errors for different frequencies of testing potential showed that this error decreased with frequency, the optimum value of frequency being zero. This relation is shown in Fig. 7, where the maximum errors at different frequencies for a 60-mile length of 19-gauge, non-loaded cable are plotted. However, with zero frequency, the sensitivity to unbalance for an impedance bridge network is also zero, increasing as the frequency increases.

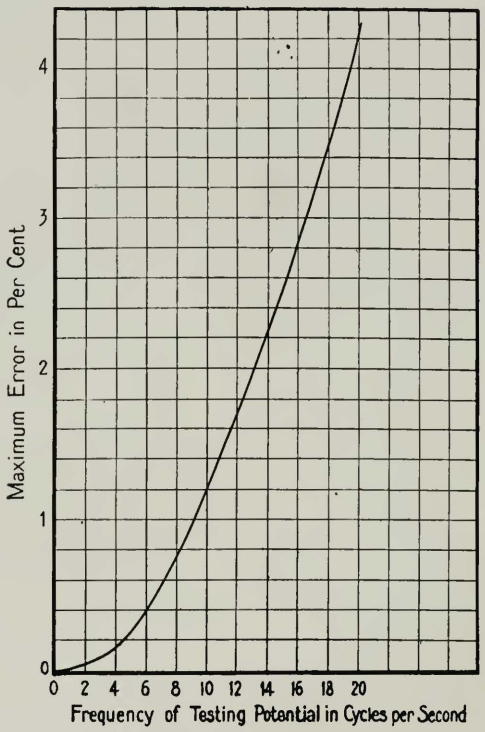


Fig. 7—The maximum hyperbolic error for various frequencies of testing potential.

The problem, then, was one of selecting a frequency which would be low enough to make the hyperbolic errors negligible for all cases of lines as regards length, gauge, and loading, and at the same time provide a sensitivity which would be sufficient to permit an accurate balance of a bridge. From this standpoint it may be observed that for a decreasing frequency of testing potential the maximum rate of decrease of hyperbolic error appears at about four cycles as shown by Fig. 7. A computation of the hyperbolic error for measurements made at four cycles on sixty miles of cable gives results as follows:

Type of Cable Circuit	*Average Percentage Error*
Extra Light or Non-Loaded 19-Gauge Cable	0.083%
Medium Heavy or Heavy Loaded 19-Gauge Cable	0.033
Extra Light or Non-Loaded 16-Gauge Cable	0.022
Medium Heavy or Heavy Loaded 16-Gauge Cable	0.090

Since these errors, at a frequency of four cycles, are for the maximum length of line which may be encountered in practice, it was considered that they might be neglected in comparison with the importance of securing a frequency of testing potential which would be high enough to give an impedance bridge a suitable sensitivity to unbalance. Hence, a computation of the sensitivity appeared to be the next step in the selection of a suitable frequency of testing potential.

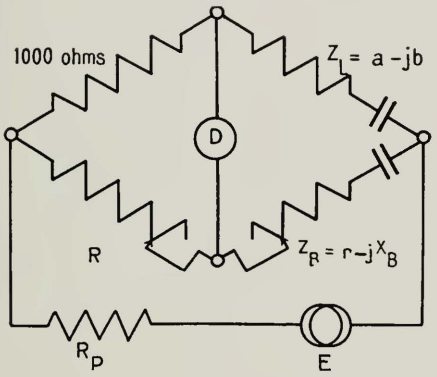


FIG. 8.

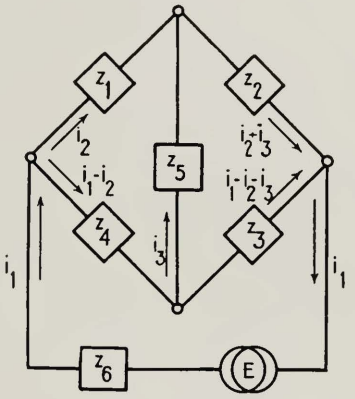


FIG. 9.

Fig. 8—Bridge used in the location of opens. E represents the low-frequency source, and R_p a protective resistance. The rheostat R and the 1000-ohm resistance may be regarded as the ratio arms of the bridge. The impedance of the line is represented by Z_L, a resistance and capacitance in series. The rheostat r is used in balancing the resistance component of the line impedance, so that the impedance angle of Z_B will equal the impedance angle of Z_L.

A very sensitive electrodynamometer, or a galvanometer equipped with an electromagnetic field, was used as a detector in the impedance bridge network shown in Fig. 8. The sensitivities for several frequencies of testing potential were computed for this modified form of the De Sauty bridge. The condition for balance in this impedance bridge network is

$$RZ_l = 1000Z_B$$

or

$$R(a - jb) = 1000(r - jX_B).$$

Collecting reals and imaginaries,

$$Ra = 1000r$$

and

$$Rb = 1000X_B.$$

If measurements R_1, r_1 and R_2, r_2 are made on the bad and good wires respectively,

$$R_1a_1 = 1000r_1$$

and

$$R_2a_2 = 1000r_2;$$

also

$$R_1b_1 = 1000X_B$$

and

$$R_2b_2 = 1000X_B,$$

$$R_1b_1 = R_2b_2,$$

or

$$\frac{R_1}{R_2} = \frac{b_2}{b_1}.$$

Yet the design of a suitable bridge is not concerned alone with balanced condition, but rather with the sensitivity and ease of balance with slight unbalances present. An indication of the probable sensitivity is afforded by a solution of this bridge network to determine the phase and magnitude of the galvanometer unbalance current with respect to the impressed voltage. These have been obtained from the equation

$$i_3 = \frac{E(z_2z_4 - z_1z_3)}{\begin{vmatrix} -z_5 & (z_1 + z_4) & -z_4 \\ (z_2 + z_3 + z_5) & (z_2 + z_3) & -z_3 \\ -z_3 & -(z_3 + z_4) & (z_3 + z_4 + z_6) \end{vmatrix}} \tag{3}$$

in which the denominator is a determinate, and the symbols are those used in Fig. 9.

Since the condition for balance is

$$z_1z_3 = z_2z_4$$

we have, using the notation of Fig. 8,

$$1000(r - jX_B) = R(a - jb)$$

and, equating reals and imaginaries,

$$1000r = aR, \tag{4}$$

$$1000X_B = bR. \tag{5}$$

This impedance bridge can be balanced in two ways: by varying r and X_B, keeping R constant, or by varying R and r, keeping X_B constant (Fig. 8). The effect is essentially the same in either case. When R is varied instead of X_B, the only difference is that the balance of the bridge is disturbed for both the real and imaginary components. This fact necessitates a correction of r each time R is changed in securing a balance of b against X_B. If X_B is varied, the balances of r against a, and X_B against b, are independent functions. In practice, it is easier to vary R and keep X_B constant, but for the purpose of theoretical discussion it lends clarity to consider X_B to be variable from the condition of balance.

From the equation (1) above the impedances of different lengths of line may be computed. For the purpose of designing a suitable impedance bridge arrangement, it is sufficient to consider the 19-gauge, non-loaded cable only, as the effect of loading on the general line characteristic is small at the frequencies employed. A number of impedance values representing different lengths of 19-gauge, non-loaded

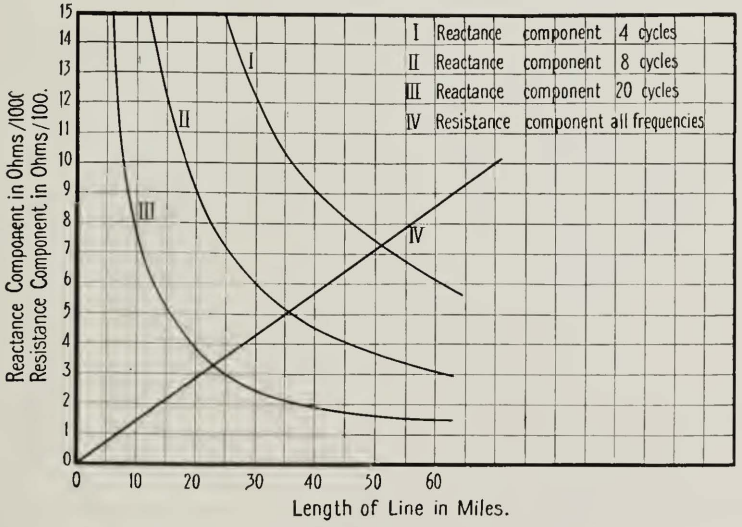


Fig. 10—Characteristics of the resistance component and capacitance component of the impedance of lengths 19-gauge, non-loaded cable.

cable up to sixty miles, at three frequencies, viz., twenty, eight and four cycles, were selected and the condition of balance of the impedance bridge calculated for each case from equations (4) and (5). Curves of these impedances are shown in Fig. 10, where the reactances and resistances at the three chosen frequencies are plotted.

Since the sensitivity of the bridge network should be a maximum when the unbalance is small, i.e., when a balance is about to be secured, this condition is the one with which the sensitivity calculation is concerned. The capacitance X_B of the bridge network was assumed to vary 10% from the condition of balance and the galvanometer

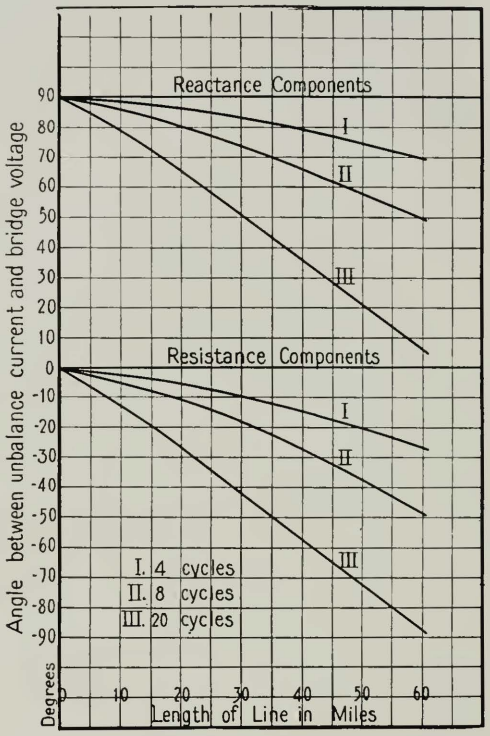


Fig. 11—The variation of the angle between the bridge voltage and the unbalance current for the reactance component and resistance component of non-loaded 19-gauge cable in lengths varying from zero to sixty miles. The curves represent characteristics for frequencies of four, eight and twenty cycles.

unbalance current, i_3, was then calculated from equation (3) for each length of line at each frequency. Both the magnitude and phase angle were found. Similarly the resistance r was assumed to vary 10% from the condition of balance, X_B remaining balanced, and another set of galvanometer unbalance currents was determined. Such calculations are particularly tedious, involving successive additions and subtractions, multiplications and divisions of complex quantities. The results of these calculations are shown in Figs. 11, 12 and 13. Fig. 11 represents the variation in phase angle of i_3 with

respect to the bridge potential, Fig. 12 the magnitude of i_3, and Fig. 13 the relative sensitivities obtained with different frequencies and different lengths of line. These sensitivities are proportional to i_3 (Fig. 12) and to the cosine of the angle between i_3 and the field current

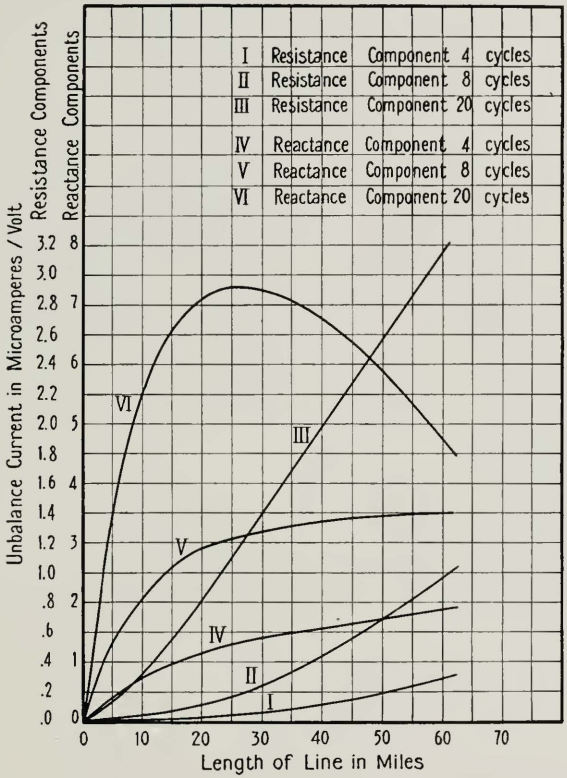


Fig. 12—Magnitude of the unbalance current for measurements, on the impedance of non-loaded 19-gauge cable, at frequencies of four, eight, and twenty cycles.

of the alternating-current galvanometer. In the calculation of the curves of Fig. 13 the field current of the galvanometer was assumed to be in phase with the bridge potential for the case of an unbalance in r, and leading the bridge potential by ninety degrees for the case of an unbalance in X_B.

Referring to Fig. 11, it is seen that for short lengths of line the unbalance current caused by unbalancing r is almost in phase with the voltage, while the unbalance current due to unbalancing X_B leads the voltage by approximately ninety degrees. As the length of line increases, the phase angles tend to lag from these positions, due to

the effect of the convergent variation of the resistive and reactive components of the line impedance, that is, the resistance increases and the reactance decreases with increase in length of line. This lag is greater for the higher frequencies. The total variation for sixty miles of cable measured at twenty cycles is practically ninety degrees which means that for a given field current the sensitivity using twenty cycles must approach zero with some length of line between zero and sixty miles. This condition is illustrated in Fig. 13, where the sen-

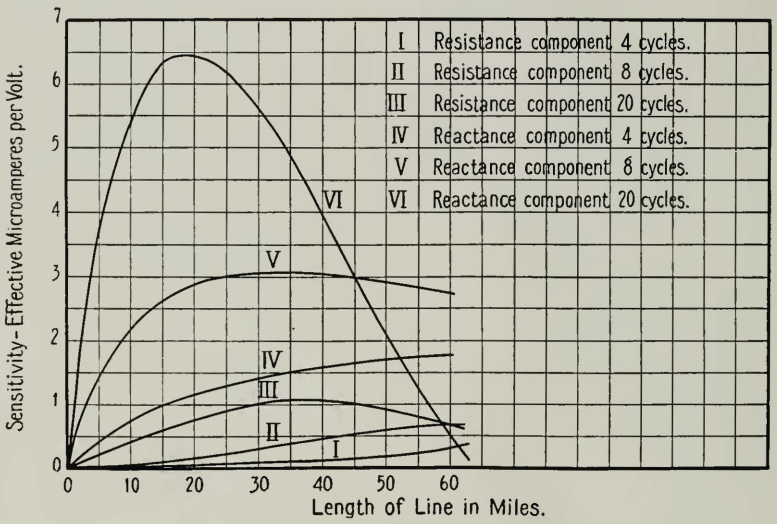


Fig. 13—Sensitivity to unbalance for the bridge network for impedance measurements on lengths of non-loaded 19-gauge cable. These curves represent sensitivities for frequencies of testing potential of four, eight and twenty cycles.

sitivity for twenty cycles is a maximum at twenty to thirty miles, but falls rapidly toward zero at the longer lengths of line. With the lower frequencies and the setting of field current used, the general effect is an increase in sensitivity as the length of line increases, and a decrease in sensitivity with decrease in frequency. This decrease is due to the decrease in reactance with increase in length of line (Fig. 10).

It would appear that provision should be made for shifting the phase of the field current through ninety degrees, its two positions being respectively in phase with the bridge potential and ninety degrees leading. Thus the two components, r and X_B, could be balanced independently except for the shift in phase with different line lengths. This phase shift is small with a frequency of testing potential of four cycles.

A frequency of four cycles was selected as the optimum as regards the size of hyperbolic error discussed above, the sensitivity available, and the amount of phase shift with increase in length of line. The curve showing the change of hyperbolic error with variation in frequency (Fig. 7) shows four cycles to be at or near the critical point of the curve. The sensitivity at four cycles has the advantage of being sufficient but not excessive. To a large extent, the condition of phase shift (with increase in length of line) governs the ease of securing a balance over the range of line lengths. The ideal arrangement would be one in which the field current could always be placed in phase with the component of the bridge unbalance current it was desired to eliminate. Such a quality is not characteristic of the type of bridge used; however, a desirable approximation of such an arrangement is obtained when a four-cycle frequency of testing potential is used.

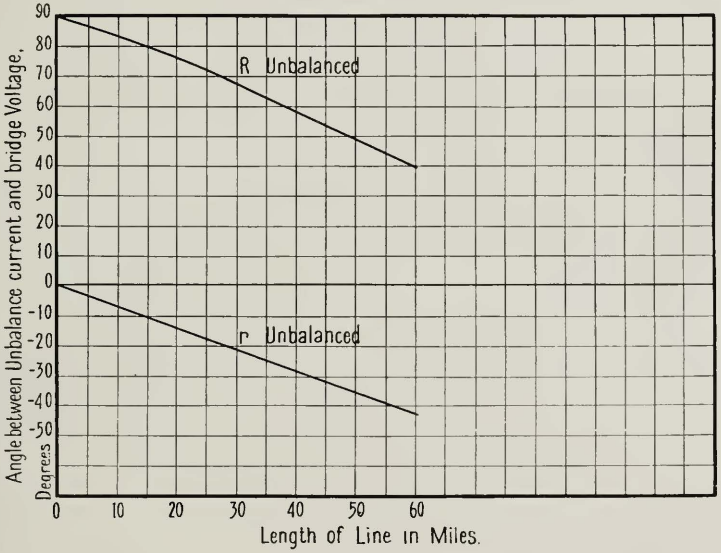


Fig. 14—Calculated variation in the angle between unbalance current and bridge voltage for unbalance in the capacitance component R and for unbalance in the resistance component r of the impedance of non-loaded 19-gauge cable. Testing potential of 4-cycles.

In order to check the assumption (stated above) that the bridge could also be balanced by keeping X_B constant and varying R and r, without materially changing the conditions of balance, another set of galvanometer unbalance currents was calculated with X_B constant and R and r varied respectively 10% from the condition of balance.

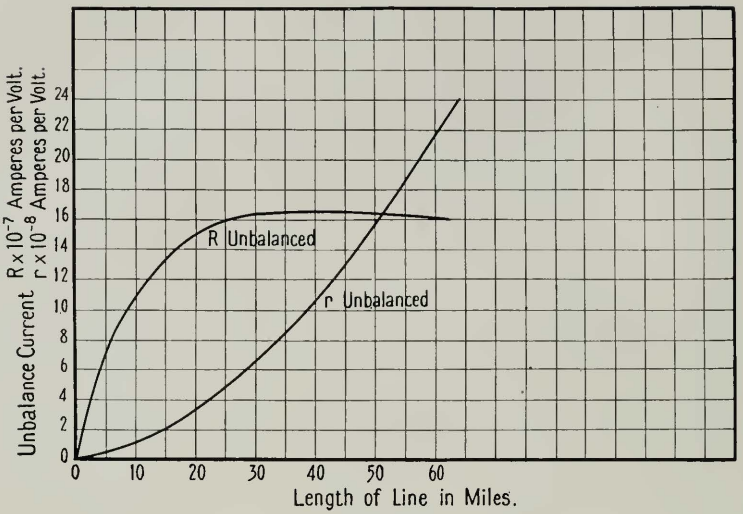


Fig. 15—Calculated magnitude of the unbalance current for unbalance in the capacitance component R and unbalance in the resistance component r of the impedance of non-loaded 19-gauge cable. Testing potential 4-cycles.

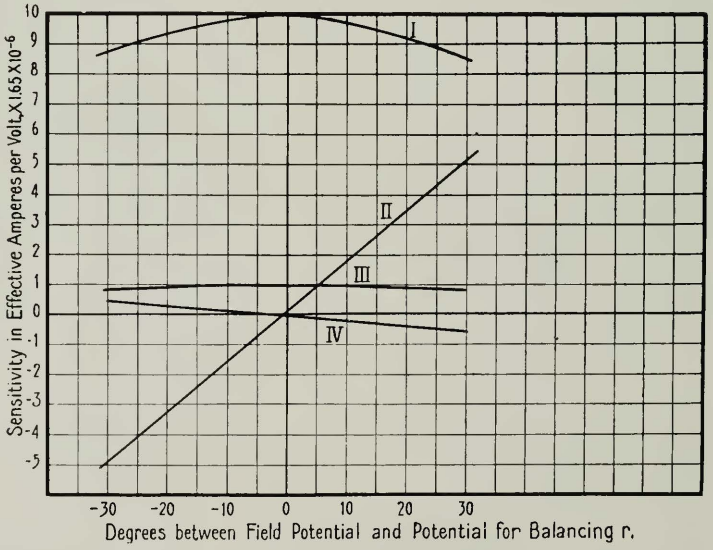


Fig. 16—Calculated sensitivity of the bridge at four cycles on a 50-mile length of non-loaded, 19-gauge cable. Sensitivities for unbalance in the component:

I—R with the phase of testing potential applied for balancing R.

II—R with the phase of testing potential applied for balancing r.

III—r with the phase of testing potential applied for balancing r.

IV—r with the phase of testing potential applied for balancing R.

The chosen frequency of four cycles was used in this calculation. These curves of phase angle and current magnitude are shown in Figs. 14 and 15, and correspond to those shown in Figs. 11 and 12, calculated by the other method.

Since R increases with increase in length of line, the sensitivity per ohm change in R will be better throughout the range of lengths if the sensitivity for a given change in R is a maximum when R is a maximum. Taking fifty miles as the average total length of line, and

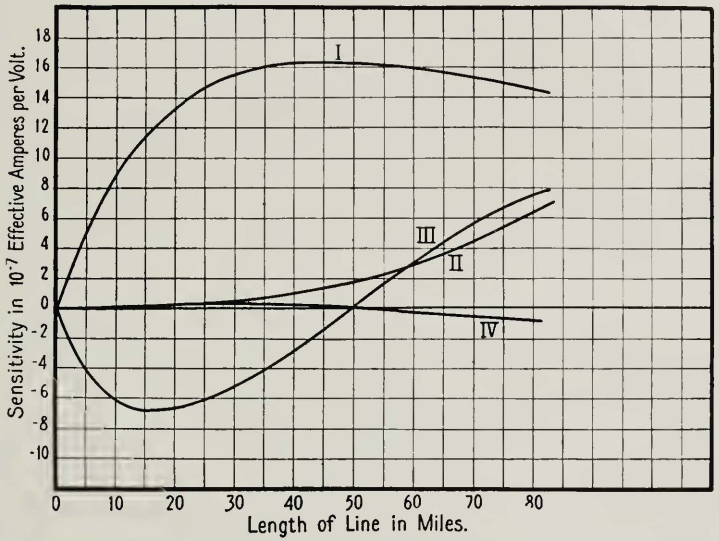


Fig. 17—Calculated sensitivity for lengths of non-loaded 19-gauge conductor using the impedance bridge arrangement which gave the maximum sensitivity for curve I in Fig. 16. Sensitivity for unbalance in the component:

I—R with the phase of testing potential applied for balancing R.
II—r with the phase of testing potential applied for balancing r.
III—R with the phase of testing potential applied for balancing r.
IV—r with the phase of testing potential applied for balancing R.

assuming two bridge potentials ninety degrees apart, a set of sensitivity curves was calculated for this length of line, the phase of the field potential being varied on either side of the bridge testing potential. The lag of the field current from the field voltage calculated from the field inductance and resistance was found to be about forty degrees. The resulting sensitivity curves are shown in Fig. 16, and these indicate a maximum sensitivity for R and r with the field potential in phase with the potential used to balance r.

With the assumed conditions as determined by this calculation,

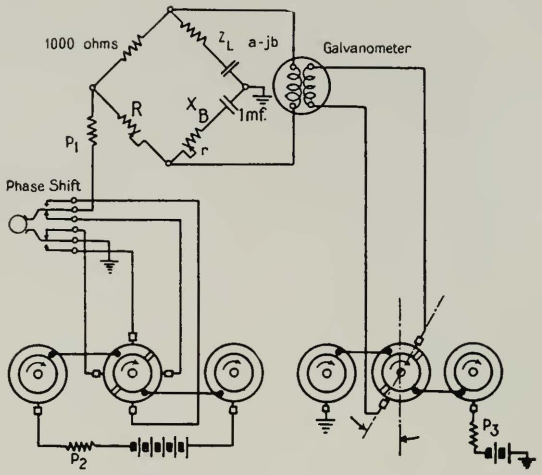


Fig. 18—Impedance bridge circuit showing the arrangement used in applying testing potentials in quadrature.

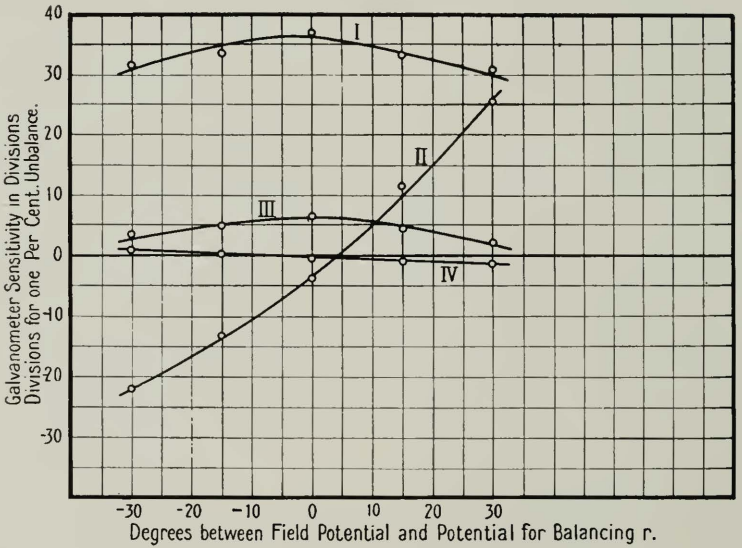


Fig. 19—Observed galvanometer sensitivities which are comparable to the calculated sensitivities of Fig. 16. Sensitivity for unbalance in the component:

I—R with the phase of testing potential applied for balancing R.
II—R with the phase of testing potential applied for balancing r.
III—r with the phase of testing potential applied for balancing r.
IV—r with the phase of testing potential applied for balancing R.

viz., two bridge potentials ninety degrees apart and a field current lagging one bridge potential forty degrees, a complete set of sensitivity curves was calculated for different lengths of line from zero to eighty miles. These are shown in Fig. 17. It should be noted that the sensitivity for detecting an unbalance in R is a maximum at the desired length of fifty miles. The sensitivity for an unbalance in r

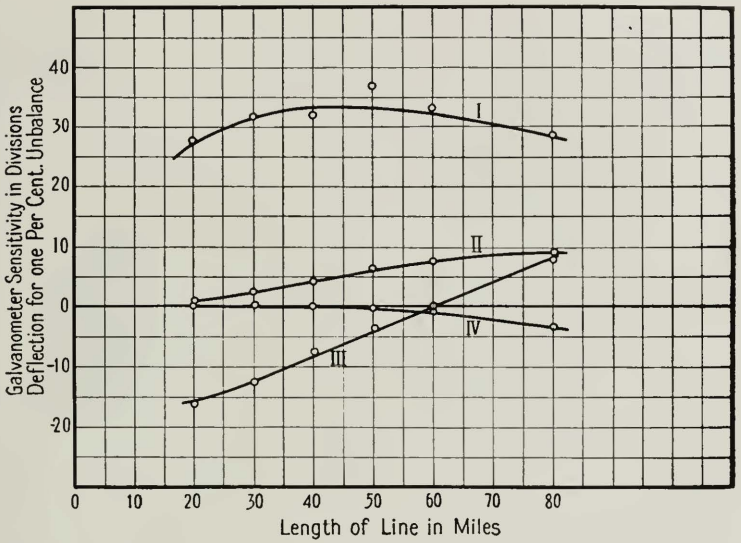


Fig. 20—Observed galvanometer sensitivities which are comparable to the calculated sensitivities of Fig. 17. Sensitivity for unbalance in the component:

I—R with the phase of testing potential applied for balancing R.
II—r with the phase of testing potential applied for balancing r.
III—R with the phase of testing potential applied for balancing r.
IV—r with the phase of testing potential applied for balancing R.

is low at the shorter lengths of line, but increases as the length of line increases. It may be noted that in both Figs. 16 and 17 the sensitivity curve for changes in R passes through zero when the testing potential is applied for balancing r. Likewise the sensitivity curve for changes in r passes through zero when the testing potential is applied for balancing R. This point is where the field current and galvanometer unbalance current are ninety degrees out of phase. Naturally this point coincides with the point of maximum sensitivity for the normal potential arrangement. As stated above, it would be ideal if these "reverse sensitivities" could be zero and the "true sensitivities" could be maximum throughout the entire range of lengths. Reference to Fig. 17 will show that the reverse sensitivity

for r is practically zero throughout the range of lengths. This fact in itself is significant. Assuming r to be set on zero, R could be varied to secure an approximate balance, using the proper testing potential. This balance would be fairly accurate since the reverse sensitivity for r is quite low throughout. Shifting bridge potentials ninety degrees, r could be adjusted almost to the proper point since R is practically



Fig. 21—Commercial design of motor-driven commutator which is used to supply 4-cycle alternating potentials for the impedance bridge.

correct. The balance of R and r can then be refined as often as is necessary to secure a perfect balance. Since r is not used in any calculation, and since its effect on R is small once an approximate balance of R and r is obtained, the need for an accurate balance of r is small.

A series of observations made with experimental apparatus arranged as shown in Fig. 18 gave the sensitivity curves shown in Figs. 19 and 20. The observed sensitivity characteristics shown by these curves agree very favorably with the theoretical values shown by the curves of Figs. 16 and 17.

By way of summarization, it may be observed that in the process of developing a suitable method for the location of faults in telephone cables a definite sequence of steps has been taken to provide an effective treatment of the problem:

1. In establishing requirements for a suitable method, a study was

made of the historical development of the art. The effectiveness of the art was compared with the needs of present practices.

2. Preliminary to the development of a suitable method, an analysis was made of the errors which may enter into the determination of an open location. These errors were classified for treatment or correction during the development.

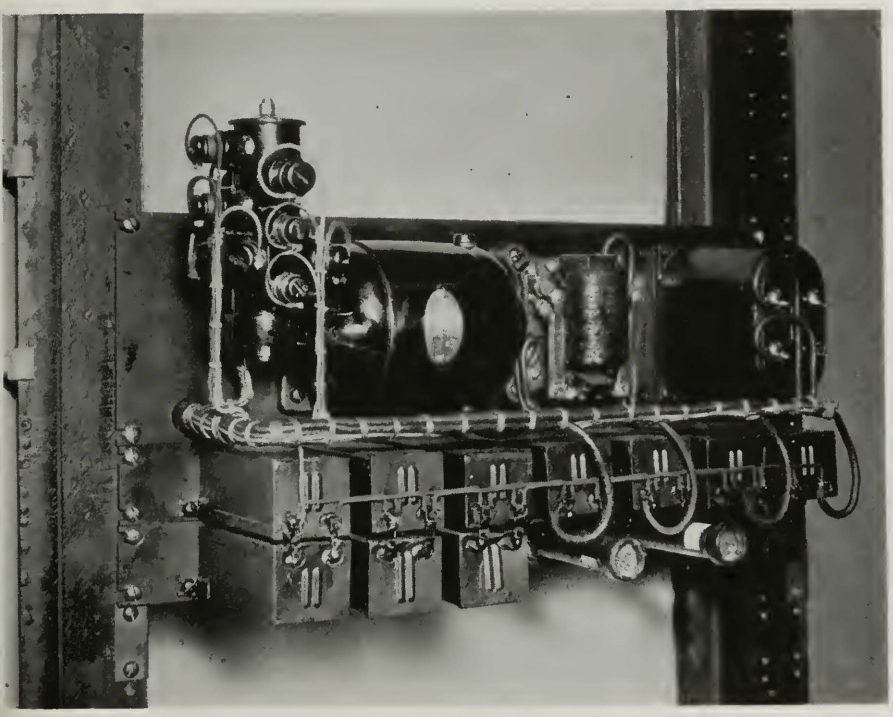


Fig. 22—4-cycle motor-driven commutator with associated equipment.

3. A mathematical analysis was made of the electrical characteristics of the circuits on which open locations may be required.

4. A suitable method was devised and an associated impedance bridge circuit was developed which, in the light of the recent research, most consistently and economically met the requirements for the determination of open locations.

After having completed an analysis of the problem and demonstrated the practicability of the proposed methods, it remained to develop applications of these methods for practical use. Equipment was designed to develop a low frequency source of alternating potential

by reversing a testing battery. The device developed for this purpose is a 4-cycle, motor-driven commutator shown in Fig. 21. By studying the curves used in the selection of a suitable frequency of testing potential, it may be observed that the selected value of four cycles is not critical, in fact a variation of ± 25 per cent may be allowable in

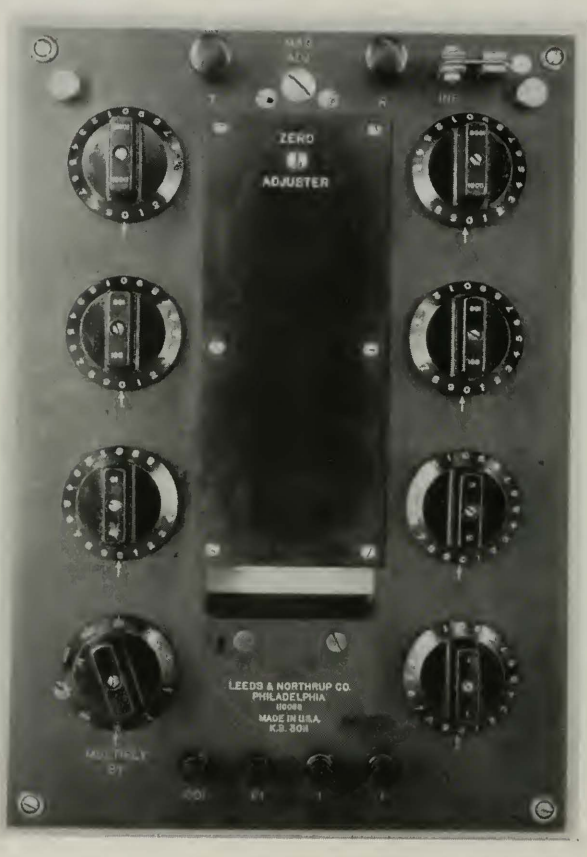


Fig. 23—Panel assembly of the impedance bridge.

different machines. However, in practical application, the accuracy of results obtained depends upon a comparison or ratio of two impedance measurements which cannot be made simultaneously. Since impedance varies with frequency, it is important that the frequency of testing potential should remain constant while the two measurements are being made. This requirement is met for the short time required for two measurements. The assembly of the 4-cycle commutator with associated apparatus is shown in Fig. 22.

The galvanometer and the equipment required in the modified form of the De Sauty bridge have been incorporated in a compact unit which is shown in Fig. 23. This bridge, while being particularly adapted to the 4-cycle impedance measurements required for open location tests, is also applicable for direct-current bridge measurements. Assembly details for the bridge arrangement are shown in Fig. 24.

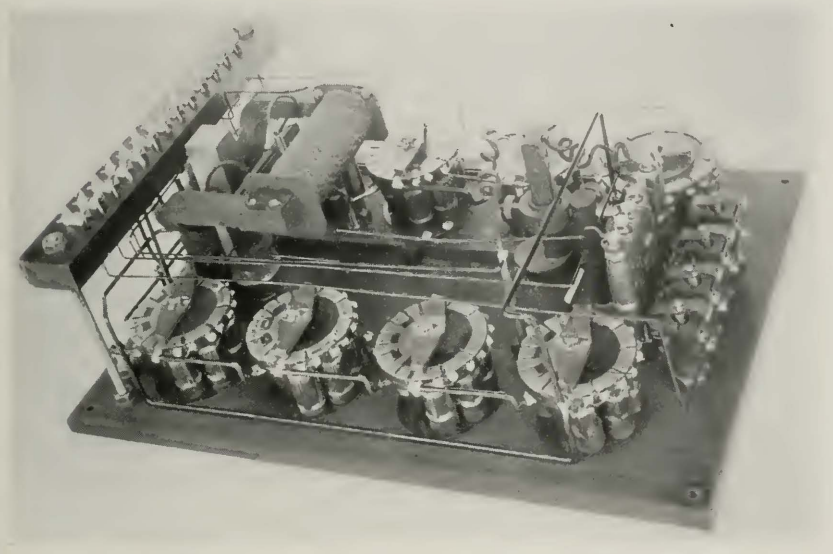


Fig. 24—Internal assembly of the impedance bridge.

The alternating-current galvanometer is sufficiently sensitive to be directly actuated by any significant unbalance of the impedance bridge, so that it is unnecessary to use an amplifier, rectifier, or other converting apparatus which may be difficult of adjustment or maintenance.

A few additional features are outlined as some of the significant results of this development of an improved open location method.

It has been shown that the error, caused by the deviation from the straight line relation between sending end admittance and physical length of line, has been reduced to a value which may be neglected as being less than the required accuracy of the open location method.

As a result of an analysis of errors which are introduced by line irregularities, methods were devised for applying corrections for all such errors which are of a magnitude sufficient to interfere with the desired accuracy.

This impedance bridge arrangement employs several features which are a distinct improvement on the methods previously utilized for this purpose. In order that the impedance angles of the impedance networks may be balanced, a variable resistance has been connected in that branch of the bridge which contains the comparison capacitance. This balance of the impedance angles of the impedance network was found to be important in obtaining a steady performance of the alternating-current galvanometer. A system has been devised for separately balancing the resistance components as well as the capacitance components of the reactive networks of the impedance bridge. It was found that this arrangement gives a maximum sensitivity to each component and permits of a very rapid balance of the bridge.

Contemporary Advances in Physics—XII. Radioactivity

By KARL K. DARROW

IN the year 1896, which fell near the beginning of the great transformation of modern physics, Henri Becquerel heard that Roentgen had discovered strange rays proceeding from an electric discharge-tube while the discharge was passing and the glass walls of the tube were phosphorescing. Suspecting that the new rays were connected with the phosphorescence, Becquerel tested samples of some of the substances which naturally phosphoresce. It happened that one which he tested was a compound of uranium. He wrapped the sample in paper to shut in the light of its phosphorescence, and set it beside a photographic plate; for the rays of Roentgen had disclosed themselves by acting on such plates. Becquerel had made a happy guess; for the compound affected the plate. Yet his original idea was altogether wrong; for the effect had nothing to do with the phosphorescence of the compound, it was due to the uranium itself and faithfully reappeared when other and non-phosphorescent compounds were used instead, and even when a piece of the pure metal was set beside the plate. It was an instance of a fallacious idea having guided a keen observer to a great discovery—not the first in the history of physics, and assuredly not the last.

Thereupon Pierre and Marie Curie, having verified that the effect of any quantity of any compound of pure uranium is strictly proportional to the amount of uranium in it, noticed that the effect of certain natural rocks and minerals containing uranium was much greater than that which their content of the metal should produce. Suspecting that there was some constituent of the rocks having the same property as the uranium but in a degree much greater, they set about the task of getting the uranium and the inert matter out of the way and isolating the more potent substance. It was a long task; to speak of "winnowing" the pile of rock would be to use a comically feeble metaphor, and as for the proverbial needle in a haystack, it could have been extracted with incomparably less trouble than the few hundredths of a gramme of the active substance which were latent in the ton of raw material. Eventually the Curies did liberate it, or rather them, for there were several active substances; and one of them was named radium, and their strange property was called radioactivity. This was the first of the words containing the magic syllables *radio*, syllables which are one of the special symbols of our epoch; were the

literature of these times to disappear all but a few scraps, posterity could date them by the appearance of that word, as Latin manuscripts are dated through containing some word or some trick of style that came into use at a definite moment of history. A word must bear almost magical connotations, to enter so thoroughly into popular usage; and the phenomena of radioactivity endowed it with these in abundance, with suggestions of rays piercing all matter, and inexhaustible stores of energy, and transmutation of elements, and influences having power even over life and death. Wonderful to relate, the suggestions for once were justified by the truth.

From 1898 onward there was a tremendous rush of investigators into the new field, and in a few years there were explorers of almost every conceivable aspect of radioactivity—chemists ascertaining the chemical properties of the radioactive elements. physicists observing their physical properties, and a great host of students investigating the numerous and striking effects of the rays. The subject presently became so wide that books on radioactivity written before and during the War resemble treatises on the contemporary physics of their dates of publication; for the new rays seemed to be able to invade all the provinces of physics as easily as they could penetrate matter in all its forms.

Eventually, however, it became clear that many of the topics classed at first with radioactivity should be removed into other fields of science. The radioactive elements all have their places in the Periodic Table, and their chemical properties are what should be expected from elements thus placed; peculiar as radium is in its one famous feature, there is nothing abnormal about its chemical reactions, and they may justly be relegated to the handbooks of chemistry and to the manuals written for those who wish to prepare or purify the element. The same thing is true of the physical properties of radium; nothing in its optical spectrum suggests that it is other than an ordinary member of the second column of the Periodic Table, nothing in its X-ray spectrum intimates that it is more than just the 88th member of the Procession of the Elements. None of these needs to be taken into account in the study of radioactivity.

The various effects of the rays which the radioelements emit are likewise quite irrelevant. At the beginning it was natural and proper for every writer to describe all that was known of the actions of the alpha-rays, the beta-rays and the gamma-rays, after having said that these are the three kinds of rays which radioactive substances emit. Indeed it was essential, for at first there was no way of defining the rays, much less of ascertaining their real nature, except by considering *en bloc*

everything that was known about their actions. This condition prevails no longer. It is established that alpha-rays are atoms of helium each bearing a charge $+2e$; that beta-rays are electrons, that gamma-rays are composed of electromagnetic radiation. Information about the first two belongs to the vast body of doctrine concerning the properties of fast-flying electrified particles; information about the last belongs to the science of the properties of radiation. I do not mean to imply that the information is redundant. One can produce in the laboratory fast-flying electrons, but none so fast as the fastest beta-rays; swift positively-charged atoms, but none nearly so swift as the alpha-particles; electromagnetic waves of many wavelengths, but none nearly so short as the shortest to be found among gamma-rays. The knowledge acquired from studying the properties of the rays is exceedingly important, and if the radioelements had not been discovered, most of it would not been acquired so early, and some of it would still be unattainable; but it is not knowledge of radioactivity.

What then is knowledge of radioactivity? So far as now appears, we know all that can be known about the radioactivity of a radioelement if we know what are the speeds of the alpha-particles emitted from it, if any; what are the speeds of the electrons emitted from it, if any; what are the wavelengths of the electromagnetic waves which it emits, if any; how many of each kind of particle (for we may speak of the waves as particles also, meaning by "particle" a quantum) are emitted from a given number of atoms in a given time; and what element or elements result from these processes. Apparently, if we could know all of these things for a particular radioelement, we should know everything which determines its peculiar actions upon the outside world. This unfortunately is not the same thing as being able to solve the problems of predicting all of these actions or understanding them; but these problems are now transferred out of the field of radioactivity into the field of the science of fast-flying charged particles and short-wave radiation. Let us leave them there, and restrict the field of radioactivity to the speeds of the particles and the frequencies of the waves which issue from each radioelement, and the rates at which they come forth, and the condition of the atoms they leave behind.[1]

[1] The specific statements made in this article are derived chiefly from three recent synopses of the data of radioactivity: the National Research Council bulletin *Radioactivity*, by A. F. Kovarik and L. W. McKeehan; the *Manual of Radioactivity*, by G. v. Hevesy and F. Paneth; and the relevant articles by St. Meyer, L. Meitner, W. Bothe and O. Hahn in volume 22 of the new Geiger-Scheel *Handbuch der Physik*. As these are all well supplied with bibliographies (and so likewise, I presume, is the new edition of Meyer and von Schweidler's *Radioaktivität*) I have omitted references to individual papers except a few published since 1923. At several places I venture to refer under the name "Introduction" to my *Introduction to Contemporary Physics* for topics not falling within the field of radioactivity as here defined.

During the composition of this article I have had the advantage of frequent consultation with my colleague Dr. L. W. McKeehan.

Already in expressing these restrictions, certain principles of radioactivity have been implicitly assumed; it is necessary to state them explicitly.

In the first place, I have spoken of the radioactivity of the elements alone; this is permissible, because radioactivity is definitely a property of individual elements. This does not mean merely that radioactivity is a property of a limited number of elements in certain states and a limited number of compounds of these and other elements, as seems to be true of ferromagnetism. It means that wherever there is a particular radioactive element, free or compounded, gaseous or liquid or solid, the characteristic rays of that element are emitted in a degree proportional to the amount of the element and not affected in the least by its condition or its state of combination. A given amount of radium emits the same kinds of rays at the same rate whether it is a piece of pure metal, or is combined with chlorine in radium chloride, or with sulphur and oxygen in radium sulphate. A given amount of radon emits rays of the same sort at the same rate whether it is gaseous as at normal temperatures, or frozen by submerging its enclosing tube in liquid air. Samples of some of the radioelements have passed through combination after combination in the chemical laboratory, being released from one compound only to enter into another; their activity was meanwhile being measured by the most delicate available tests, but it was never found to be affected in any perceptible degree. There is no other property of an element, excepting mass, of which this can be said without reservation.[2]

The indifference of radioactivity to the state of combination of the elements which display it extends also to all their other circumstances. In modern laboratories it is feasible to subject pieces of matter to very powerful, severe and violent agencies; heat enough to melt any element, cold enough to freeze any substance, electric fieldstrength high enough to tear electrons out, high magnetic fields, intense illumination, bombardment by multitudes of fast moving charged particles—and all of these have been tried to some extent, some to the utmost humanly possible extent, upon radioactive elements; but in every instance the radioactivity has remained constant without detectable variation, inaccessible and immune to all the powers within human control or under human observation.[3]

[2] It can be said of the higher-frequency emission-lines and absorption-edges of the X-ray spectra of the elements, but not unreservedly; for since the lower-frequency lines and edges of an element do vary slightly but perceptibly when its state of combination is altered, there is a strong presumption that the higher-frequency spectra will likewise be found to vary as soon as the accuracy of the measurements is increased say five- or tenfold.

[3] Influences of sunlight upon radioactivity are reported now and then in the *Comptes Rendus;* but it seems exceedingly unlikely that something immune to every other known agency should be susceptible to this particular one.

Sooner or later, in expounding almost any topic in physics, one arrives at a place where the introduction of an atom-model greatly simplifies what remains to be said. In the present article, this is the place.

Physicists commonly employ an atom-model in which a certain number of electrons are arranged around a nucleus bearing a charge equal in magnitude and opposite in sign to the sum of their charges. For any particular element the number of electrons assigned to its atom-model is equal to its atomic number, which can be obtained from any modern chart of the Periodic Table. In such a chart the elements are arranged in the order of their atomic numbers from 1 to 92, composing what I shall call the *procession of the elements*—a procession from which only two are now missing. In dealing with an element of high atomic number—all of the radioactive elements are of this character, ranging in atomic number from 81 upwards [4]—the electrons are assigned to various locations, some being close to the nucleus and others intermediate and others at the periphery of the atom-model. In fitting the various regions and divisions of the atom-model to the various properties of the element which it represents, the outermost electrons are assigned to the task of accounting for those properties which vary exceedingly with the state of chemical combination and with the other circumstances of the element; for being at the surface of the atom they should be most exposed to outer influences. The inner electrons, being partly shielded, are used to account for such properties as the X-ray frequencies, which depend so little upon the circumstances of the element that their variations are scarcely perceptible or not at all. The nucleus is the best shielded of all, and it receives for its quota the two properties which within the accuracy of experiment are immune from change—radioactivity and mass.

There are additional reasons for assigning mass and radioactivity to the nucleus. As for the mass: since the sum of the masses of the electrons constituting an atom-model never attains 1/1800 of the known mass of the atom, the balance which the nucleus must take is practically the whole of it. Again, there are experiments which show that a single chemical element may have several kinds of atoms differing in mass and yet quite alike in chemical properties, in their line-spectra, in their X-ray spectra; since these similarities require that the same nucleus-charge and the same number and arrangement of electrons be imposed upon all these atoms, the outstanding difference in their masses must be ascribed to their nuclei.[5] Again, there are slight

[4] Except potassium and rubidium (compare footnote 13).

[5] *Introduction*, pp. 29–39, 65–66.

differences between the band-spectra of compounds involving such atoms, which are well explained by attributing to the several atoms identical nucleus-charges and electron-systems, but nucleus-masses standing to one another in the same ratios as the observed masses of the atoms do.[6] But I must not give all the evidence for the nuclear atom-model, or this article will be swamped.

Among the reasons for ascribing radioactivity to the nucleus, the primary one has already been introduced—radioactivity, like mass, is unalterable; and another has already been stated, though without mentioning its relevance to this question. Certain radioactive elements emit charged atoms of helium; and since outside of the nucleus nothing except electrons is provided in the atom-model, these charged atoms must be supposed to proceed out of the nuclei. This argument could not be used upon the radioelements which emit electrons; but even for these there are reasons for suspecting that some of the electrons which issue from them do not come out of the family of electrons surrounding the nucleus, but from some other place. For instance, it is possible and usual to pry electrons out of various locations in the circumnuclear family; but when this is done, the resulting "ionized" atoms promptly take in one electron or as many more as they have lost, and revert to their original state and nature. This does not happen with the radioactive atoms which emit beta-rays; the departure of the electron effects an irreversible change, the atom is altered for good and all. It does not however acquire a permanent positive charge; it takes on an electron and makes good its loss of charge. This is best explained by supposing that the original atom lost an electron originally located in the nucleus, and added one to the circumnuclear family, keeping its net charge equal to zero but undergoing a rearrangement of its charges.

By accepting the idea that certain of the charged particles emerging from a radioactive element issue from the nuclei of its atoms, it is possible to express and explain very simply a celebrated law of radioactivity which was discovered by Fajans and Soddy in the early days of the nuclear atom-model and helped greatly to establish it.

When an atom of a radioelement of atomic number Z emits an alpha-particle with its charge $+ 2e$, its nuclear charge diminishes by that amount. It becomes an atom with nuclear charge $(Z - 2)e$ and Z electrons. The diminished nuclear charge cannot hold the entire electron-family; two of its members depart, and the atom becomes an atom of nuclear charge $(Z - 2)e$ and $(Z - 2)$ circumnuclear or orbital electrons. The radioelement changes into an element two steps farther down in the procession of the elements.

[6] *Introduction*, p. 400.

This is the first half of the displacement-law of Fajans and Soddy. It signifies that *the emission of α-particles by a radioelement is the sign of a transmutation of that element into the next but one of those preceding it in the procession of the elements.* If the properties of this latter element are known already, the law can be tested with all the accuracy desired. Polonium stands two places after lead in the procession; it emits alpha-particles; its atoms should turn into atoms possessing all the chemical qualities of lead, and they do. If a radioelement which lies two steps ahead of an element not previously known is discovered to emit alpha-particles, we are still not without information as to the qualities of the element into which it should transmute itself. For if we know the column of the Periodic Table in which the original element lies, we know also the column in which the element two steps ahead of it should lie; and the chemists know what features are common to all the known elements of that column and presumptively extend also to the unknown one. Radium lies in the second column of the Periodic Table; it emits alpha-particles; it should be transmuted into an element lying in the "zero" column. That element was not known until after radium was discovered; but it was known that the other elements of the zero column are inert gases, and consequently that the one into which radium transmutes itself should be an inert gas. This is verified; and as a general rule it is verified that when a radioelement emits alpha-particles the substance left behind possesses the particular chemical features of the elements belonging to that column of the Periodic Table to which the element two steps preceding the original one belongs. From this fact of experience it is only a short step to the first part of the Fajans-Soddy displacement-law—and a step which is put quite beyond criticism by the relations presently to be cited which connect the atomic weights of the radioelements.

The second half of the law relates to the other radioelements, those which eject electrons from their nuclei. When an atom of atomic number Z emits an electron from its nucleus, the nuclear charge increases to $(Z + 1)e$, which is sufficient to hold another electron beyond the Z electrons of the original family. The atom does pick up another electron which enters into the circumnuclear set (*not* into the nucleus); and it becomes an atom of nuclear charge $(Z + 1)e$ and $(Z + 1)$ orbital electrons. *The radioelement changes over into another which is one step farther up the procession of the elements.* This is the second half of the displacement-law of Fajans and Soddy.

The evidence for this second part is extensive; but on the whole it is not so imposing as the evidence for the first part. Largely this is

due to the difference between the two types of emission. Alpha-particle emission is violent and unique; positively-charged particles moving with a speed like theirs are not produced in any other way known to man. Beta-particle emission is considerably less violent, and there are so many known processes for producing fast-flying electrons that one must always keep in mind the possibility that some of the electrons proceeding from radioelements may be due to one or another of these; in fact, many certainly are. Perhaps the best way to state the evidence is this: every radioelement which does not emit alpha-particles transmutes itself into an element lying one step farther up the procession,[7] and all but one (actinium) of these elements is known to emit electrons, all of which agrees with the assumption that in each of these transmutations one electron is extruded from each participating nucleus. Stated thus, it may not sound very convincing; but if the second part of the Fajans-Soddy law were not true, we should hardly have failed thus far to find something definitely inconsistent with it.

Were gamma-rays without an accompanying beta-particle or alpha-particle to be emitted from a nucleus we could scarcely call the result a transmutation, since it would not affect the nuclear charge nor the electron-family of the atom. There is no reason for denying that this might happen; but I am not aware that it is known ever to happen, except in cases of nuclei which have just previously undergone a transmutation—cases which we shall eventually examine.

If now each radioelement is passing over into another element, one step before it or two steps behind it in the procession according as it emits beta-rays or alpha-rays—then it must be possible to draw up *genealogies* of radioelements, series of elements of which each member is transmuted out of the foregoing and transmutes itself into the following one. All of the known radioelements fall into one or another of several such series. To represent all these relations, and one more, it is convenient and suitable to draw a graph in which the atomic numbers of the elements are laid off horizontally, and their atomic weights are laid off vertically. Each element is represented by a point upon this graph; when the element transmutes itself it moves to another point, two units to the left if an alpha-particle is emitted and one to the right if the change is a beta-ray change. Now the emission of an alpha-particle involves the departure of four units of mass from the nucleus which it leaves; the loss of an electron however involves a loss

[7] More precisely, into an element having the chemical features distinguishing the column of the Periodic Table containing the element lying one step farther up the procession than the original one.

of only 1/1850 of a unit of mass, which is quite inappreciable.[8] Hence in a transmutation of the former sort, the point representing the element in the graph moves four units downward as well as two to the left; in one of the latter sort, the point simply slides horizontally to the right. The meaning of Fig. 1 will now be clear.

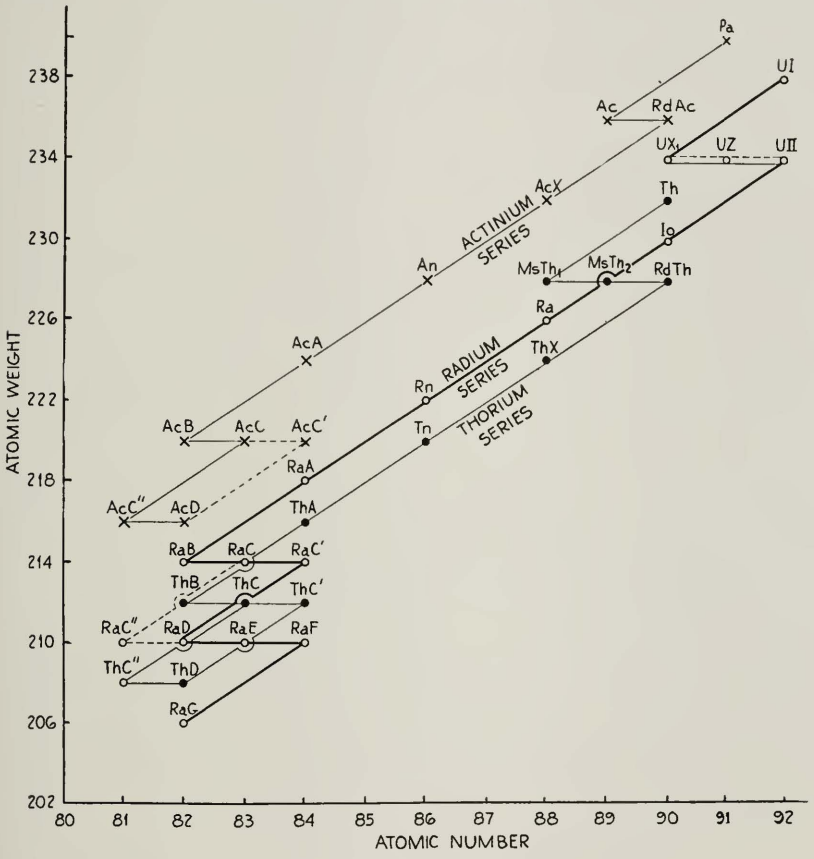


Fig. 1. Genealogies of the radioelements

(The actinium series is plotted some distance above the others for legibility, but should almost certainly lie lower.)

The lines in Fig. 1 which represent the family trees of the radioelements descend in zigzags, which signifies that the "decline and fall"

[8] These masses are given in terms of the unit of mass in which atomic weights are measured, of which 16 constitute the mass of an oxygen atom. Were the mass of an electron appreciable in these measurements, we should have to allow for the electrons added to or lost from the circumnuclear family to balance the change in the nucleus-charge. But then we should also have to make decisions about the mass to be assigned to the energy carried away by the particles and the waves.

of a radioactive atom does not proceed continually downward to lower and ever lower atomic numbers, but is interrupted by occasional partial recoveries. Whenever there are three consecutive transmutations of which two involve the emission of beta-rays and one that of alpha-rays, the element in which the third ends has the same atomic number as the element from which the first originates. In each of the three lines of descent made visible in Fig. 1 there are instances of this; the one including radium for instance touches three times at atomic number 82, the one commencing with thorium twice. In the sense in which I have thus far used the word "element," the element 84 recurs three times in one series and twice in the other. Here is an ambiguity which the time has come to dispel.

The ambiguity in the use of the term *element* is a question of words, but not wholly a linguistic, much less a trivial one; it is such a question as arises when a field of knowledge is expanded and enriched to such an extent that its old vocabulary ceases to be adequate. This particular question arose after the discovery that certain substances differing in radioactivity are very much alike in their chemical properties—another of the facts which the atom-model is especially adapted to explain.

Consider, for an example, the three elements radium B and radium D and radium G, which lie upon the same line of descent, the "radium series." The first transmutes itself into the second, and the second transmutes itself into the third, each in a three-stage process involving the departures of two electrons and an alpha-particle (with the order of their exits we are not now concerned) from the nucleus. The mass of the third is four units less than that of the second and eight units less than that of the first; in radioactivity also they differ. But all three are alike in nuclear charge, and hence in the size and presumably in the arrangement of their circumnuclear families of electrons; and hence the presumption arises, that in their physical and chemical properties apart from radioactivity and mass they should be quite alike.

This presumption about the chemical properties is confirmed by the fact that RaB and RaD and RaG cannot be separated from one another by chemical means after they are once mixed. In general, whenever two of these "radioelements" coinciding in atomic number are subjected to any of the very considerable variety of agents in the chemists' armory, they respond in so nearly, if not exactly, the same way that there is no method known for taking one and leaving the other. Crystallization out of a mixture of salts of two such elements merely produces crystals containing the two salts in the same proportion as the liquid; sublimation merely produces a deposit containing the elements

in their original ratio; electrolysis does not favor one above the other, nor does osmosis, and if a mixture of the two elements or of salts of theirs is presented to an absorbent or an adsorbent or a solvent, it accepts them in the same proportions as they are presented, while any element willing to react with either reacts in precisely the same degree with the other.[9] Similarity such as this goes far beyond the interresemblances of the alkali metals, for instance, or even those of the rare-earth elements, difficult as the task of separating these latter from one another sometimes proves; it is not similarity merely, it amounts to identity.

As for the presumption that radioelements sharing the same atomic number should be alike in what are loosely termed the "physical properties," it is more difficult to test. In fact, there seems to be no instance of two such elements, each radioactive and each obtainable quite unmixed with the other in quantities large enough for such experiments. The three elements sharing atomic number 86 are all gaseous at ordinary temperatures, but they are too scanty and two of the three are much too fugitive for making accurate comparative measurements of such qualities as viscosity or elasticity or ionizing potential. The elements sharing atomic number 82 are, as I shall presently bring out, mostly stable, and upon them it is possible to test the expected coincidence in optical line-spectra and X-ray spectra, which is verified except for certain very minute (but unexplained!) differences in the wavelengths of certain lines. The band-spectra of these elements (more precisely, of their compounds) display slight differences which are beautifully explained by the contemporary theory of band-spectra, involving as it does a participation of the nucleus with its mass in the production of the bands.[10] Mixtures of two of the elements sharing atomic number 90 (thorium and ionium) display precisely the same optical spectrum as pure thorium.[11] In addition to these observations, a great many have been made upon the physical

[9] There is a huge literature of the attempts to separate elements of identical atomic number and to discriminate between their chemical properties, for a review and bibliography of which I refer again to von Hevesy and Paneth (*l.c. supra*, chapter XII).

[10] E. S. Bieler, *Nature*, **115**, p. 980 (1925).

[11] This is vividly illustrated by a passage in the classical treatise upon radioactivity which Rutherford wrote in 1912. Boltwood had isolated from uranium ores a sample of thorium oxide which emitted, along with the alpha-particles from the thorium, a considerable number coming from ionium. Russell and Rossi produced its arc spectrum and "the spectrum of thorium was obtained, but not a single line was observed that could be attributed to ionium. On the assumption that ionium has a life of 100,000 years, the preparation should have contained 10% of ionium. Since probably the presence of 1% of ionium would have been detected spectroscopically, it would appear that the ionium was present in small amount, indicating that the life of ionium must be much less than 100,000 years." As a matter of fact there was probably more than 10% of ionium in the mixture; but its spectrum lines were identical with those of thorium.

properties of other non-radioactive elements which share particular atomic numbers and are mixed together in varying proportions; and they establish that such "elements" are indistinguishable except in such properties as are influenced to a measurable extent by the mass of the nucleus.

These facts make it necessary to redefine the word *element*, which in its long journey through the centuries from Lucretius has modified its meaning time and time again to keep pace with the gradual refinement of scientific thought, though all the while it kept its spelling intact. These are the alternatives: *either* to confer the status of a separate "element" upon each substance (apart, of course, from the compounds!) possessing a distinctive mass and radioactivity of its own, so that there may be several distinct elements sharing a given set of chemical properties— *or* to link the term "element" with a characteristic set of chemical and physical properties, with a specific atomic number and position in the Periodic Table, so that a given element may be an ensemble of several different kinds of matter differing in radioactivity or mass or both. Reasons of science require that one or the other of these alternatives be chosen, but the actual choice is determined by reasons of language and expediency. These reasons—I will not pause to develop them—favor the second alternative. Inconvenient though it may be to refer to RaB and RaD and RaG and ThB and several other radioactive substances as the same element, the inconveniences entailed by the other policy would in the end be immensely greater. One element to each atomic number, one place in the Periodic Table to each element—this is the choice which the prior usage and the associations of the word *element* recommend; and some other name must be selected to distinguish the substances which share a common atomic number but differ in mass or radioactivity or both.

Such a name is Soddy's word *isotope*, constructed out of Greek words to signify "in the same place." Radium B and RaD and all the other substances which appear in the column labelled "82" in Fig. 1 are isotopes of the element 82; radon and thoron and actinon are isotopes of the element 86. In these eleven places of the Periodic Table extending from 81 to 92, the individual isotopes enjoy names of their own, the elements are best known by their numbers. The names *thallium*, *lead*, *bismuth* and *uranium* are, it is true, generally attached to the elements 81, 82, 83 and 92; but the first three of these names are used by some people to mean the elements in question and by others to designate only those of their isotopes which are not radioactive, and there is danger of confusion.[12] Elsewhere in the Periodic Table, where

[12] The names *polonium*, *radium*, *actinium*, *thorium* and *protactinium* signify par-

all the isotopes of each element are stable, the elements have individual names and the isotopes are designated only by their masses. The elements 81, 82 and 83 have some isotopes which are radioactive and others which are not; thus the word "radioelement" is misleading, and should be replaced by "radioactive isotope." Consistency indeed requires that one speak of the successive members of a family of radioactive substances not as consecutive elements, but as consecutive isotopes of diverse elements. At this point however consistency almost ceases to be a jewel. I can find no satisfactory compromise, and will hereafter refer to the various radioactive materials simply as "substances"—so bringing to an end this long analysis of words, which is justified only in so far as it may have concentrated the reader's attention upon the facts underlying them.

We return to Fig. 1.

The radioactive substances are grouped into three main lines of descent or sequences, commonly called *series*. Each of these throws off one or two branches, which however cannot be followed far; these I will discuss further on, pausing here only to mention that one of the three main sequences, the *actinium series*, is believed by many to branch in this manner out of the *uranium-radium series*. This however is not certainly established, and it is suitable to regard these two and the *thorium series* as independent sequences, which between them comprise all the known radioactive isotopes among the elements.[13]

Uranium and thorium, the first elements of the series to which they have given their names, are even yet after all the aeons of the earth's existence to be found in abundance among its rocks. This practically proves that uranium, at least, disintegrates with exceeding slowness; for all the other known elements are lighter than it is, and consequently there is none of them out of which the steadily-dwindling supply of uranium might be replenished by transmutation. We shall presently learn methods of estimating the duration of uranium, by which it is shown to be truly colossal.

The atomic weights of uranium and thorium are known, and amount to 238.18 and 232.12 respectively. From these it should be possible

ticular isotopes of the elements 84, 88, 89, 90 and 91 respectively, but are sometimes used as names for these elements—another dangerous source of misunderstanding. The name *niton* was formerly used for the isotope *radon* of element 86, and might well be used for this element now that the isotopes are individually named.

[13] Apart from the elements potassium and rubidium, which will continually demand to be mentioned as exceptions unless they are disposed of once for all at this point. Let it be stated, then, that these elements emit electrons, so feebly however that they are much less active than even uranium, which ranks among the least radioactive of all the known radioactive substances; and that no one has identified the substances into which they are transmuted, though presumably those are isotopes of calcium and strontium respectively. Cf. an account of the radioactivity of these elements by A. Holmes and R. W. Lawson: *Phil. Mag.* (7) **2**, pp. 1218–1233 (1926).

to deduce the atomic weights of all the other members of the two sequences; thus, a radium atom is what is left behind after a uranium atom has ejected three alpha-particles (mass, 4 apiece) and two electrons (mass negligible) and its atomic weight should therefore be 226.18. Here we meet a troublous fact. The value of the atomic weight of radium, as measured by no less an expert than the celebrated Hönigschmid, is 225.97 with an uncertainty believed not to exceed three units in the last place. This might be explained by supposing that the *element* uranium as found in nature is a mixture of several isotopes in relatively large proportions, only one of which is the parent of the uranium-radium series, while the others may be stable or perchance the ancestors of the other series; indeed it is hard to think of any other adequate explanation.[14]

All three of the sequences terminate in isotopes of the element 82, commonly known (but remember the caution on page 110!) as lead. It is a curious fact that the most rare and precious of all substances should die away by self-transmutation into the one which serves as the symbol for everything which is commonplace, dull and cheap. The atomic weights of the terminating isotopes of the radium and thorium sequences may be guessed in the same manner as that of radium from that of uranium. Starting from radium and from thorium respectively and noting that an atom of radium is destined to eject five alpha-particles and an atom of thorium six during the transformations whereby they turn into atoms of RaG and ThD respectively, we calculate the values 206.0 and 208.1 for the atomic weights of these two isotopes of element 82. Now nearly every sample of lead that has ever served for an atomic-weight determination has yielded a value near 207.2. Yet, when the lead-content of certain minerals rich in uranium and its posterity and deficient in thorium was extracted and investigated, the atomic weights of these samples were found to lie extremely near to 206—some of the values recorded are 206.046, 206.048 and 206.08. On the other hand, samples of lead extracted from various minerals rich in thorium and poor in uranium displayed abnormally high atomic weights, values attaining in some instances to 207.9. These are data much more dramatic than the customary outcome of the tedious process of determining an atomic weight; one wonders vainly what chemists would have felt, if they had been published before

[14] What is commonly called "uranium" contains not only the ancestor of the uranium-radium series, but also one of its descendants, which however is not present in sufficient amount to affect the atomic weight. This is the reason for inserting the words "in relatively large proportions" in the above sentence. The fact that the atomic weight of uranium is not integral might be taken to suggest that it is a mixture of integral-weight isotopes. Aston's latest experiments on stable elements of non-integral atomic weight show, however, that the premise does not necessarily lead to the conclusion.

radioactivity was discovered. They disclose the only known instance of distinct stable isotopes of an element being found separately from one another in nature. Whether "ordinary lead" of atomic weight 207.2 is a mixture of these two isotopes, or contains still others, is as yet an unsolved question.[15]

The three series resemble one another not only in the nature of their terminal substances, but in other regards as well. The substance in the radium series known as ionium, the member of the actinium series called radioactinium, the member of the thorium series named radiothorium, are isotopes all three of the element 90; and these three substances evolve through the same succession of transformations, alpha-ray emissions and beta-ray emissions following after one another in the same order. The nth descendants of these three substances, for each value of n from 1 to 6, are isotopic with one another—a statement which will probably be made clearer by Fig. 1 than by these words. This parallelism, which from the grandchildren of Io and RdAc and RdTh onward is reflected in the names of the substances, includes also the "branchings" which occur in each sequence at the substance labelled C—radium C and actinium C and thorium C. It is limited in its range, for the earlier parts of the three sequences are by no means alike, while the radium sequence continues onward for three stages longer than the two others. Something within the radium atom impels it to continue evolving even after it has twice taken and left the atomic number which it is destined eventually to take and keep, although the atoms which were once actinium or thorium are contented to stop at the atomic number 82 when for the second time they reach it.

The phenomenon of *branching*, which I have twice casually mentioned, is worthy of a few paragraphs. It signifies that a certain proportion of the atoms of such a substance as (for instance) thorium C transmute themselves in one fashion, the remainder in another. Sixty-five per cent of the atoms of ThC extant at any moment are destined to emit beta-rays and become atoms of a substance ThC′ lying one step further up the procession of the elements; the other thirty-five per cent eventually emit alpha-particles and become atoms of ThC″ placed two steps further down the procession. Such a "dual transmutation" occurs also at RaC and at AcC—an instance of the parallelism just mentioned, which however does not extend to the relative frequency of the two modes of transformation; 9996 out of ten thousand atoms of RaC, but only three out of a thousand atoms of

[15] Not however a definitely insoluble question, since the Thomson-Aston method of resolving mixtures of isotopes (*Introduction*, pp. 14–29) and measuring their individual masses should be applicable to lead—that is to say, certain difficulties have thus far prevented it from being applied to the very heavy elements, but these difficulties may not prove insuperable.

AcC, transmute themselves by ejecting electrons. As the disintegration of a sample of any of these substances proceeds, the relative proportions of the atoms disintegrating in the two ways remain unchanging. This makes it seem inadvisable to describe ThC (for instance) as a mixture of two distinct substances; rather it appears that the atoms may be all alike, but the destiny of each particular atom is a matter of "chance," with the chances favoring one type of disintegration over the other by nearly two to one. This is not the only circumstance in radioactivity which suggests the operations of "chance."

The substances labelled C′ and C″, which result from the dual disintegration of any of the three substances labelled C, differ in atomic weight and in atomic number, and in radioactivity as well; for the C′ substances which were born out of beta-ray transformations emit alpha-rays, while the C″ substances which resulted from alpha-ray transmutations send forth beta-rays. Consequently their immediate descendants, the two grandchildren of each C-substance, are isotopes with one another—and isotopes which should be alike not only in atomic number but in atomic weight as well. Is there any respect in which they differ? We cannot tell. Both of the grandchildren of ThC are apparently non-radioactive and stable; probably they are one and the same isotope of lead. Both grandchildren of AcC likewise seem to be stable. The predominant grandchild of RaC is the radioactive substance RaD; but in this case the number of atoms of RaC electing the less popular path of disintegration is so exceedingly small that we can neither discern any distinctive radiation to be ascribed to a substance isotopic with RaD but distinct from it, nor yet conclude from our failure that no such substance exists. Concerning the fourth of the known branchings, which occurs at UX_1, the state of affairs is the same as with RaC; we can neither detect more than one kind of grandchild, nor be sure that there is only one. In this case, by the way, both modes of transmutation of the parent element involve the emission of beta-rays.

Although among the four substances which are known to disintegrate in two alternative ways there is thus none for which both of the two lines of posterity can be traced through more than two generations, it is believed by many that there must be a fifth such substance in the uranium series, from which the actinium series goes off as a branch while the main proportion of the atoms continue evolving down the radium sequence. The reason for this idea is that in the ores of uranium the members of the actinium sequence are as a rule to be found about three per cent as abundantly as the members of the radium sequence. This fact could be deduced by assuming that

uranium II suffers a dual alpha-ray disintegration, about 97 per cent of the atoms transmuting themselves into ionium and the other 3 per cent into the mysterious substance uranium Y which is always found mixed with uranium, and which is known to emit beta-rays and hence to pass over into an isotope of element 91 which may well be protactinium, the first known member of the actinium series. On following out these presumptive transformations in Fig. 1 the reader will see that they would lead to the actually-observed result; but that is not quite the same thing as proving that the observed result is attained in just that way. The branching may occur elsewhere in the posterity of uranium; or the observed constancy of the ratio of actinium to radium in the rocks may mean that actinium and its family all descend from a separate isotope of element 92, not concerned in the production of radium. Much light would be shed upon this question if someone would only determine the atomic weight of even one member of the actinium sequence—an achievement which would settle at once those of all the others, and is most eagerly awaited.

Having dealt with the filiation of the radioactive substances, having specified the substance from which each is born and the substance to which it gives birth, and the sort of particle which is emitted in each process of transmutation, it remains to specify the rates at which the transmutations occur, and the speeds of the particles which are emitted, and the wavelengths of the quanta of radiation which sometimes come out also, and how many there are of these. The fundamental assumptions of the theory of radioactivity, which the experiments have sustained, require that in a transmutation only one alpha-particle or one beta-ray be emitted from the nucleus of one self-transmuting atom; but there is no such limitation upon the radiation-quanta, nor upon the electrons incidentally ejected from the circumnuclear family.

The rate of transmutation of every radioactive substance, so far as we know, is governed by the famous exponential law which signifies that *equal fractions perish in equal times*—that if one were to take a sample of the substance and determine the quantities extant at two instants an hour apart, and also those existing at two other instants an hour apart, and at any number of pairs of instants separated by intervals of one hour, then the mutual ratios of the two measured values of all those pairs would be the same. Half of any sample of thorium C transmutes itself in one hour; half the remainder in the next hour; half the remainder in the next hour, and so forth *ad infinitum* (or, to speak more carefully, up to the limit of the observations).

This law is described by the following formula relating the quantity Q of the substance existing at any time t, and the quantity Q_0 existing

at any other time t_0 (provided that no replenishment of the supply is taking place!):

$$Q = Q_0 \exp\left(\frac{t_0 - t}{\tau}\right). \tag{1}$$

Furthermore the rate dQ/dt at which the substance is being transmuted at any instant is related to the amount Q existing at that instant as follows:

$$dQ/dt = -\frac{Q}{\tau} = -\frac{Q_0}{\tau}\exp\left(\frac{t_0 - t}{\tau}\right). \tag{2}$$

These formulæ contain only a single constant characteristic of the substance. Nothing simpler could be desired. A phenomenon that

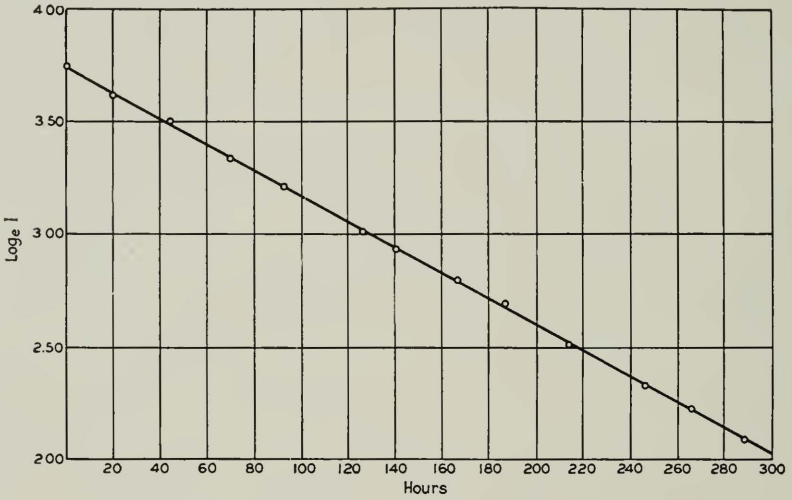


Fig. 2. Decay-curve of radium E (R. F. Curtiss)

(Being plotted on logarithmic paper, the graph of the exponential law is a straight line.)

can be described by a formula involving only one constant which has to be varied to distinguish one case from another is a rare gift of nature.

While the equations (1) and (2) are naturally valid whatever the unit in which we choose to measure Q, it is desirable as a rule (and necessary, in comparing the radioactivity of different substances) to express Q either in gramme-molecules, or in actual numbers of atoms. In some places I shall use N as a symbol for Q measured in the latter manner.

The exponential law is a law of chance. It may be expressed by saying that the chance of an atom disintegrating within a given time-

interval is precisely the same, whichever atom one chooses to consider and whenever the instant at which one chooses to let the given time-interval begin. I will quote a passage from Poincaré [16], taking only the liberty of writing 'nucleus' where he wrote 'atom.' "If we reflect on the form of the exponential law, we see that it is a statistical law; we recognize the imprint of chance. In this case of radioactivity, the influence of chance is not due to haphazard encounters between atoms or other haphazard external agencies. The causes of the transmutation, I mean the immediate cause as well as the underlying one (*la cause occasionnelle aussi bien que la cause profonde*) are to be found in the interior of the atom [*read*, in the nucleus]; for otherwise, external circumstances would affect the value of the coefficient in the exponent. . . . The chance which governs these transmutations is therefore internal; that is to say, the nucleus of the radioactive substance is a world, and a world subject to chance. But, take note! to say 'chance' is the same as to say 'large numbers'—a world built of a small number of parts will obey laws which are more or less complicated, but not statistical. Hence the nucleus must be a complicated world. . . ." I shall make no further allusion to theories of radioactivity.[17]

The constant τ may be interpreted as the time-interval during which the fraction $\frac{e-1}{e}Q$ (or approximately $0.632Q$) of any initially-present quantity Q of the substance would undergo its change. It is greater by the factor $1/\log_e 2$ (or approximately 1.44) than the *half-period* of the substance, the interval (designated by T) during which one half of the initially extant atoms are transmuted. It is also the average duration of the life of a single atom. All of these statements may be proved without difficulty from the formula (1). From the similarity between (1) and (2) it follows that the rate at which transmutations occur in an unreplenished sample of a radioactive substance, and the rate at which rays shoot out of such a sample, and the intensity of all the effects which the rays produce, vary exponentially with time; and the constant T which is the half-period for the extant quantity of the substance is likewise the half-period for all of these. The constant τ likewise has the same meaning for them all, and so does its reciprocal

[16] *Dernières Pensées*, pp. 204–205; he credits Debierne with the idea.

[17] Further and very valuable evidence that the transmutations of individual atoms are governed by the "laws of chance" operating within their own nuclei is furnished by the variations or fluctuations (*Schwankungen*) of the numbers of alpha-particles emitted from a sample of any radioactive substance in consecutive equal time-intervals very short compared with the half-period of the substance (*v.i.*). These are precisely analogous to the fluctuations in thermionic emission known by the name of "Schroteffekt" (*Introduction*, p. 10). Consult an article by K. W. F. Kohlrausch in *Ergebnisse der exakten Naturwissenschaften*, 5 (1926).

λ, which is called the *disintegration constant*, and is often specified instead of τ or T.

The values of the half-period T for the various radioactive substances are collated in the accompanying Table, which contains also the names of the substances, their usual symbols (those used in Fig.1), the symbols embodying their atomic numbers proposed by Kovarik and McKeehan, and the types of particle which they emit from their nuclei.

SYMBOLS, NAMES AND HALF-PERIODS OF THE RADIOACTIVE SUBSTANCES

The first column of this Table contains the usual symbols for the substances; the third, their usual names; the fourth, their half-periods as collated by A. F. Kovarik and L. W. McKeehan (*l. c. supra*); the fifth, the nature of the particles which they emit at transmutation. In the second column, the symbols proposed by Kovarik and McKeehan are given; each is composed of the atomic number of the substance, of a symbol denoting the series to which it belongs, and sometimes of a second numeral which, when the substance is an isotope of one or more others in the same series, denotes whether it is the first, second or third of these isotopes reached in the course of the transmutations. In cases of branching, the less common of the two resulting substances is italicized. The annotation *est.* signifies that the half-periods in question are estimated by extrapolating the Geiger-Nuttall relation (*v.i.*). The abbreviations *s*, *m*, *d*, *a* stand for second, minute, day, year.

URANIUM-RADIUM SERIES:				
UI	92UI	Uranium I	$4.6 \cdot 10^9 a$	α
UX_1	90UI	Uranium X_1	$24.5d$	β to UX_2, β to UZ
UX_2	91U	Uranium X_2	$1.138m$	β
UZ	*91Ua*	*Uranium Z*	*6.69h*	β
UII	92UII	Uranium II	$1.2 \cdot 10^8 a$ (*est.*)	α
Io	90Ra	Ionium	$7.43 \cdot 10^4 a$	α
Ra	88Ra	Radium	$1.69 \cdot 10^3 a$	α
Rn	86Ra	Radon, radium emanation	$3.810d$	α
RaA	84RaI	Radium A	$3.0m$	α
RaB	82RaI	Radium B	$26.8m$	β
RaC	83RaI	Radium C	$19.5m$	β to RaC′, α to RaC″
RaC′	84RaII	Radium C′	$10^{-6}s$	α
RaC″	*81Ra*	*Radium C″*	*1.32m*	β
RaD	82RaII	Radium D	$16a$	β
RaE	83RaII	Radium E	$4.85d$	β
RaF	84RaIII	Radium F, polonium	$136.3d$	α
RaG	82RaIII	Radium G, radium lead apparently stable		
THORIUM SERIES:				
Th	90ThI	Thorium	$1.3 \cdot 10^{10} a$	α
MsTh1	88ThI	Mesothorium 1	$6.7a$	β
MsTh2	89Th	Mesothorium 2	$6.20h$	β
RdTh	90ThII	Radiothorium	$1.90a$	α
ThX	88ThII	Thorium X	$3.64d$	α

Tn	86Th	Thoron, thorium emanation	54.5*s*	α
ThA	84ThI	Thorium A	0.145*s*	α
ThB	82ThI	Thorium B	10.6*h*	β
ThC	83ThI	Thorium C	60.6*m*	β to ThC′, α to ThC″
ThC′	84ThII	Thorium C′	$10^{-11}s$ (*est.*)	α
ThC″	*81Th*	*Thorium C″*	*3.20m*	β
ThD	82ThII	Thorium D, thorium lead apparently stable		

Actinium Series:

Pa	91Ac	protactinium	$1.6 \cdot 10^4 a$	α
Ac	89Ac	actinium	20*a*	β
RdAc	90Ac	radioactinium	18.9*d*	α
AcX	88Ac	actinium X	11.2*d*	α
An	86Ac	actinon, actinium emanation	3.92*s*	α
AcA	84AcI	actinium A	2.00*s*	α
AcB	82AcI	actinium B	36.1*m*	β
AcC	83Ac	actinium C	2.16*m*	α to AcC″, β to AcC′
AcC′	*84AcII*	*actinium C′*	*.009s*	α
AcC″	81Ac	actinium C″	4.71*m*	β
AcD	82AcII	actinium D, actinium lead apparently stable		

UY, K, Rb not assigned to series. They emit beta-rays, and their half-periods are given respectively as 24.6*h* (St. Meyer, *l. c.* footnote 1), $1.5 \cdot 10^{+12} a$ and $10^{11} a$ (Holmes and Lawson, *l. c.* footnote 13).

To measure a disintegration-constant seems an easy task, since one has only to choose the most convenient effect of the rays of the substance in question, and measure it at sufficiently many times to establish a sufficiently long arc of its decay-curve. Yet there is, I suppose, no other problem of which the general solution involves as many of the typical difficulties of research in this field; partly because some of the half-periods to be measured are so exceedingly short and some so tremendously long, largely because no radioactive substance ever exists by itself. Some can be separated completely from their ancestors, but none can ever be totally isolated from its posterity, especially since its rate of producing its posterity is the very thing which is being measured. Its own gradually-declining rays are mixed with the gradually-augmenting rays of its descendants, and while the specific effects of the former can indeed in some cases be distinguished from those of the latter, this is often difficult and sometimes impracticable. Frequently the observer is required to deduce the half-periods of individual substances from observations upon a continually-changing mixture; and most of the mathematical formulæ used in the study of radioactivity are developed out of equations (1) and (2) for interpreting such observations, or inversely for predicting the evolution

of a mixture of substances of which the initial composition is taken for granted. There is no better way of conveying a notion of the methods by which radioactivity was and is studied than to describe how some of the known half-periods were actually ascertained.

The simplest of all the cases are those in which a substance which can easily be separated from its ancestors transmutes itself into one which either is not radioactive at all, or else decays so slowly that the rays which it emits are not strong enough to interfere with the observations on the rays of its parent. The penultimate substances of the various series are candidates for this class, but the only one among them which is abundant enough and lasts long enough to be easily isolated from its ancestors is radium F, otherwise known as polonium. This therefore is the classical instance of a substance of which the decay-curve is determined directly from observations on rays of its own. Another is radium E, of which the half-period is so short (about 5 days) and the half-period of its daughter-substance so long (more than four months) that its decay-curve can be traced practically as if it changed into a stable element (Fig. 2).

Almost as simple are certain cases in which a radioactive substance is isolated both from its ancestors and from its posterity, and then the growth of its immediate descendant is measured. This method is available when the parent-substance is much longer-lived than its child, so that the rate at which atoms of the latter come into being is practically constant throughout the period of observation. Let B represent this rate; let Q represent the quantity of the daughter-substance extant at any time t, the time being measured from the instant when the isolation of the parent-substance is perfected, so that $Q = 0$ at $t = 0$; let λ stand for the disintegration-constant of the daughter-substance, so that the rate at which its atoms are disappearing through transmutation is equal to λQ. The net rate of growth of the daughter-substance is therefore

$$dQ/dt = -\lambda Q + B \tag{3}$$

from which we obtain by integration

$$Q = \frac{B}{\lambda}(1 - e^{-\lambda t}), \tag{4}$$

so that the quantity of the daughter-substance, and the intensity of its rays vary as exponential functions of time with the disintegration-constant standing in the exponent. This function, it is true, rises from zero to a positive final limiting-value instead of falling to zero from a positive initial value, as the decay-curve would; but the value

of λ is determined from it quite easily, and as a matter of fact the decay-curve itself can be obtained merely by plotting as function of time the difference between Q and the limiting-value ($= B/\lambda$) which Q approaches as t increases indefinitely. Determining a half-period from a rate of growth is therefore mathematically the same process as determining it from a rate of decay. This is one of the ways in which the half-period of uranium X_1 is measured; and the standard method for determining that of radium is based partly upon it, as we shall presently see.

Eventually the grandchild and the remoter posterity of the parent-substance must make their presence known. This is not always a disadvantage. Letting λ_1 and Q_1 stand for the disintegration-constant and the extant quantity of the daughter-substance, λ_2 and Q_2 for those of the granddaughter, we have as basis for the theory these equations:

$$dQ_1/dt = B - \lambda_1 Q_1, \qquad dQ_2/dt = \lambda_1 Q_1 - \lambda_2 Q_2, \tag{5}$$

integrating which, and supposing that at $t = 0$ the parent-substance has just been isolated so that the building-up of the two descendants from zero is just commencing, we obtain for Q_1 the expression (4) with λ_1 in the place of λ, and for Q_2 the function

$$Q_2 = B\left[\frac{1}{\lambda_2} + \frac{1}{\lambda_1 - \lambda_2} e^{-\lambda_1 t} - \frac{\lambda_1}{\lambda_2(\lambda_1 - \lambda_2)} e^{-\lambda_2 t}\right], \tag{6}$$

which to second approximation is equivalent to

$$Q_2 = \frac{1}{2} B\lambda_1 t^2. \tag{7}$$

The amount of the grandchild therefore should increase at first as the square of the time elapsed, whereas the amount of the child increases proportionally to the time. There are instances, in the history of the study of radioactivity, of a substance being regarded as the child of another until measurements were made upon its rate of growth in an isolated sample of its putative parent, whereupon through its conformity to (7) it was proved to be the grandchild and not the child. The question whether radium comes directly out of uranium II, or out of an intermediate substance, was settled in this fashion; and by observing a sample of uranium II at intervals over a period of almost twenty years, and measuring the radium which was being developed within it, Soddy was able through equation (7) to calculate the half-period of this intermediate substance (ionium).

The method used in deriving the equations (4) and (7) can always be

applied to any number of consecutive radioactive substances; there are always just equations enough to determine all the constants and describe completely the future history of any mixture of the members of a single family line, provided that their relative proportions in the mixture are specified for some particular moment. Even with only three substances the behavior of the mixture may be extraordinarily complicated; but there are simpler cases which are instructive.

If for instance one sets aside a substance with a much longer half-period than any of its posterity possesses, the extant quantity of each and every one of the descendants will first increase and then begin to decrease, and eventually diminish along the same exponential curve as the long-lived ancestor itself—not because the half-periods of the descendants are actually changed, but because of the partial balancing between the decay and the replenishment of each. Thus the half-period of the long-lived ancestor may be determined by plotting against time the total intensity of all its rays and all the rays of its descendants, or that of any particularly convenient kind of ray emitted by any member of the family. The most carefully measured and accurately known of all disintegration-constants, that of radon, is usually determined in this way; its half-period amounts to four days, those of its three next descendants radium A and radium B and radium C to only a few minutes each, so that after isolating a sample of radon and waiting a few hours one can set up any device for measuring the gamma-rays of radium C, plot their decay-curve, and from it determine a value of λ which is not that of radium C, but that of radon.

If in such a case as the foregoing the long-lived ancestor is so very long-lived that no appreciable decrease in its rate of transmutation can be detected over a period of years, then eventually the quantities and the radiations of all of its descendants assume values which likewise do not change appreciably for years; "radioactive equilibrium" is attained. In a unit of time, equal numbers of atoms are transmuted out of each substance into the substance following, into each substance out of the one preceding. Representing by N_n the number of atoms of the nth member of the series (counting the very long-lived ancestor as the first) extant in the mixture in radioactive equilibrium, by λ_n its disintegration-constant, and remembering that $\lambda_n N_n$ is the rate at which its atoms perish by transmutation, we have the chain of equations:

$$-dN_1/dt = \lambda_1 N_1 = \lambda_2 N_2 = \lambda_3 N_3 = \lambda_4 N_4 = \ldots \qquad (8)$$

from which, if we know the relative quantities of any two members in

a mixture in equilibrium, and the half-period of either, we can determine the half-period of the other.[18]

This method could be applied to estimate the half-period of radium, which is so long that in the years since it was first isolated no sample has yet become perceptibly feebler in emitting its rays, while the half-periods of its descendants are all much shorter, and that of its child is only 3.82 days and is rather accurately known. However, the volume of radon gas in equilibrium with one gramme of radium (about the largest quantity of radium which has ever been gathered together in one place) is at normal temperature and pressure only about .0006 cc, and the measurement of so small a quantity of gas is inevitably so inexact that this method cannot compete even with the not-very-accurate alternative methods which we shall presently meet. However, its results do not disagree with theirs.

The most fascinating application of this method is made upon the rocks of the earth, which have presumably been existing so long that there has been ample time for the longest-lived member of the uranium-radium series to attain equilibrium with all of its descendants. As it happens, the longest-lived member of this series is the first, uranium I. Probably this is no mere accident; if uranium is the descendant of less lasting ancestors, they would all be gone by now. However that may be, it is a fair presumption that at least the older rocks of the earth have been formed and buried long enough for the uranium in them to have attained to equilibrium with its descendants. The ratio of the concentration of uranium to the concentration of any member of its posterity, radium for example, should then be equal to the reciprocal of the ratio of their half-periods. Great numbers of samples of rock from all over the world were analyzed by Rutherford and his pupils, and in the laboratories of France and Germany; and for a large proportion among them the ratios of the radium content to the uranium content were found to lie close to one another, and to a mean value which Rutherford assigns as $3.40 \cdot 10^{-7}$. Accepting this as the equilibrium-ratio, and 1690 years as the half-period of radium, we obtain for the half-period of uranium the truly colossal figure of 4.4 billions of years! This value is substantiated, as we shall presently see, by an altogether different method.[19]

[18] In some of the older rocks of the earth, uranium and its descendants have attained mutual equilibrium, and the value of λN for uranium in such a rock is equal to the rate at which the inert end-product (RaG) of the series is accumulating, so that by measuring the amount of RaG already accumulated and the amount of uranium still remaining one can estimate the age of the rock. Consult O. Hahn, *Handbuch der Physik*, **22**, pp. 289–306.

[19] This is a fortunate circumstance, as it gives greater confidence in rejecting the data obtained with samples of rock which yield values of the radium-to-uranium ratio differing considerably from $3.4 \cdot 10^{-7}$. In some cases these deviations may be

One of the two best methods for determining the half-period of radium is a combination of this last-named method with one of those which I described earlier. Let us suppose that a sample of ionium, equal to the amount which would be in equilibrium with one gramme of uranium and $3.40 \cdot 10^{-7}$ grammes of radium, is purified of its original radium-content and set aside for occasional observations of the rate of growth of fresh radium in it. Representing by N_1 the number of ionium atoms in the sample (which diminishes in so small a proportion that we may consider it constant), by N_2 the number of radium atoms extant at time t after the new supply begins to grow, by λ_1 and λ_2 the disintegration-constants of these two substances; translating equation (4) into this notation, and remembering that the rate of transmutation of the parent substance which was there called B is now (measured in atoms transmuted per second) equal to $\lambda_1 N_1$, we have

$$N_2 = \frac{\lambda_1 N_1}{\lambda_2}(1 - e^{-\lambda_2 t}). \tag{9}$$

Represent by N_{20} the number of atoms of radium which would be in radioactive equilibrium with the sample of ionium, that is to say, the number of atoms in $3.40 \cdot 10^{-7}$ grammes of radium; by equation (8) we have

$$\lambda_1 N_1 = \lambda_2 N_{20}, \tag{10}$$

so that equation (9) may be transformed into one containing no constants except the known one N_{20} and the object λ_2 of the investigation. The gain is still greater; developing the exponential function in (9) as a power-series in t and retaining only the first term, we have

$$N_2 = N_{20}(1 - e^{-\lambda_2 t}) = N_{20}\lambda_2 t + \text{terms of higher order.} \tag{11}$$

This means that we need to trace the growth-curve of radium out of ionium only so far as is necessary to determine its initial slope, the initial rate at which the radium increases before its own transmutation begins to tell. This as it happens is all that there has yet been time to trace, so that this combination of the two methods is the only way yet available of interpreting the growth-curves.[20] After a sample of ionium has been kept for a century or two, it may be possible to trace a long enough arc of the curve to determine by the first method. After

ascribed to the comparative youth of the rocks, in others to the selective action of flowing water and other geological agents in removing some and leaving others of the members of the radioactive family.

[20] This method, it will be perceived, is essentially a measurement of one and hence of all of the terms $\lambda_n N_n$ which are equated in equation (9); the rate of growth of radium out of ionium being ascertained, it is possible to calculate the value of λ_n for any substance in the radium series for which N_n, the quantity in equilibrium with the preassigned quantity of ionium, can be measured.

our descendants have solved the other problems of physics, they may be able to entertain themselves by keeping records of the behavior of long-lived radioactive substances, and so determining half-periods with an accuracy improving from millennium to millennium.

Another and the most picturesque of all the ways of determining a disintegration-constant consists in counting the atoms which in a measured quantity of the substance disintegrate in each second. It sounds almost unbelievable that this should be feasible; but it is really practicable to count the alpha-particles which proceed from a radioactive substance, for they make individual visible scintillations upon a fluorescent screen placed across their paths. If this device is inconvenient, one can measure the total charge which the alpha-particles carry into a chamber arranged to receive them, and divide it by the specific charge borne by each, which is very accurately known. The particles and consequently the transmuted atoms having been counted, it is necessary to weigh the substance which is emitting them; and this requirement is less easy to fulfil, being fulfillable in fact only for three substances—radium, and the long-lived ancestors thorium and uranium. Dividing the mass of the weighed sample by the mass of an atom, and dividing the quotient into the number of alpha-particles emitted per second, we obtain the value of λ_1. This of course does not prove that the transmutation is actually proceeding according to the exponential law; that is proved only for certain substances of which the half-periods amount to a few months, days or hours. Nevertheless we assume it, and multiply the reciprocal of λ_1 so measured by $log_e 2$, and call the product the half-period. The values thus obtained are close to 1700 years for radium, agreeing well with the results of the method just above described; 4.7 billions of years for uranium I, agreeing with the result derived from the relative proportions of uranium and radium in the rocks; and 22 billions of years for thorium.

There are yet other ways of estimating half-periods, some of them very ingenious. Extremely short-lived substances require special methods. Thoron, a gas with the half-period of fifty-four seconds, is blown with a measured velocity through a tube along which various electrodes are placed for measuring its activity as it flows past them. Actinium A, of which the half-period is only .002 second, is projected upon the rim of a rapidly revolving wheel, and whirled past various instruments which measure its activity at successive points of its transit through space and time. The projection is due to a very simple but none the less striking natural phenomenon; when an alpha-particle is fired out of an atom of its parent-substance actinon, the residual particle—the atom of actinium A—rebounds or recoils like the

gun which fires a shell. The speed of this "recoil atom" is calculable, standing as it does to the speed of the ejected particle in the inverse ratio of their masses; and it has been utilized for measuring an excessively short half-period, that of RaC′, which amounts to only 10^{-6} second; a tube was oriented so that some of the recoil atoms flew along it, and their activity at various points of their flight was measured as in the case of thoron.[21]

Many of the half-periods, finally, are determined by analyzing the curves which represent the variation in time of the rays from continually-changing mixtures of growing and decaying substances: curves which presumably can be represented as sums of three, four or even more terms like the exponential terms in equation (6), not however independently known—that is to say, their coefficients and their exponents must be determined by inspecting the activity-curve itself and trying to build one like it. This operation sometimes requires a great deal of skill and discernment and intuition. It seems little short of marvelous that all the radioactive substances of the known series should have been recognized and their half-periods measured. That they have all been recognized there can be little doubt; for let us consider what it would imply if another substance lay undetected between (let us say) radon and radium A. There would have to be not one such substance but three, one of them emitting alpha-rays and the two others beta-rays—for otherwise the displacement-law of Fajans and Soddy would be broken. But if there were an undetected alpha-ray-emitting substance between radon and radium A, the atomic weight of the latter would be eight units below that of the former, instead of only four as we now suppose; and this difference of four units would follow step by step all the way down the radium series, ending in a to-be-expected value of 202 instead of 206; which would vitiate the excellent agreement between the latter figure and the observed atomic weight of the samples of element 82 contained in the uranium ores. The same argument can be used in the thorium series; in the actinium family the basis for the argument is lacking, but the parallelism between this and the other two families conduces to the same belief. It is all but certain, therefore, that the explorers of radioactivity have done their work so thoroughly that no substance yet remains unknown in the direct genealogical line from uranium I to radium G, nor in that from thorium to thorium D, nor between radioactinium and actinium D.

[21] This experiment was first performed by J. C. Jacobsen, and later by A. W. Barton, whose paper (*Phil. Mag.* (7) **2**, pp. 1275–1282; 1926) should be consulted for details. It is a very delicate one, especially as the atoms recoil because they have emitted not alpha-particles but electrons, which are comparatively light and are emitted with various speeds.

We turn to the rays themselves.

The alpha-rays are particles of mass $6.60 \cdot 10^{-25}$ gramme and positive charge $2e$ or $9.55 \cdot 10^{-10}$ electrostatic unit. The particles emitted from different radioactive substances differ, so far as we know, only in their initial speeds. The range of variation is astonishingly small; the slowest known alpha-particles issue from their sources (atoms of uranium I) with a speed of $1.423 \cdot 10^{9}$ cm/sec, the fastest [22] emerge from atoms of thorium C′ with a speed of $2.069 \cdot 10^{9}$ cm/sec. The differences in speed between different alpha-particles emerging from a substance are imperceptibly small.

As a rule the speed of the alpha-rays from a substance is neither measured nor quoted directly; one measures by preference their *range in air*, a thing which can be defined because alpha-particles ionize air (and other substances) more readily the more slowly they are moving, until their speeds drop below 10^{8} cm/sec and they suddenly cease to ionize altogether. Consequently, if alpha-rays shoot out from a bit of radioactive substance into environing matter, the concentration of the ions which they produce increases steadily and rapidly from the emitting substance outwards, up to a distance where it attains a sharp maximum and then suddenly falls to zero.[23] This distance is the range in the material in question; it is greater the faster the alpha-rays, varying as the cube of their initial speed. It is a property of the alpha-rays and not of the substance which emits them, and I should not have introduced it here but for a certain relation between ranges and half-periods, and as a pretext for showing some pictures of pleochroic haloes.

These haloes occur in certain ancient minerals, chiefly mica; they are systems of concentric spheres of discoloration, of which the pictures represent cross-sections. No one could imagine what they were when they were first discovered; but the explanation is simple and beautiful. Particles of uranium in some cases, of thorium in others, bubbles of radon in yet others, were caught ages ago and held in the points which were to become the centres of the haloes; the spheres of discoloration are the regions of maximum intensity of ionization, where the alpha-rays emitted from the central source were slowed down to their speed of optimum ionizing-power and were on the verge of

[22] Among the particles issuing from samples of thorium C and producing scintillations on fluorescent screens, very occasional ones (one in ten thousand, or fewer) have a much greater range than the rest, or than the characteristic particles of other substances. A few corpuscles of abnormally long range issue from samples of radium C. It is a controversial question whether these particles come from nuclei disintegrating in a rare and abnormal manner, or from nuclei struck and broken by alpha-particles ejected from other atoms, as sometimes happens. Even the published data are not all in accord, and it is unsafe to make further statements.

[23] *Introduction*, pp. 200–204.

ceasing from ionization altogether. The radius of the outer boundary of every such sphere is the range, in mica, of the kind of alpha-rays which caused it. All of the alpha-ray-emitting descendants of the initially-imprisoned substance form their individual spheres; in the cross-sections of the best haloes one can discern nearly all of the rings due to uranium I and its seven alpha-ray-emitting descendants on the direct line to radium G, or those of the seven members of the thorium

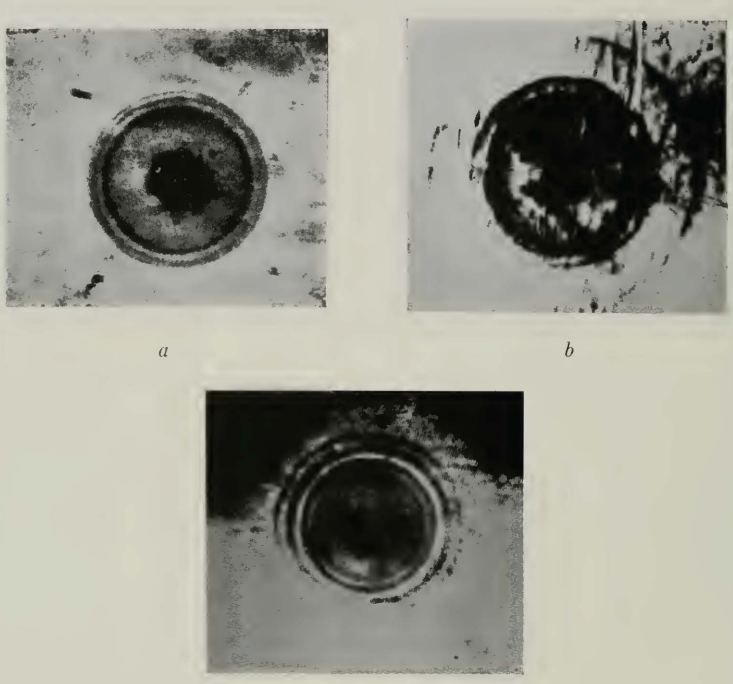


Fig. 3. Pleochroic haloes (B. Gudden, *ZS. f. Physik*)

a. Rings of UI and UII (innermost, merged into a single broad ring), Io, and Ra.
b. Rings of RaF (innermost), Rn, RaA and RaC′.
c. Rings of various substances of the uranium-radium series. Magnifications 665, 500, 480 respectively.

series which disintegrate in this way. There are no extra rings in these haloes, which strengthens the presumption that no radioactive substances in either series lie undetected. But there are also haloes of which the rings have not the proper radii to be identified with any known radiating substance. Are these possibly evidence for the prehistoric existence of others belonging to other series, all of which were too short-lived to survive into the days of scientific research, but disappeared with the dinosaur and the pterodactyl?

There is an interesting and important relation between the initial speeds of alpha-rays and the half-periods of the substances which emit them. One varies as an exceedingly high (negative) power of the other, so that when speed is plotted versus half-period upon logarithmic plotting-paper the resulting curve is a straight line; or, rather, three parallel straight lines, one for each of the three series. This remains true if we plot any power of the speed (for instance, the third) against

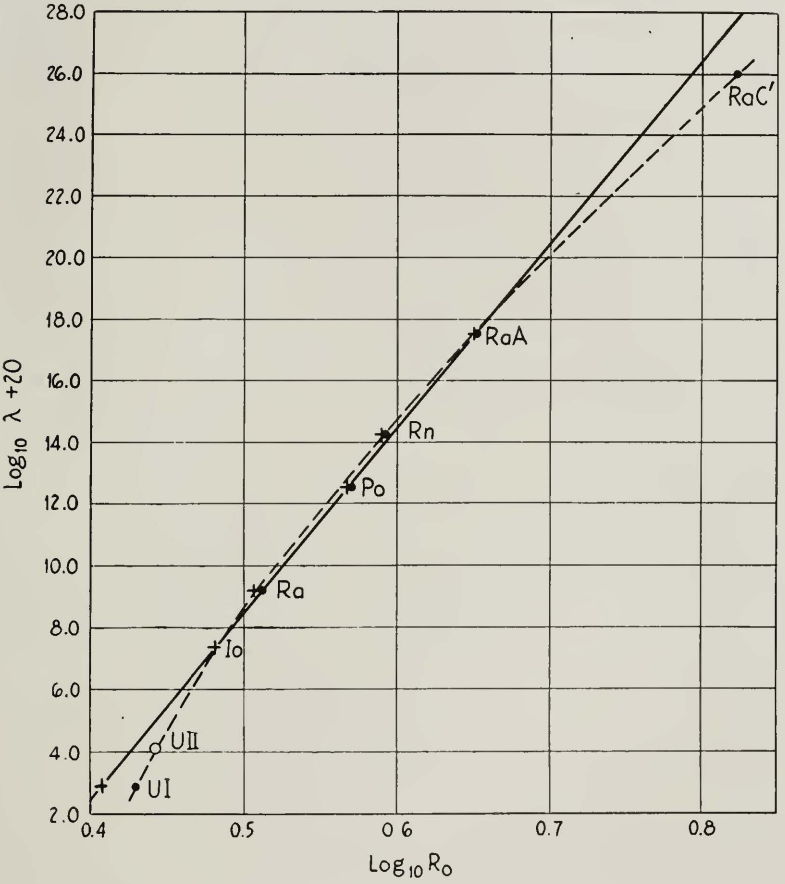


Fig. 4. The Geiger-Nuttall relation

Data for the uranium-radium series; the values for alpha-particle range denoted by dots are taken from pleochroic haloes, those marked by crosses from experimental data. The value of half-period for UII is not known, but is placed by interpolation upon the smooth curve. The straight line represents the best attainable approximation by a linear relation between logarithms of half-period and range; the smooth curve is that drawn by B. Gudden, from whom the data are taken.

any power of the half-period (for instance, the power -1), so that the logarithm of the range of the emitted particles varies linearly with the

logarithm of the disintegration-constant of the emitting substance. This is the way in which this, the "Geiger-Nuttall" relation, is usually expressed:

$$\log \lambda = A + B \log R. \tag{12}$$

The constant B is given (by Hevesy and Paneth) as 53.9 for all three series of radioactive substances, which signifies that the disintegration-constant varies as the fifty-fourth power of the range of the ejected particles! I do not know of any other relation between physical variables in which so high a power occurs; radioactivity, like astronomy, is the home of colossal numbers. The constant A varies from one series to another; it is given as -37.7 for the radium series.

The Geiger-Nuttall relation, like most simple formulae, is a mere approximation. For the radium series its degree of accuracy is illustrated by Fig. 4; the curve drawn through the various points is not quite a straight line. In the actinium series there is a jolt; the point for actinium X lies quite away from the place it should occupy on the straight line drawn to fit closest to the points for the other members, and in fact the half-period of AcX is shorter than that of RdAc, though its alpha-particles are slower. A straight line can be drawn to pass near the points for the remaining members of this series, and another to pass near the points for the descendants of thorium, about as successfully as the line in Fig. 4 fits the points for the radium family. Extending the line drawn for the thorium family to the value of the range for the fastest of all alpha-particles,[24] those of thorium C′, one obtains by extrapolation for the half-period of this substance the fantastically small value 10^{-11} second. There is no discernible prospect of verifying this by direct measurement, and in quoting it one should remember the risks of extrapolation.

An alpha-particle is a helium nucleus; when it acquires two electrons, the combination is a helium atom. Helium therefore is a daughter-substance of every radioactive substance which transmutes itself by emitting alpha-rays.

Passing over from alpha-rays to beta-rays, we take at once a great step backward from the clear to the obscure.

The great trouble arises from the fact that beta-rays are electrons, and electrons exist both in the atom-nuclei and in the electron-systems which surround them, or at least they come out of both localities. Whereas the emergence of an alpha-particle from a substance is a clear sign of the transmutation of one atom of that substance, the emergence of a beta-particle need not mean anything of the sort; it may

[24] Reservation being made for the particles mentioned in footnote 22.

simply mean that an alpha-particle, or a gamma-ray quantum, or a different beta-particle coming out of one atom-nucleus operated on its way out the expulsion of that electron from the outer electron-family of that atom or some other. To take one instance only: radium C and radioactinium both emit beta-rays and alpha-rays together, but in the former case there is as we have seen a dual transmutation, in the latter the beta-rays appear to be electrons torn out of the electron-shells surrounding the atom-nuclei as the alpha-particles pass by on their way out. Electrons have the same charge and the same mass, whatever their origin; although it is essential to distinguish how they originate in all these cases of beta-ray-emitting substances, there is no way to make the distinction except by performing experiments on distribution-in-speed of the beta-rays and invoking various theories, not always of the highest order of reliability, to interpret the results. This is the reason why, as Meitner says, the beta-rays actually emitted from self-transmuting nuclei "are the least clarified point in the entire problem of the radioactive transformations."

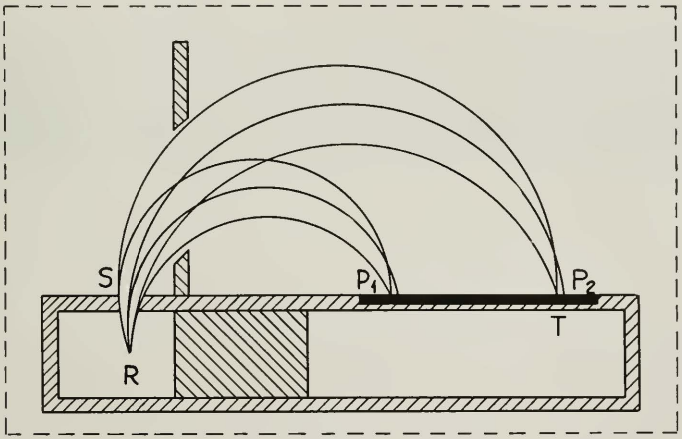


Fig. 5. Apparatus for photographing beta-ray spectra

(After C. D. Ellis and H. W. B. Skinner. Source at R, magnetic field and photographic plate perpendicular to plane of paper, which the plate intersects along P_1P_2.)

In attacking the beta-rays the first thing to do, and indeed the only thing which can be done by experiment, is to determine their distribution-in-speed—the function which gives the relative number of electrons issuing from the substance with speeds comprised between any preassigned limits. The process consists in isolating, by a proper system of narrow perforations and slits, a narrow beam or pencil of electrons, and applying to this pencil a magnetic field which bends the paths

of the electrons (see Fig. 5). The slower the electron, the more its trajectory is curved; if the beam comprises particles of more than a single speed, it is spread into a fan, and a photographic plate placed across the path of the fan records the "magnetic spectrum" of the electron-beam. If the beam comprises several groups of electrons, each with its own sharply marked and definite speed, each group falls upon a distinct part of the plate; if the slit limiting the beam is long and narrow, the groups form long and narrow discolored bands upon the plate, and these bands or "lines" constitute an electronic line-spectrum. The appearance of lines in a magnetic spectrum is taken as practically convincing evidence that the electrons in question issue from the circumnuclear electron-families of the atoms, not from their nuclei.

For this view there is direct evidence of a very convincing character: namely, that beta-ray spectra containing the same lines can be elicited from ordinary stable elements not undergoing transmutation, by the simple process of playing gamma-rays upon them from the radioactive substance in question. Take a sample of the substance, and envelop



Fig. 6. Part of the beta ray spectrum of radium B

(After Ellis and Skinner, *Proc. Roy. Soc.* Range 0.037 to 0.054 millions of equivalent volts.)

it in a metal sheath thick enough to stop all of the electrons or alpha-particles issuing from it. Some of the gamma-rays will pass through the sheath, for generally some of them (not necessarily all) are more penetrating than any other radiations which the substance can emit. Let these fall upon another piece of metal nearby; apply a magnetic field to the electrons expelled from this metal, or indeed to the electrons which the gamma-rays expel through the outer surface of the sheath enclosing their source; photograph the resulting spectrum. If the atomic number of the irradiated metal does not depart too far from that of the radioactive source—if for instance the irradiated metal is uranium or lead or platinum or tungsten—the spectrum of the electrons expelled from it will resemble the natural beta-ray spectrum of the source, closely enough so that strong lines of the one spectrum can obviously be identified with corresponding strong lines of the other. Corresponding lines in the two spectra may or may not coincide with

one another; that depends on the kind of metal irradiated; a prominent line in the spectrum of (for instance) radium B will be composed of electrons having energy somewhat greater than that of the electrons forming the corresponding line in the spectrum elicited from uranium, somewhat less than that for the corresponding line from platinum. But if the irradiated metal be isotopic with the substance into which the radioactive source is being transmuted, corresponding lines will be found to coincide with one another. One obtains a beta-ray spectrum having many lines in common with that of radium B, by allowing the gamma-rays to play upon and expel electrons from a piece of a metal isotopic with radium C—that is to say, bismuth.[25]

Whether one uses an atom-model or not, these facts suggest that some at least of the electrons emerging from a radioactive substance are hurled out by some sort of a secondary process operated upon the already-transmuted atoms by the accompanying gamma-rays, working in the same manner as they work upon atoms exposed to them outside. This suggestion becomes much more precise when the atom-model is invoked; for the contemporary model is designed to give a vivid explanation of the lines in the electronic spectra elicited by X-rays and gamma-rays playing upon the atoms of the stable metals.

Every such line is composed of electrons extracted from a particular group, in the circumnuclear electron-family of the atom, by radiation of a particular frequency. Think of the most tightly-bound electrons of all, the so-called K-electrons, to be imagined as lying or revolving closer than any of the others to the nucleus. Merely to extract one electron of this set, a definite amount of energy W_K must be imparted to the atom. Conceive a beam of radiation of frequency ν pouring over a multitude of similar atoms; to each it communicates either no energy at all, or else a definite amount of energy equal to $h\nu = 6.57 \cdot 10^{-27}\nu$. If this "quantum" unit of energy exceeds W_K, and if the radiation extracts a K-electron from an atom, the liberated electron will fly away with a kinetic energy equal to the excess of the imparted energy $h\nu$ over the extraction-energy or "binding-energy" W_K.

$$\text{Kinetic Energy} = T = h\nu - W_K. \tag{13}$$

This equation determines the initial speed of the departing K-electrons.[26]

[25] *Introduction*, pp. 184–192; to this I refer also for reproductions of some very beautiful photographs of beta-ray spectra taken by J. Danysz and M. de Broglie.

[26] If the speed v of the electrons is inferior to $3 \cdot 10^9$ cm/sec, it is permissible to set for T the familiar expression $\frac{1}{2}mv^2$, putting for m the "rest-mass" $m_0 = 9 \cdot 10^{-28}g$ of the electron. Otherwise it is necessary to take account of the dependence of the mass of the electron upon its speed, preferably by using the formula derived from the

If there is but one frequency in the inflowing radiation, the spectrum of the emitted electrons will contain one line composed of what were formerly K-electrons. It will contain others, composed of electrons which originally belonged to other and less firmly-bound sets within the atoms. We distinguish, in order of decreasing binding-energy, K and L_{I} and L_{II} and L_{III} and M_{I} and M_{II} and M_{III} and M_{IV} and M_{V} and still further classes of electrons. The electron-spectrum due to radiation of a single frequency attacking atoms of a single kind comprises a line for each of these classes (apart from those, if any, for which the binding-energy exceeds the quantum-energy $h\nu$ so that the radiation cannot detach them) and the speed of the electrons composing each line is determined by an equation like (13), with the appropriate extraction-energy W_{LI} or W_{LIII} or whichever it may be inserted in place of W_{K}. If there is more than one frequency in the incident radiation, each produces its own system of lines. These statements are proved, and the binding-energies are determined for all the classes of electrons and most of the kinds of metallic atoms, by irradiating metals with X-rays of which the frequencies are known, for they can be separately measured.[27] To ascertain the binding-energy of, let us say, the L_{II} electrons of platinum, one has only to look into the standard tables.

Now we have seen already that the physical and chemical properties of each radioactive substance, so far as they are known, are almost exactly like those of its stable isotope (if it has one); and with this rule the resemblance between the beta-ray spectrum of a radioactive substance and the electronic spectrum which its gamma-rays elicit from its stable isotope most admirably conforms. When a line in the former spectrum obviously corresponds to a line in the latter, both presumably are composed of electrons extracted from the same level by the same radiation. The same gamma-rays are working upon atoms isotopic with one another, and therefore endowed with electron-

Theory of Relativity, to wit:

$$T = m_0c^2\left[\frac{1}{\sqrt{1 - v^2/c^2}} - 1\right] = h\nu - W_K$$

X-rays generated by artificial means never have frequencies so high that the electrons which they expel move rapidly enough for the simple substitution $T = \frac{1}{2}mv^2$ to be inadequate; but the frequencies of some of the gamma-rays are so great that the electrons which they extract even from the K-layers of massive atoms depart with speeds much exceeding $3 \cdot 10^9$ cm/sec. J. Thibaud has made direct measurements of a certain gamma-ray frequency and of the speed of the electrons which it ejects from a certain group of known extraction-energy, which are compatible with one another and with equation (13) if the relativity-formula for T is used, but decidedly incompatible if T be set equal to $\frac{1}{2}m_0v^2$ or to the once well-known expression derived by Abraham (J. Thibaud, *L'effet photoélectrique composé;* Paris (Masson) 1926).

[27] *Introduction*, pp. 192–195, 273–282.

families classified into identical classes with identical binding-energies. There is an evident difference; the atoms in the latter case are ionized by radiation poured upon them from without, in the latter by processes which occur within their own nuclei. (Whether in the latter case a wave-train does actually leave a nucleus, and enjoy a real existence during the brief time before it reaches the circumnuclear electron which it is destined to eject, is a question to which it is not easy to give a

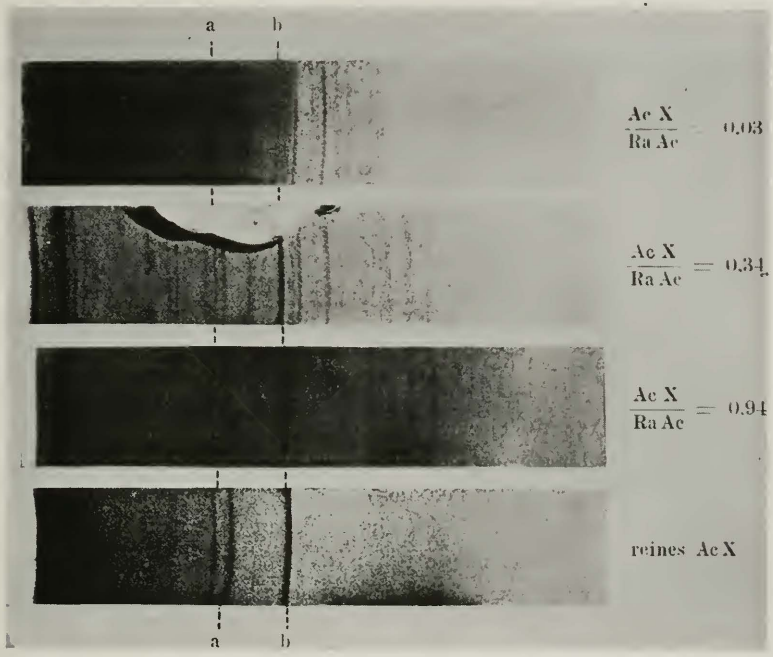


Fig. 7. Beta ray spectra of radioactinium and actinium X

(After O. Hahn and L. Meitner, *ZS. f. Physik.* The three upper pictures represent portions of the beta-ray spectrum photographed respectively a few hours, 6 days, and 20 days after the preparation of a pure sample of RdAc, in which the daughter-substance AcX was steadily growing; the lowest, the corresponding portion of the spectrum of a sample of AcX with its descendants. The lines which diminish in intensity from top to bottom are those of RdAc, those which increase belong to AcX and its descendants (note especially the lines marked *a* and *b*).)

sensible answer!) But the difference does not affect the energies of the ejected electrons; only their number, for, as seems natural enough, the beta-rays expelled from atoms of which the nuclei are emitting gamma-rays are much more abundant than those which an equal amount of gamma-radiation extracts from atoms on which it falls from without. Corresponding electron-groups have the same energy.

This is why we know in some cases, and suppose in the others, that when the electrons issuing from a radioactive substance constitute the lines of a line-spectrum they are not themselves coming from the nuclei; they are merely the signs of gamma-rays coming from the nuclei.

This discovery disposes of one potential objection to the displacement-law of Fajans and Soddy. Radioactinium (for instance) is a substance which emits alpha-rays and passes over into a substance two steps farther down the procession of the elements, as the displacement-law requires; but it also emits beta-rays, and since no alternative product one step farther up the procession has been discovered, the displacement-law would be gravely threatened if it were necessary to suppose that these come from the nuclei. There is no such necessity; and since the beta-rays display a line-spectrum, it is intrinsically all the more likely that they come from the circumnuclear electron-shells.

En revanche the character of the thus-far-analyzed beta-ray line-spectra makes it all the more difficult to understand what becomes of the electrons which must truly be emitted from the nuclei, in the transmutations in which the daughter-substance is displaced one step up the procession from its parent. When a substance is undergoing a transmutation of the other kind, the alpha-particles which its atoms emit all have very nearly the same speed. One would certainly expect that the electrons emitted from the nuclei of all atoms of radium B at their instants of transmutation emerge with the same speed. If so, they should compose a sharp line in the beta-ray spectrum of radium B. Now there are certainly some lines in this particularly rich spectrum which have not yet been definitely and exactly explained by the theory which I described before; but it appears that none of them is very prominent, and most of the experts refuse to admit that any one of them is composed of electrons coming forth direct and unretarded from the nucleus. There are other substances which display beta-ray spectra comprising but a few lines, one of which some physicists believe to contain the nuclear electrons.

If the nuclear electrons are not to be assigned to the lines, there remains but one alternative; they must be identified with the electrons composing the continuous beta-ray spectrum which underlies the lines and intervenes between them. The best way to study this spectrum is to dispense with the photographic plate, and set a Faraday-chamber to receive the electrons, with its aperture somewhere in the plane which the plate formerly occupied; if then the magnetic field is continuously varied, the spectrum slides across the aperture, and at each particular value of the fieldstrength the electrons of a particular limited speed-range pass into the chamber and are counted (more precisely, the total charge which they bear is measured, which comes to the same thing).

Curves obtained in this way are copied in Fig. 8. The peaks are the traces of lines (not so many as a photograph would show, for the method in this respect is not so delicate) rising up not from the zero-level but from a smooth sweeping curve, carried (hypothetically) in

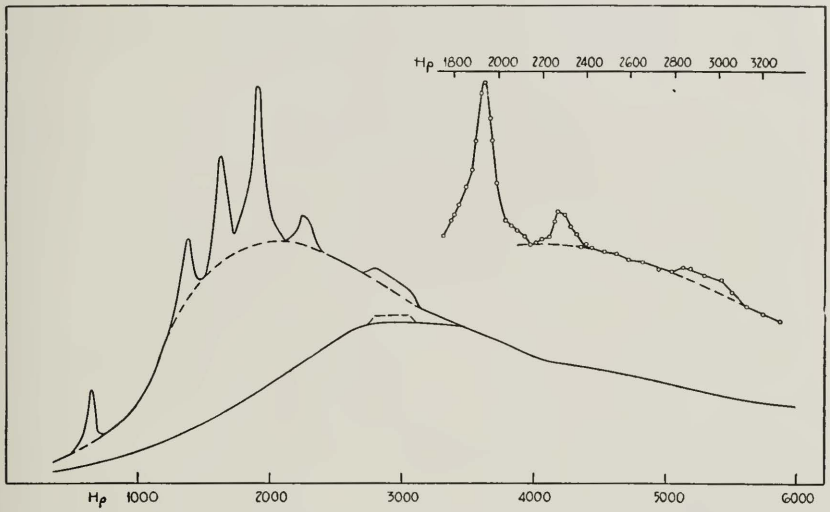


Fig. 8. Beta-ray spectra measured with a Faraday chamber
(Lower curve for RaC, upper curves for RaB + RaC. After R. W. Gurney.)

dashes across the base of the peaks. This is the distribution-curve of the electrons forming the continuous spectrum; integrating it, one obtains the total number of these electrons.[28] This number has been measured for radium B and radium C by Gurney; it amounts to somewhat more than one electron per self-transmuting atom.[29] A much smaller number, had such a one been found, would have rendered untenable the notion that all the nuclear electrons go into the continuous spectrum; the result proves that there is no such obstacle, not at least in these cases. The beta-ray spectrum of radium E consists of a single diffuse band; there are no lines. Emeléus counted the emitted electrons and found a value equivalent to 1.1 electrons per self-transmuting atom.[30] Perhaps then it is a quality of the nuclear

[28] This statement is not exact. The curves may be transformed into true distribution-curves, resembling them but not identically like them, by processes involving allowances for the geometry of the apparatus. The area under these true curves must then be found by integration, and gives the total charge borne by the electrons, the quotient of which by e is the desired number of electrons.

[29] R. W. Gurney, *Proc. Roy. Soc.*, **A 109**, pp. 540–561 (1925).

[30] K. G. Emeléus, *Proc. Camb. Phil. Soc.*, **22**, pp. 400–404 (1924). It is frequently pointed out that RaE emits no perceptible gamma-rays, a fact which makes it seem additionally probable that all the electrons which it emits come from the nuclei. This does not prove anything, as it is conceivable that gamma-rays are emitted which extract electrons from the electron-layers with such efficiency that no appreciable fraction of them escapes unconverted.

electrons that they issue from their atoms with widely and irregularly scattered speeds. If this is true, the presumption is that they escape from the nuclei with equal speeds, the differences resulting from experiences of theirs during the transit through the circumnuclear electron-shells. But it must not be forgotten that a continuous electronic spectrum appears together with the lines, when the gamma-rays from a radioactive substance fall upon one of its stable isotopes; and some allowance must certainly be made for this.

Refreshingly in contrast with the status of this perplexing question is the condition of another, for years the subject of a fervid controversy. Are the gamma-rays from a self-transmuting atom emitted before or after the transmutation occurs? There is only one way of settling this question, and perhaps the question itself ought to be so phrased as to bring this way into prominence. Granting the theory of beta-ray line-spectra expounded in these pages, and granting that certain lines in a certain spectrum have been recognized as being composed of electrons expelled by gamma-rays of one and the same frequency from various K, L, M classes in the circumnuclear electron-family, do the energy-values of these lines show that the electrons come from atoms as yet untransmuted, or from atoms which have already undergone their transmutation—from the atoms of the parent, or those of the daughter-substance? There is no forceful *a priori* reason for expecting either of these alternatives rather than the other; the question must be put to experiment.

If one knew with all desirable accuracy the frequency of the gamma-ray responsible for a particular set of lines, and the class of electrons contributing each line—if one knew for instance that a certain line is composed of K-electrons extracted by gamma-rays of a known frequency ν, one would measure the speed of these electrons, calculate their kinetic energy, subtract it from $h\nu$, identify the difference with the binding-energy W_K according to equation (11), and consult the standard tables to locate the element possessing that value of the extraction-energy for its K-electrons. But there are few gamma-rays of which the frequencies are independently known, and for these the values are not very accurate; so that this method is not generally available.

If however two lines are composed, the one of K-electrons and the other of L_1 electrons ejected by gamma-rays of the same though unknown frequency, then the difference between the values of kinetic energy for the electrons of the two lines is equal to the difference between W_K and W_{L1}; and as this difference varies from element to element, one can consult the tables to locate the element for which the

difference between the K and the L_I extraction-energies agrees with the measured value. This is the usual method.

There is still another way, which may be explained by describing an experiment performed by C. D. Ellis and W. D. Wooster.[31] They enclosed a sample of radium B mixed with radium C in a rather thick-walled platinum tube, and deposited a thin layer of the same mixture upon the outer surface of the tube. The thin layer contributed the beta-ray spectrum of radium B and radium C. The beta-rays from the substances inside the tube were stopped by its walls, but the gamma-rays went through and expelled electrons from the platinum, which mingled with those from the covering film; so that upon the photographic plate there appeared side by side the spectrum-lines composed of electrons extracted from atoms of radium B and radium C by their own gamma-rays, and the spectrum-lines composed of electrons extracted from atoms of platinum by gamma-rays of the identical frequencies. Side by side there appeared, for instance, the lines due to K-electrons extracted by the same radiation from radium B and from platinum. The electrons from the radioactive substance had less energy than those from the platinum, for more had been spent in extracting them; the difference between the values of kinetic energy of the electrons was equal to the difference between the values of the K binding-energy for the atoms, with sign reversed; the K binding-energy for platinum is known, that of the other atom is calculated at once.[32]

The six or eight investigations, performed by these methods upon diverse substances by various physicists during the past two years, have all come to concordant results. The atoms from which the electrons of the beta-ray line-spectra are detached are the atoms of the daughter-substances; the gamma-rays are emitted, or at least they act (and it would be a daring person who would say that they exist for a while before they act!) after the transmutation occurs. The controversy

[31] *Proc. Camb. Phil. Soc.*, **23**, pp. 844–848 (1925). There are several important articles in this (November, 1925) number of the *Proceedings* which deal with the problem of the emission of gamma-rays, secondary X-rays, and electrons emitted from the nucleus or ejected from the circumnuclear family by these rays.

[32] All the methods require the observer to guess which lines are composed of electrons from the K-class, which of electrons from the L_I class, and so forth; and this is the major difficulty of the problem, for there is nothing intrinsically distinctive about the lines. In many cases, especially when there are several gamma-ray frequencies and a multitude of beta-ray lines, it is necessary to proceed by trial and error, assigning a line first to one class of electrons and then to another, and finally adopting the systematization which leaves the smallest number of lines unexplained or at odds. Sometimes only one out of the three L classes yields a perceptible number of electrons; there is a rule, which if general is very valuable, that when the product of h into the frequency of the gamma-ray exceeds the extraction-energies of all the L classes very greatly, then the L_I class is the only one out of which electrons enough are extracted to make a noticeable line.

is settled, and the triumphant side is that of which Meitner was the protagonist. Evidently we must conceive that the electron departing from a nucleus leaves it in a very unstable state, from which it speedily passes over into a comparatively though not absolutely stable state by one or a series of transitions, of which the gamma-rays are the manifestations.

We have still the gamma-rays to consider. Throughout this article I have taken it for granted that the gamma-rays are electromagnetic waves of definite frequencies. The evidence that they are electromagnetic waves has been known so long that it need not be rehearsed. The classical way of determining the frequency of such a wave is to measure its wave-length. With ordinary light-waves this is effected by dispersing them with a prism or diffracting them with a ruled grating. It used to be thought that these methods do not avail with X-rays, because of the shortness of their waves; natural crystal gratings, in which close-ordered files of atoms play the rôle of the rulings in artificial gratings, are used to diffract these. Applying the crystals to gamma-rays, one meets the same difficulty as the discoverers of X-ray met when they applied prisms and ruled gratings; the waves are mostly too short to be diffracted appreciably by natural crystals. The gamma-rays are spread out into a spectrum, and sometimes lines are discernible in the spectrum (Fig.9); but the line of shortest wave-length thus far measured (so far as I know) is at 0.052 Angstrom units or 52 X-units, and there are certainly many others at much shorter wave-lengths which the crystal spectroscope does not diffract far enough outward to be located. Recently people have renewed the attempt to measure wave-lengths of X-rays by the methods appropriate to visible light, and have attained values of astonishing accuracy; perhaps it is not too much to hope that a comparable advance in technique will bring the shortest gamma-rays into the scope of crystal gratings.



Fig. 9. Gamma-ray spectrum of radiothorium and thorium B

(Wavelengths of the lines 52X, 145X, 168X from right to left. After J. Thibaud, *l. c.* footnote 26.)

The usual method for estimating the frequencies of gamma-rays consists in guessing the class to which the electrons forming a beta-ray line originally belonged; taking its binding-energy from the tables; measuring the kinetic energy of the electrons forming the line; adding

it to the binding-energy, and dividing the quotient by h. This is in a sense the reverse of the usual process of ascertaining which is the element from which the beta-ray line-spectrum proceeds; and as a matter of fact the two have frequently been carried out as parts of one single investigation. The line-spectrum is photographed and its lines are measured, and then the student works over the data until he succeeds in setting up a hypothetical gamma-ray spectrum in which not too small a fraction of the beta-ray lines are explained by the action of not too great a number of gamma-ray frequencies upon the K and L and M classes of electrons, and reversely there is no too obtrusive case of a beta-ray line being predicted from his hypothetical gamma-spectrum and failing to appear.

For illustration I will quote some actual results. Meitner and Hahn located forty-nine lines in the beta-ray spectrum of radioactinium; among these, thirty-seven could be attributed to the action of one or another of twelve gamma-ray frequencies upon one or another of nine classes of electrons. In the spectrum of actinium X they observed twenty-nine lines and explained fourteen of them by postulating seven frequencies. Happily there are much more perspicuous cases. The spectrum of radium D consists of only a few lines—five, according to L. F. Curtiss,[33] whose measurements show that four may be supposed to consist of electrons ejected from the L_{I}, L_{II}, M_{I}, and N_{I} shells by a single gamma-radiation of wave-length 0.26A, while the energy of the electrons forming the fifth line is not perceptibly different from the quantum-energy $h\nu$ of the rays themselves. These last electrons may have been extracted from the outer layers of the atoms, where the binding-energy is so small that it makes but an insignificant deduction from the energy transferred to the electron. Another instance is that of radium itself, of which the three lines composing the beta-ray spectrum may be ascribed to a single gamma-ray of wave-length 0.066A expelling electrons from the K group, the L_{I} and the M_{I} group. Such cases as these are so simple that the theory in general and the wave-lengths calculated for the gamma-rays in particular are almost beyond all question.

Certain of the gamma-ray frequencies thus determined, and some which are directly measured with the crystal spectroscope, are found to agree with characteristic X-ray frequencies of the atoms whence they come. This is true of the solitary gamma-ray which is necessary and sufficient to explain the beta-ray spectrum of UX_1, and of two of the rays postulated by Meitner to account for the spectrum of RdAc

[33] L. F. Curtiss: *Phys. Rev.* (2), **27**, pp. 257–265 (1926). A previous investigation by L. Meitner (*ZS. f. Physik*, **11**, pp. 35–54; 1922) had led to substantially the same conclusion regarding the gamma-ray spectrum.

and two of those for AcX. This is precisely what was to be expected; for in the contemporary atom-model, the characteristic X-rays of an atom are conceived to arise from its circumnuclear electron-family, and to arise after and because an electron has been evicted from the family —a cause which the primary gamma-rays, or the alpha-rays or the electrons coming out of the nuclei, can themselves supply. The electrons expelled by these "secondary" gamma-rays or X-rays (the latter term is now preferred, in all cases where the identification can be surely made) are ejected as the *fourth* stage of a complicated process: first, the primary quantum or particle departs from the nucleus, then a tightly-bound electron is ejected from the electron-family, then a rearrangement of the remaining electrons brings about the emission of an X-ray, which in turn expels the loosely-bound electron. (It seems unlikely, as I intimated before, that the four stages are really separate; probably the passage from the initial state to the final takes place in a single operation, in a flash; but one does not see how to conceive that single operation without resolving it into four.) Since even the primary gamma-rays are emitted after the transmutation, the secondary X-rays *a fortiori* must come from the atoms of the daughter-substance; and this they do.[34]

The gamma-ray spectra thus far mapped out consist of from one to fourteen frequencies, not counting the secondary X-rays; the palm, in this respect, is awarded to radium C. The highest frequency thus far recorded is $5.4 \cdot 10^{20}$, corresponding to a wave-length of 5.57X (5.57×10^{-11} cm) and a quantum-energy amounting to $3.54 \cdot 10^{-6}$ erg or 2.22 millions of equivalent volts; it has twenty times the frequency of the highest X-ray known, and twenty times as great an energy in each quantum as is required to tear the most tightly-bound electron from the family of the most massive atom. It emerges from the nuclei of atoms which have just transmuted themselves out of radium C into radium C′. The fastest electrons forming a definitely-known line in a beta-ray line-spectrum occur in that of thorium C″; their speed amounts to 0.986 of that of light, their energy to almost $2.5 \cdot 10^6$ equivalent volts; but there are still faster ones in the continuous spectrum of radium C, which extends at least as far as to 0.998 of the speed of light. The energy of the alpha-particles of the various substances which emit them

[34] The strongest single piece of evidence is the measurement upon two radiations of radium B, performed with the crystal spectroscope by Rutherford and Wooster (*Proc. Camb. Phil. Soc.*, **23**, pp. 834–837; 1925) who found that the difference between the angles at which they were diffracted from the crystals agreed closely with that to be expected for two prominent X-ray lines of the *L* series of the daughter element (atomic number 83) and disagreed unmistakably with that to be expected for the parent element. This invalidated a contrary result obtained in 1914, which long had stood as an obstacle in the way of the conclusion that gamma-rays are emitted after the transmutation.

ranges from somewhat over four to somewhat under nine millions of equivalent volts. The greatest amount of energy which men have yet succeeded in loading upon a single charged particle or crowding into a single quantum of radiation lies well below the first million of equivalent volts; it still lay well below the first hundred thousand, ten years after radium was discovered. The step from the tens of thousands to the millions is a great one; this supplement voluntarily offered by Nature, transcending immensely the greatest amounts of energy which men can concentrate into a compact parcel, is chiefly responsible for the advances in the understanding of energy and matter which radioactivity made possible.

The advances have indeed been great. Consider what ensued from the discovery of the alpha-rays alone. With alpha-particles Rutherford explored the interiors of atoms, and the results of his explorations led him to the nuclear atom-model. The particles themselves he proved to be atom-nuclei of a certain element, and they established the amounts of electric charge which must be assigned to the atoms of that element and all the others. The nuclear atom-model in turn supplied Niels Bohr with the substructure of his theory; and Bohr's theory, together with the phenomena which it inspired men to seek and find, forms the half of contemporary physics. In the edifice of modern physical theory, the alpha-particle is the cornerstone. Had Nature not dispersed the radioactive substances through the rocks of the earth, had there not been one or two of them long-lived enough to survive and maintain a supply of their descendants until man arrived and became scientific—or if the faint outward signs of the radioactivity latent in the rocks had been overlooked, or having once been noticed had been left unstudied—in any of these cases, centuries more might have passed before a proper foundation was located for the edifice. That is the prime reason for honoring those who detected radioactivity, and then did not rest until they had brought it fully into the light. Theirs is an illustrious history, and one not without pathos; for some of those who had worked with the greatest zeal found themselves in later years the prey of a terrible and inexorable disease; like Prometheus in the myth, they were consumed for having brought benefits to the human race. Even yet the benefits which they gave have not been fully exploited; marvelous things may still be discovered, in the process of understanding the actions of the rays on living matter. But that will be another story, and a long one.

Dynamical Study of the Vowel Sounds
Part II

By IRVING B. CRANDALL

SYNOPSIS: Comparative studies based on oscillographic records of the principal characteristics of vowel, semi-vowel, and consonant sounds, have contributed much to an understanding of the mechanism of speech. Analyses of the frequency spectra of vowels show almost invariably two principal resonance peaks which fact is suggestive of a double resonator to produce them.

The present paper is concerned with the mechanism of the double resonator system and a mathematical treatment thereof. Based on the volume, shape and coupling of the resonating chambers, some models of cardboard, tube and plasticene were made, and with which some experimental tests in the production of vowels were carried out. The best success was had with the sound ä (father) while fair results were obtained with the sound ō, ă and ĕ.

INTRODUCTION

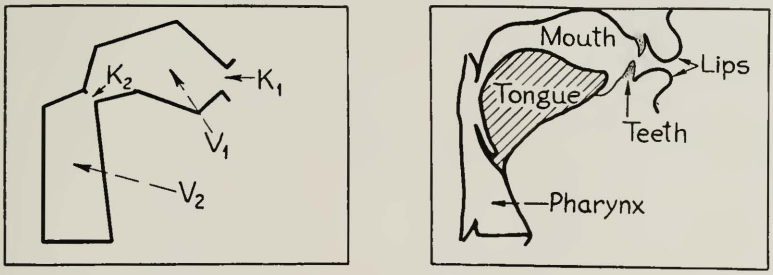
IN two earlier papers [1] a diagram has been given of the frequency spectra of the vowel sounds, based on analyses of a large number of accurate oscillographic records. In addition, there was given, in the second of these papers, a comparative study of the principal characteristics of vowel, semi-vowel and consonant sounds, and an account of certain studies made by other investigators whose methods and results have contributed to our understanding of the mechanism of speech.

Among the more original of recent contributions are those of Sir Richard Paget,[2] who has successfully employed multiple resonators to simulate almost all the vowel and consonant sounds. In getting together the material for the second paper from the Bell Laboratories, Paget's results for the *vowel sounds* were compared with ours only in a general way, and not in so detailed a manner as was followed in the discussion of consonant and semi-vowel sounds in that paper. It may be permissible to return to a consideration of the vowel sounds in the present paper, following Sir Richard Paget's idea of the double resonator as the instrument for vowel production. Indeed Sir Richard has pointed out to us that, since our own data on the spectra of the vowel sounds show almost invariably two principal resonance peaks, there must be a double resonator to produce them, thus harmonizing our results, at least for the male voices, with his own.

[1] I. B. Crandall and C. F. Sacia, "Dynamical Study of the Vowel Sounds," also "The Sounds of Speech," BELL SYSTEM TECHNICAL JOURNAL, III, 1924, p. 232; bid., IV, 1925, p. 586.

[2] *Proc. Roy. Soc.*, A102, 1923, p. 752; ibid., A106, 1924, p. 150.

Fig. 1*a* is a diagram of a double resonator. The volumes of the chambers are respectively V_1 and V_2; the conductivities [3] of the orifices are K_1 and K_2. In this structure the outer orifice corresponds to the mouth (see Fig. 1*b*), the outer cavity to the buccal cavity, the



1*a* 1*b*

Fig. 1*a* and 1*b*—Diagram of the mouth-pharynx system

inner orifice to the constriction between the soft palate and the back of the tongue, and the inner cavity to the pharynx. The source of sound in the back of the inner chamber is of course the glottis, or rather the periodic puffs of air to which the glottis gives rise, and we may remark that at resonance the apparatus is *driven at a node* (or pressure maximum) which is a condition for maximum efficiency. In Paget's models, a small opening was made at the back for the source of sound, which was a loosely stretched strip of rubber, mounted in a slit, and blown by an air stream. To be successful, in connection with the resonator model, in producing a vowel sound artificially, such a source must of course generate a sound whose fundamental is somewhere near that of the human larynx, and which has in addition a very extended range of harmonics; that is, for a bass voice, we should need a fundamental frequency of about 100, and additional sound energy scattered through the frequency range up to 4,000 or 5,000 cycles.[4]

[3] The average mass of air which surges to and fro in the orifice of a resonator is $\rho S^2/K$, in which ρ is the density of air, S the area of the orifice, and K the conductivity. K is a linear quantity, proportional to the width of the orifice, and is a measure of the ease of flow of fluid through it. It may be defined as the ratio of the (velocity) potential difference, between the two ends of the orifice, and the flux or current ($S\xi$) flowing through the orifice.

[4] Sacia suggests that a source of sound giving a saw-toothed wave (rip saw tooth: one slanting and one vertical side) should be ideal for driving vowel resonators. (An experiment with such a device will be described later.) This wave shape corresponds to a fundamental and full retinue of harmonic tones, and should be of service in many ways in acoustic experiments.

Physical Features of the Mouth-Pharynx System

It is a curious fact that most of our data on the shape of the mouth cavities, position of the tongue, etc., for producing the different vowel sounds have been obtained by students of phonetics. There are of course excellent drawings of the mouth structure, in a few typical positions, given in the literature of anatomy; but for the finer differences, from one vowel sound to the next, we must rely on other sources. I know of no determination, for example, of the actual volumes of the mouth and pharynx, in any position for a typical individual, nor have I succeeded, by consulting anatomical experts, in obtaining the desired data.

In Fig. 2, there are shown certain conventional drawings, in median section, of the human mouth-pharynx region. These are taken from Rippmann's "Elements of Phonetics" (London, Dent, 1914) and were taken in turn by Rippmann from an article by Dr. R. J. Lloyd. In

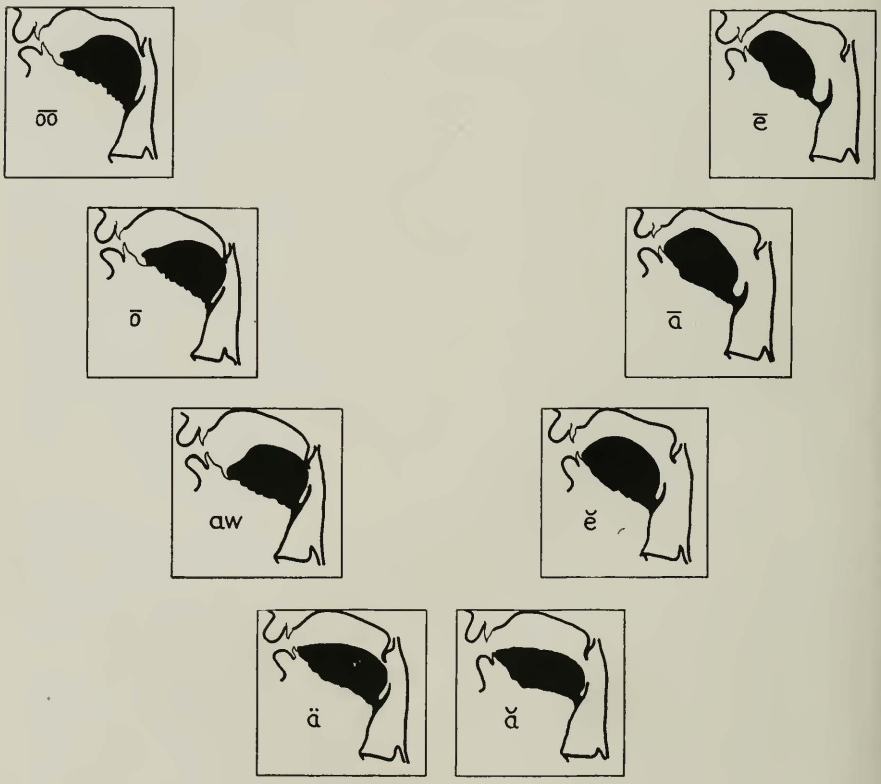


Fig. 2—Diagrams of vocal cavities for various vowel sounds

drawing conclusions from such diagrams as these, we must take care, of course, to use only the broadest features revealed by the series.[5]

It is evident that for the sounds on the left leg of the usual triangle (Fig. 3) (with the exception of short *u*), the inner orifice (that between the back of the tongue and the soft palate) is much constricted, and we

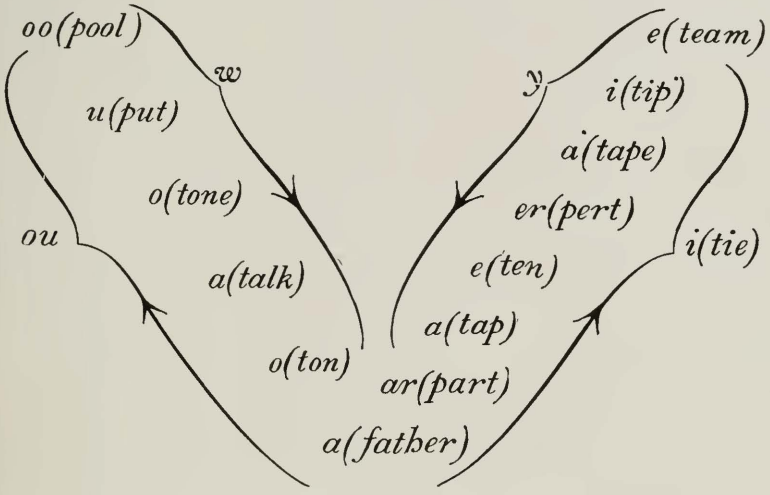


Fig. 3—Conventional vowel triangle

have here a loosely-coupled system to deal with. Also, due to the rearward position of most of the tongue structure, the mouth cavity appears larger than the pharynx. Here we must realize the horizontal width of these cavities, as well as their vertical extent. For the sounds on the right of the triangle, the tongue goes forward in such a way that the front cavity becomes the smaller of the two, and the connecting (inner) orifice becomes larger; the system then becomes closely coupled.

For some of the sounds it is not a difficult matter to get fair values for the conductivity of the mouth opening; this is approximately a circle, or an ellipse of moderate eccentricity in these cases. (The con-

[5] I understand that Prof. G. Oscar Russell, Director of the Phonetics Laboratory, Ohio State University, has made a remarkable series of clear X-ray photographs of tongue and mouth positions, for the various vowel sounds, some of which he has kindly shown me. He has worked out a special technique for making these pictures, and is now engaged in a thorough study of them, which will ultimately be published in monograph form. Unfortunately it is not possible to reproduce the pictures here; but it may be stated that the series follows (but in a more systematic way) the general course exhibited by the Rippmann diagrams shown in Fig. 2 of the present paper. The comparison between the results sketched in the present paper for the mouth cavities and results later to be published by Professor Russell should make a most interesting study.

ductivity of the circle is its diameter; we may take the conductivity of the ellipse as roughly equal to that of the circle of equal area.) In some cases, however (as for example, long $\bar{e}$), where teeth and lips are nearly closed together, the conductivity is certainly less than it appears on merely viewing the opening between the lips; hence a smaller value must be used. The conductivity of the inner orifice is even more uncertain, but in getting at this we are aided to some extent by a theoretical principle which will be given later. The diagrams at least offer some guidance in placing the various conductivities in thei order of relative magnitude.

The most serious lack of data relates to the volumes V_1 and V_2. I have made attempts to fill the mouth with water, and then measure this volumetrically, but of course this gives no hint of the volume of the pharynx. From these experiments, and other considerations, it seems that for an adult male the total volume $V_1 + V_2$ should be something over 100 cm.^{3}, and *nearly constant* for all the vowel sounds. That is to say, the change in V_1 and V_2 consists largely in a shift of volume from V_1 to V_2 (or vice versa) by the movement of the tongue; a proposition not so unreasonable anatomically, because competent advice states that a muscle, in taking its various shapes, preserves the same volume. Finally one would expect a somewhat larger total volume with the mouth wide open, for certain sounds, but this is partially compensated by the flattening of the cheeks in that position.

For the purposes of this study we shall consider $V_1 + V_2 \doteq 120$ cm.^{3} as one of the given data. But it may be stated in passing that these volumes should be much more accurately determined, preferably by anatomical experts.

It would be interesting to compare the results we shall obtain, for the dimensions of the resonator systems, with the actual data of Paget's resonators. But, on account of the four variables involved (K_1, K_2, V_1, V_2), there is no solution of a given case that is unique—that is to say, there are several combinations of different elements possible which will produce a given pair of natural frequencies. Hence such comparisons would often tell us little. Besides, in most cases it is impossible, from the figures given by Paget, to determine the sizes of his resonators, though their shapes are well shown in his drawings. Paget sometimes frankly imitated the structure of the mouth-pharynx system—not necessarily to scale—but sometimes, as in producing double (uncoupled) resonance by resonators in parallel, his models bore no relation to the structure of the natural system.

Spectra of the Vowel Sounds

We shall take as fundamental data the average spectra of the vowel sounds (for male voices) as given by the writer's previous work with C. F. Sacia, and as given in Sir Richard Paget's chart. We thus treat Sir Richard's data as if they had been obtained analytically, and not synthetically, for the sake of taking the mean values of the two most complete series of data available, to get a better basis for calculation.

The two principal resonant frequencies for each sound are given in Table I. The lower characteristic frequency is denoted by $\omega_1/2\pi$; the other by $\omega_2/2\pi$. These characteristic frequencies are also shown in the chart of Fig. 4.

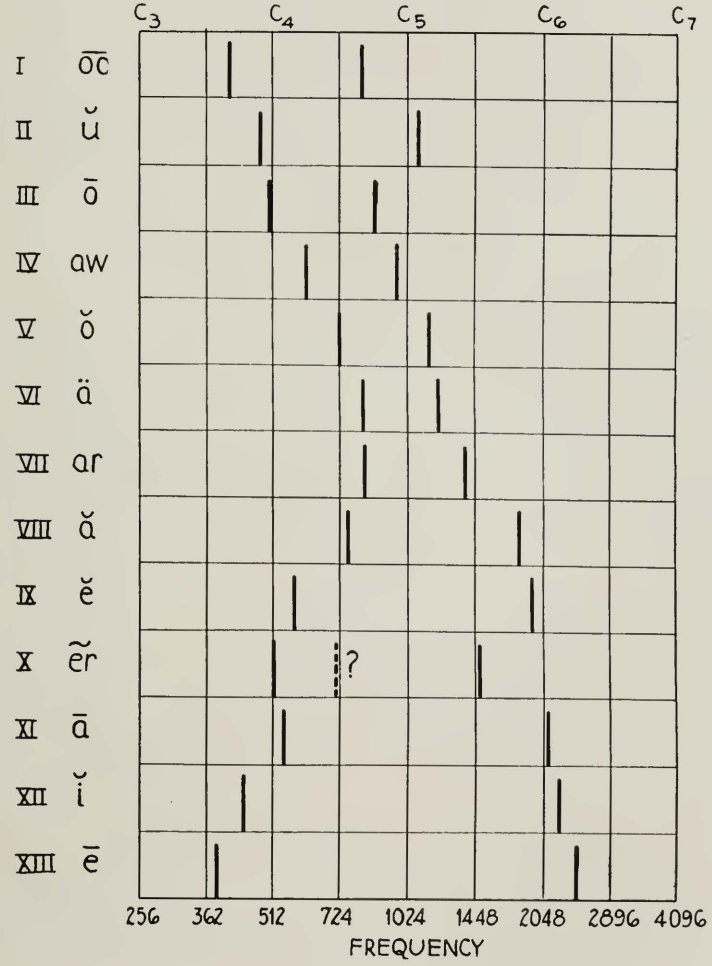


Fig. 4—Principal resonant frequencies, vowel sounds

TABLE I

NATURAL OR CHARACTERISTIC FREQUENCIES OF THE VOWEL SOUNDS

(*Male Voices*)

Sound	$\omega_1/2\pi$				$\omega_2/2\pi$			
	Crandall and Sacia	Paget (centered about)	Mean	Equiv. pitch	Crandall and Sacia	Paget (centered about)	Mean	Equiv. pitch
I. *oo* (*pool*) . . .	431	383	407	$G_3\sharp$	861	724	793	$G_4\sharp$
II. *u* (*put*) . . .	575	362	473	B_3-	1,149	966	1,058	C_5+
III. *o* (*tone*) . .	575	430	502	C_4-	912	790	851	A_4
IV. *a* (*talk*) . . .	645	558	602	$D_4\sharp$	1,024*	886	955	B_4
V. *o* (*ton*) . . .	724	703†	713	$F_4\sharp-$	1,218	1,116†	1,167	D_5
VI. *a* (*father*) .	861	790	825	$G_4\sharp+$	1,149	1,254	1,202	$D_5\sharp$
VII. *ar* (*part*) . .	861	767	814	$G_4\sharp$	1,290	1,491	1,390	F_5+
VIII. *a* (*tap*) . . .	813	703	758	G_4-	1,825	1,824	1,825	$A_5\sharp$
IX. *e* (*ten*)	609	527	568	D_4-	1,825	1,932	1,879	B_5-
X. *er* (*pert*) . . .	542 700‡	470	506 700	C_4	1,448	1,534	1,491	G_5-
XI. *a* (*tape*) . .	609	470	540	$C_4\sharp$	2,048	2,169	2,108	C_6+
XII. *i* (*tip*)	512	362	437	A_3	2,170	2,298	2,234	$C_3\sharp+$
XIII. *e* (*team*) . .	431	332	381	G_3	2,435	2,434	2,435	$D_6\sharp+$

* Poorly resolved, in our charts.

† In Paget's notation, for the sound *o* as in *not*.

‡ Considering *er* to have triple resonance.

The main resonances of most of these sounds are so pronounced that it is not at all difficult to take the correct data from the original charts, gnoring the less-essential minor peaks. In only one case (*a* as in *talk*) does our original chart fail to resolve the two principal peaks, but they are partially resolved even in this case, so that there is no great uncertainty in the figure given. In the case of the sound *er*, a third frequency is shown in the diagram. Reasons will be given later for considering this sound to be produced by a system of three degrees of freedom.

MECHANISM OF THE DOUBLE RESONATOR SYSTEM

There are two ways of studying the action of the double resonator at its resonant frequencies. If we drive the back of the inner chamber with a source of prescribed motion, then the greatest motion in the orifices will be obtained when the driving point impedance of the system *as viewed from the back of the inner chamber* is infinite. Or, equally, if we drive the system (by sound waves, say) from the front orifice, then the greatest motion will be obtained for those frequencies for which the driving point impedance of the system *as viewed from without* is zero. By either method we should be able to deduce the natural frequencies of the system; the second method is chosen here because it involves less labor.

In Rayleigh (II, p. 191, eq. 12) it is shown that the natural frequencies of a double resonator of the type described are the roots ω_1, ω_2, of

$$\omega^4 - \omega^2(n_1^2 + n_2^2 + n_{12}^2) + n_1^2 n_2^2 = 0, \tag{1}$$

in which

$n_1 = c\sqrt{\frac{K_1}{V_1}}$, the natural frequency of the outer resonator, with inner orifice closed;

$n_2 = c\sqrt{\frac{K_2}{V_2}}$, the natural frequency of the inner resonator alone;

$n_{12} = c\sqrt{\frac{K_2}{V_1}}$,

and c is the velocity of sound. Equation (1) is easily obtained by writing the equations of motion of the system, for zero applied forces and zero damping, and placing the determinant of the coefficients of the amplitudes or velocities equal to zero.[6] (This is equivalent to placing the driving point impedance, as viewed from the front orifice, equal to zero.) If $n_{12} \doteq 0$ (the case of a very constricted inner orifice), the roots of (1) are simply n_1, n_2.

We neglect damping in the system in order to get an easily-managed solution for the natural frequencies. Damping arises in two ways: (1) from sound absorption by the soft (tissue) lining of the cavities, and (2) by radiation from the mouth. Both are very variable, that due to radiation particularly so because of the considerable change in size of the mouth opening from one vowel sound to another. A great deal can be learned of the mechanism of the system by studying only the natural frequencies, and although it is not entirely impracticable to solve the problem with an allowance for radiation damping, we shall ignore this here.

The general procedure in this study will be to take as known from the vowel spectra the actual natural frequencies ω_1, ω_2 of the system, and to find the most reasonable values for the four quantities K_1, K_2, V_1, V_2, in order that these natural frequencies may result. We thus reconstruct the hypothetical resonator, or throat-mouth system which produces the vowel sounds. If we take

$$n_{12} = c\sqrt{\frac{K_2}{V_1}} = n_1\sqrt{\mu}, \qquad \left(\mu = \frac{K_2}{K_1}\right), \tag{2}$$

[6] A typical solution of a double resonator problem is given in the author's "Theory of Vibrating Systems and Sound," Van Nostrand (1926), pp. 59–64. The double resonator as a sound amplifier is discussed by E. T. Paris, *Science Progress*, XX, No. 77 (1925), p. 68.

we may rewrite (1) as

$$\omega^4 - \omega^2[n_1^2(1 + \mu) + n_2] + n_1^2 n_2^2 = 0. \tag{1a}$$

If it were not for μ, we could determine from (1a) the ratios K_1/V_1 and K_2/V_2 from the known data ω_1, ω_2. As will appear later, we can make reasonable assumptions with regard to μ; but it is obvious that even then two further assumptions are required to fix K_1, K_2, V_1, V_2 in absolute value. These we supply by assuming a fixed total volume $V_1 + V_2$ for the system, and a certain conductivity K_1 for the mouth opening, which is the most easily observed element of the system.

Proceeding in the manner outlined, it will be possible to take the series of the vowel sounds and fit to each sound a doubly resonant system such that the whole series forms a more or less coherent group.

The following is an outline of the type of calculations required. If we write, from (1a),

$$\begin{aligned} n_1^2(1 + \mu) + n_2^2 &= \omega_1^2 + \omega_2^2, \\ n_1^2 n_2^2 &= \omega_1^2 \omega_2^2, \end{aligned} \tag{3}$$

and eliminate n_2^2, we have

$$\left.\begin{matrix} n_1^2 \\ n_1'^2 \end{matrix}\right\} = \frac{\omega_1^2 + \omega_2^2 \pm \sqrt{(\omega_1^2 + \omega_2^2)^2 - 4(1 + \mu)\omega_1^2\omega_2^2}}{2(1 + \mu)}; \tag{4}$$

also, if we eliminate n_1^2, we have

$$\left.\begin{matrix} n_2^2 \\ n_2'^2 \end{matrix}\right\} = \frac{\omega_1^2 + \omega_2^2 \mp \sqrt{(\omega_1^2 + \omega_2^2)^2 - 4(1 + \mu)\omega_1^2\omega_2^2}}{2}. \tag{4a}$$

In these equations it will be noted that (n_1^2, n_2^2) $(n_1'^2, n_2'^2)$ each represent possible combinations of simple resonators which will give, on coupling, the observed frequencies ω_1, ω_2. In other words, for given (comparable) conductivities K_1, K_2, of the two orifices, the outer resonator may be small, and the inner resonator large ($V_1 < V_2$), corresponding to the (separate) natural frequencies

$$\begin{aligned} n_1^2 \equiv c^2\frac{K_1}{V_1} &= \frac{\omega_1^2 + \omega_2^2 + \sqrt{(\omega_1^2 + \omega_2^2)^2 - 4(1 + \mu)\omega_1^2\omega_2^2}}{2(1 + \mu)}, \\ n_2^2 \equiv c^2\frac{K_2}{V_2} &= \frac{\omega_1^2 + \omega_2^2 - \sqrt{(\omega_1^2 + \omega_2^2)^2 - 4(1 + \mu)\omega_1^2\omega_2^2}}{2}, \\ n_1^2 &> n_2^2; \end{aligned} \tag{5}$$

or, if $V_1 > V_2$, we must apply the other pair of equations

$$n_1'^2 \equiv c^2\frac{K_1}{V_1} < n_2'^2 \equiv c^2\frac{K_2}{V_2} \tag{5a}$$

using the lower signs in (4) and (4*a*). Thus in reconstructing the resonator cavities from the vowel data, we must take care to use, for each particular vowel, that pair of solutions (n_1, n_2, or n_1', n_2') which places the front and rear cavities in correct order for relative size. From the discussion given above of the data on position of the tongue, sections of the cavities, etc., the application of this principle is a relatively easy matter.

The matter of fixing the coupling factor is not so straightforward. For the loosely coupled systems (*oo* to *ar*, the vowels on the left leg of the triangle, Fig. 3), it appears that the maximum allowable coupling factors μ (that is, the values of μ for which the radicals in (4) and (4*a*) vanish) are so small that it seems reasonable to adopt them forthwith.[7] In these cases we have the single solution

$$\left.\begin{aligned} n_1^2 &= \frac{\omega_1^2 + \omega_2^2}{2(1+\mu)}, \\ n_2^2 &= \frac{\omega_1^2 + \omega_2^2}{2}, \qquad n_1^2 < n_2^2; \\ \mu &= \frac{(\omega_1^2 - \omega_2^2)^2}{4\omega_1^2\omega_2^2} = \frac{K_2}{K_1}. \end{aligned}\right\} \tag{6}$$

In this situation (since the ratio V_2/V_1 is fixed if n_1^2/n_2^2 and K_2/K_1 are fixed) all the quantities V_1, V_2, K_1, K_2 are determinate as soon as we fix either K_1 or $V_1 + V_2$. The practice followed will be to set a value for K_1 and check this by noting the value of $V_1 + V_2$ to which it leads; thus by trial and error the most reasonable values for the resonator constants for the loosely coupled systems can be found. Incidentally, we shall note in all these cases that the solution requires V_1 to be larger than V_2.

The vowel short *ă* marks the transition between the loosely-coupled systems already considered and the closely-coupled systems for the sounds from short *ĕ* to long *ē* on the right leg of the triangle. Short *ĕ* is also the first vowel sound of the series to have a high frequency resonance of frequency greater than 1,500 cycles. We might be in a

[7] These values of the coupling factors are not inconsistent with the diagrams of the mouth cavities shown in Fig. 2. Aside from complicating the calculations, the effect of taking still smaller values for μ (keeping K_1 constant) is merely to lower V_2 in proportion as K_2 is decreased. For example, taking $\mu = \mu$ max. for the sound *aw*, we arrive at the solution $V_1 = 119$ cu. cm., $V_2 = 22$ cu. cm., if $K_1 = 2.1$ cm. as given in Fig. 5. Now if we take $\mu = \frac{1}{2}\,\mu$ max., we get $V_1 = 121$, $V_2 = 10$ cu. cm. Thus no great change has been made in the total volume $V_1 + V_2$, except that we get a value for V_2 which seems unreasonably small. The most satisfactory course, in the case of the loosely coupled systems, is to use the maximum allowable coupling factors.

dilemma here, as to which pair of solutions (5 or 5*a*) to apply, since solutions are possible in which the two cavities V_1 and V_2 are of comparable size in this case. It is nearly certain, however, that the front cavity, V_1, is greater than V_2 in this case, but it is not certain that the highest possible value of μ ($\mu = 1$) is the one to use. A compromise was made, setting $\mu = .80$, and using equations (5*a*) for the solution. We shall see later that a resonator built according to these specifications performs sufficiently well to justify these assumptions. With this sound we have finished with equations (6) and (5*a*) and for the last time we have $V_1 > V_2$.

For the last 5 sounds (short *ĕ* to long *ē*) the maximum possible coupling factors range from 1.75 to 9.4; it has been found advisable to shade these and use factors ranging from 1.25 to 5.0. A choice now has to be made between solutions (5) and (5*a*); and since the tongue comes so far forward in these cases, we adopt at once the first solution, according to (5), which leads to the relation $V_1 < V_2$ in all these cases.

Discussion of the Results

The calculated results are shown in the chart, Fig. 5. Because of the speculative character of some of the assumptions made it is reasonable to call attention only to certain outstanding features of the chart. Among the first seven (loosely-coupled) systems the sound *u* (as in *put*), if placed *second*, would seem definitely out of order, because of the magnitude of the coupling factor, or (what is the same thing) the greater separation of the characteristic frequencies. There is no escape from the larger inner orifice for this system, and the effect which it produces. This sound simply does not conform to the habits of its (assumed) neighbors; otherwise the first seven sounds form a coherent group. In classifying short *u* Paget takes the dilemma by the horns, and places it *first*, that is, preceding all the other sounds of this group. This arrangement is adopted in Fig. 5.

There will be noticed in the chart a tendency to expand the total volume, $V_1 + V_2$, for the rounder and more open sounds. This is in a deliberate attempt to allow for the effect of opening the mouth a little wider in these cases.

The last 5 sounds (from short *ĕ* to long *ē*) form a fairly coherent group, except for the non-conforming member *er*. Paget places *er* preceding short *ă* in the series; it seems to the writer a hybrid of the short *ĕ* (or long *ā*) and the *r* sound, but its low frequency resonance (ca. 500) requires a large volume for either V_1 or V_2, and this can only be back of the tongue (V_2) because of the contraction of V_1 when the tip of the tongue is raised for the *r* sound. If we let $K_1 = 1.5$ cm., and

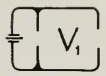
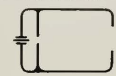
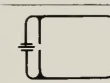
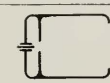
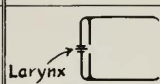
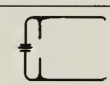
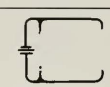
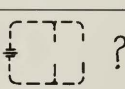
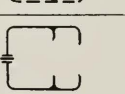
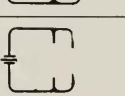
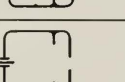
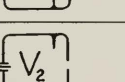
Sound	μ max.	μ used K_2/K_1	K_2	K_1	V_2+V_1	System (Schematic)	V_2	V_1
II ŭ (put)	.80	.80	.96	1.20	137		42	95
I o͞o (pool)	.50	.50	.45	.90	134		34	100
III ō (tone)	.31	.31	.45	1.50	146		28	118
IV o̧ (talk)	.23	.23	.48	2.10	141		22	119
V ŏ (ton)	.26	.26	.73	2.8	134		23	111
VI ä (father)	.15	.15	.52	3.5	126		15	111
VII ar (part)	32	.32	1.12	3.5	127		23	104
VIII ă (tap)	1.00	.80	2.0	2.5	123		23	100
IX er (pert)	1.75	1.00	1.5	1.5	118		73	45
X ĕ (ten)	2.27	1.25	2.25	1.8	117		77	40
XI ā (tape)	3.34	1.8	2.34	1.30	101		73	28
XII i (tip)	6.10	3.0	2.4	.80	102		80	22
XIII ē (team)	9.4	5.0	3.0	.60	104		83	21

Fig. 5—Schematic diagrams of doubly-resonant systems for vowel sounds

assume maximum coupling, i.e., $\mu = 1.75$, we get $V_1 = 98$ cu. cm. and $V_2 = 62$ cu. cm., which seems absurd; if we assumed for *er* a system of only *two* degrees of freedom, the most reasonable course would be to give μ a smaller value (say, unity) and solve on the basis that $V_2 > V_1$ which would give (if $K_1 = 1.5$) $V_1 = 45$ cu. cm., $V_2 = 73$ cu. cm., and $K_2 = 1.5$ cm. These data are entered (very tentatively) in Fig. 5; here again we revise the previous order, and place *er* between short *ă* and short *ĕ*.

It is not at all certain, however, from the spectra of the *er* sound (see chart, Fig. 13, in the paper "The Sounds of Speech") that it is produced by a system of only two degrees of freedom; the analyses of the female voices gave 3 definite peaks, and we note that when the tip of the tongue is raised, for this sound, there is a third cavity betewen the tongue and the lips which is doubtless significant. There will be noted, with a question mark, a third line (of frequency about 700, for the male voices) in the spectrum of *er* shown in Fig. 4. I have attempted, from the three lines shown in Fig. 4, and some simple assumptions regarding the volumes and conductivities, to obtain a rough solution, using 3 degrees of freedom for this sound; but none of these results are entered in the chart, because they appear to be unreasonable.[8]

No attempt has been made to subject the semi-vowel sounds (*l*, *ng*, *n*, *m*) to dynamical calculations. It is evident from their spectra (cf. "The Sounds of Speech") that they are produced by systems of three or four degrees of freedom, which is to be expected, if, in addition to mouth and pharynx, the tongue, naso-pharynx, or

[8] By trial and error it was hoped that some triply-resonant system could be found which would give the spectrum of *er*, as shown in Fig. 4. After solving more than a dozen of these systems, the best fit was one in which $V_1 = 31$, $V_2 = 63$, $V_3 = 31$ cu. cm.; $K_1 = K_2 = 1$ cm., $K_3 = \frac{1}{2}$ cm. The calculated frequencies for this system are 445, 890, and 1,520 cycles. The trouble with this solution is that the middle cavity (V_2, between the tongue and the roof of the mouth in this case) is the largest of the three, which does not seem reasonable. A model made to these specifications, and tried by the method described later, gave a sound something like *er* but not so satisfactorily that one could accept this as a solution. Consequently it is not entered in Fig. 5.

At first, in a number of these attempted solutions, the innermost chamber, V_3, was taken as the largest of the three. These all led to too great a separation of the two lower resonant frequencies to be acceptable.

The sound *er*, in addition to the three resonances about as shown in the chart, may contain a component of higher frequency; or it may be due to a progressive variation or modulation of the two principal frequencies shown in the chart. Some of Paget's results suggest this; and if this is so, it would be a most difficult vowel to imitate with a *fixed* resonator. It is possible that X-ray pictures may reveal some point hitherto overlooked in the mouth adjustment for this sound.

nasal cavities are brought into play. The calculations required would be too cumbersome for the present paper. It is rather a tribute to Paget's experimental skill that he was able to synthesize these more complicated sounds with resonators of more than two degrees of freedom and so arrive at their characteristics.

It is not thought that the calculations given herein suffer appreciably due to the omission of damping factors from the dynamical equations. It would be almost impossible to take correct values of damping constants from the speech spectra; there is a better chance of doing this from the records of the sounds themselves, but even so, they cannot be determined with anything like the precision of the natural frequencies.

To summarize the results, we have an idealized system of two degrees of freedom, loosely coupled for one group of the sounds, closely coupled for the remaining sounds, with fair indication of the transition between the two groups. We have the assumption of virtually constant total volume of the two cavities, and an indication of how this volume should be apportioned between them in most cases. We also have a rough determination in most cases of the conductivity of the inner orifice between the two cavities.

Some Experimental Tests

It would be of interest if we could now make models of all the systems considered, excite them in some suitable way, and establish their essential validity from the character of the sounds produced. This might seem unnecessary, on account of Sir Richard Paget's extended work; it seemed worth while, however, to attempt a few models, using cardboard tubes and plasticene for the structure.

The most success was had with the sound *a* (*f*a*ther*). A model was made to scale (Fig. 6), using the data of the chart—but of course

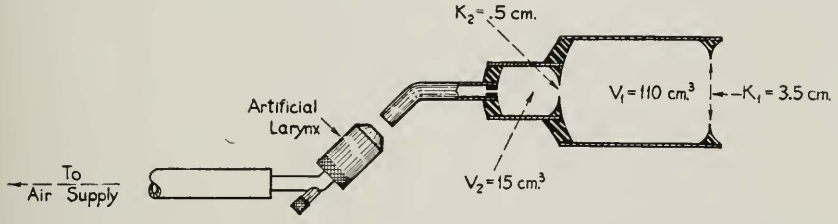


Fig. 6—Double resonator model for *ä*, and method of attaching artificial larynx

we should expect similar results from somewhat larger or smaller models, provided the ratios $K_1 : K_2 : V_1 : V_2$ were maintained; the chief point here is the variation in damping with the sizes of the orifices, and the requirement that any orifice should be smaller than

the mean dimension of the adjacent volume, in order that the usual resonator theory may apply.

The model when gently blown with a slow current of air through the small hole in the back gave a good *whispered ä*; but some difficulty was experienced in exciting it correctly for a *voiced ä*. It was first connected, at the rear, to an artificial larynx,[9] keeping the connecting hole small in order to preserve the dynamical characteristics of the main system. When the artificial larynx was blown (though it did not function well with the output orifice so small), a recognizable *voiced ä* was produced by the apparatus; but this was not as good as the whispered sound first described. (We have here the point made at the beginning: that the driving system, to imitate the vocal cords successfully, must give a low pitched tone, very rich in partials.) The artificial larynx was then replaced by a telephone receiver excited by the (rip) saw-toothed A.C. wave of 100 fundamental frequency, arranged by Mr. Sacia. A rather poor sustained *ä* sound resulted, quite deficient in volume, because of weak driving through the small hole in the back. Altogether, the artificial larynx, with its intermittent or variable excitation, came the nearest to producing a voiced *ä*; and the sound was similar to that produced by a person actually using the artificial larynx inserted in the side of the mouth, in the usual manner, for this sound.

Very fair results were also obtained with a model, built according to specifications, for the sound long *ō*. Models were next attempted for short *ă* and short *ĕ*. First, a model was made with two volumes $V_1 = 80$ cu. cm., $V_2 = 45$ cu. cm., and having the *three* openings K_1, K_2, and the hole in the rear of V_2 (for a cork fitting connecting the larynx) each about 2.5 cm. in diameter. It was thought that, when blown from the rear of V_2, it would give a recognizable short *ă* sound; and that when reversed, i.e., when the cork fitting was inserted in K_1 so that V_1 and V_2 became interchanged, it would give short *ĕ*. The result was that the sounds produced were nearly alike, and quite unsatisfactory in both cases! However, when the conductivities were modified, so that $K_1 = 2.5$, $K_2 = 2.0$, for short *ă*, and $K_1 = 2.0$, $K_2 = 2.5$ for short *ĕ*, the volumes being interchanged as before, the results were much better. As described here, the model for short *ĕ* approximates in dimensions the data entered in Fig. 5, but the model for short *ă* ($V_1 = 80$, $V_2 = 45$ cu. cm.) does not quite have the theoretical division of total volume (namely, $V_1 = 100$, $V_2 = 23$ cu. cm.) entered in the chart. The partition was therefore moved back, until this condition was obtained, with the result that the short *ă* sound was given at least as well as before.

[9] Previously described by H. Fletcher and C. E. Lane.

Attempts were also made at models for long $\bar{a}$ and short $\breve{\imath}$, using the theoretical data. These seemed to give *whispered* sounds which suggested the true ones, but were not very satisfactory when excited by the artificial larynx. It is evidently more difficult to imitate the mouth structure by such simple means, when the outer conductivity (K_1, the orifice between lips and teeth) is small, and the inner orifice K_2 is large. And in addition it is likely that the artificial larynx does not supply sufficient high frequency energy to excite these sounds properly. There is also, of course, the difficulty of applying the simple resonator theory, when the conductivity of an orifice is comparable to one of the dimensions of the adjacent volume.

Conclusion

In this paper we have attempted to visualize the mechanism of the vowel sounds, on the basis of previous work, certain simple calculations, and a few rough experiments. It appears that the vowel sounds are usually produced by a double resonator system whose behavior in itself is thoroughly understood; but this does not by any means close the subject. A most interesting field of study remains in the excitation of the resonator system, to say nothing of the various factors which produce damping in the system itself.

We know from laboratory experiments that a reed (or a simple "squawker" made of rubber strip) is by itself a very poor imitation of the vocal cord apparatus. The artificial larynx, for example, will not vibrate properly unless a tube some 15 inches long is interposed between the "larynx" and the pressure reservoir by which it is blown. Correspondingly, we should expect the wind-pipe leading from the lungs to the human larynx to have a very important rôle in fixing the lower frequencies produced by the vocal cord apparatus. The mechanical problem indicated for study in this connection is the excitation of a reed-pipe with the reed at the distant end of the pipe, an inversion of the arrangement of ordinary wind instruments.

Consider the question of damping. In the apparatus used by J. Q. Stewart [10] (tuned electrical circuits excited by an interrupter) the damping could be systematically adjusted; this is the only case I know of, in experimenting with speech sounds, in which this adjustment was possible. In ordinary mechanical apparatus damping is difficult to control. Yet, damping is a significant element in the character of the constituent vibrations of either sustained or transient vowel sounds. For example, I have already pointed out [11] the close

[10] *Nature*, Sept. 2, 1922. "An Electrical Analogue of the Vocal Organs."

[11] "The Sounds of Speech," end of § V. Refer also to Records and Fig. 14 of that paper.

similarity between the spectra of *l* and long $\bar{e}$. In the semi-vowel *l* the characteristic high frequency (if viewed as a transient) decays much more rapidly than the corresponding vibration in the $\bar{e}$ sound; this fact we have from the records themselves, but not from the frequency spectra. It may be that such phenomena as these will require a more definite adherence to the "transient" point of view in dealing with the vowel sounds, a matter previously discussed at some length.

The transitory or unstable qualities in the actual speech sounds almost defy imitation by mechanical means. There is, for example, the variation in fundamental frequency during the course of a vowel or semi-vowel sound which was pointed out in the paper "The Sounds of Speech." There is also the lengthening of the fundamental period for semi-vowels and voiced consonants as compared with vowel sounds; also the shortening of the fundamental cycle at the beginning of a voiced consonant.

Finally there is the question of classification of the speech sounds. We have already noted difficulties for some of the vowel sounds. It is likely that the vowel triangle or the arrangement of the vowels in a linear series will require modification. A satisfactory classification for all the sounds, from the dynamical standpoint, is at present an unsolved problem; but in conclusion one suggestion may be permissible. We might limit the application of the term "vowel sound" to those sounds which *can* be satisfactorily produced by the simple double resonator system. The more complicated vowel-like sounds (*l*, *ng*, *n*, *m* and possibly *r*) and some of the consonants can undoubtedly be related to systems of three or more degrees of freedom. A study of these systems is beyond the aims of the present paper; but it is to be hoped that such a study can be carried out, for the sake of the aid that mechanical theory offers in helping to visualize the mechanism of speech.

Radio Broadcast Coverage of City Areas [1]

By LLOYD ESPENSCHIED

SYNOPSIS: 1. Radio broadcasting involves a system of electrical distribution in which dependent relations exist between the transmitting station, the transmitting medium and the receiving station.

2. The attenuation and fading which attend the spreading out of broadcast waves are considered. The attenuation of overland transmission is shown to be, on the whole, very high and to vary over a wide range depending upon the terrain which is traversed. The distance at which the fading of signals occurs is found to be that at which the normal directly transmitted waves have become greatly attenuated and to depend upon the terrain traversed.

3. A field strength contour map is given of the measured distribution of waves broadcast by Station WEAF over the New York metropolitan area. A rough correlation is given between measured field strengths and the serviceability of the reception in yielding high grade reproduction. The range of a station as estimated in terms of year-round reliability is found to be relatively small. It becomes clear that the present radio broadcasting art is upon too low a power level and that higher powered stations are required if reliable year-round reception is to be had at distances as short even as 30 to 50 miles from the transmitting station.

4. The question of the preferred location of a transmitting station with respect to a city area is considered. It is shown that an antenna located upon a tall building may radiate poorly at certain wave-lengths and well at others. Surveys are presented of the distribution effected by an experimental transmitting station located in each of several suburban points. The locations are compared upon the basis of the "coverage" of receiving sets which they effect.

5. Finally, there is considered the relation which exists in respect to interference between a plurality of broadcast transmitting stations operating in the same service area. The importance of high selectivity in receiving sets is emphasized and there is given the measured selectivity characteristics for samples of a number of receiving sets.

IT is well recognized that the elements which comprise an electrical transmission system are required to function not simply as individual pieces of apparatus, but as integral parts of a whole. In the case of radio broadcasting, the absence of a common control of the two ends makes this over-all "systems" aspect less apparent than it is for wire systems.[2] Nevertheless a definite systems correlation is required between the broadcast transmitting station and each of the receivers served, as will be evident from the following:

1. The transmitter should put into the transmitting medium, without distortion and with the power called for by that medium, all of the wave-band components required and no others.

2. The transmitting medium should be capable of delivering to the receiver an undistorted wave band, reliably and stably, and with

[1] Presented at the New York Regional Meeting of the A. I. E. E., New York, N. Y., Nov. 11–12, 1926.

[2] Some other examples of such a "systems" relationship are given in "Application to Radio of Wire Transmission Engineering," published in the *Proceedings* of the Institute of Radio Engineers, October, 1922.

sufficient strength to enable the received waves to stand well above the level of the ever present interfering waves.

3. Finally, the receiving set should pass with the necessary volume all of the wave components required to reproduce the program signal and should sharply exclude all others.

The rapid apparatus development borne in by the vacuum tube has brought the art to the point where it is now physically possible to meet quite fully the terminal requirements. The apparatus development, in fact, has outstripped our knowledge of the transmitting medium itself, and we are now in the position of possessing apparatus possibilities without knowing very definitely the limitations and requirements placed upon their use by the intermediate link. Only within the last few years have methods become available for measuring radio transmission and thereby placing it upon a quantitative basis.

Such measuring means have been applied to the study of radio broadcast transmission from certain stations in New York City and in Washington, D. C. The earlier results of this measurement work have already been published. It is the purpose of the present paper to present results of a systematic study which has been made of the coverage which can be effected of the radio broadcast listeners of the New York metropolitan area and in so doing to portray something of the general systems requirements of radio broadcasting.

The Character of Radio Broadcast Transmission

The ideal law for broadcast distribution would be one whereby the transmitted waves are propagated at constant strength over the zone to be served and then fall abruptly to zero at the outer boundary. All receivers within the area would be treated to signals of equal strength and no interference would be caused in territories beyond.

The kind of law which nature has actually given us involves a rapid decadence in the strength of the waves as they are propagated over the service area, and then, instead of a sharp cut-off, a persistence to great distances at field strengths which, although often too low to be generally useful, are sufficient to cause interference in other service areas.

This situation is illustrated in Fig. 1. The upper curve shows the relation between intensity and distance; the lower portion, the interpretation of this curve in terms of areas of reception. The attenuation traced by the heavy line of the curve is that of the component of the radiation which is propagated directly along the earth's surface. It is this radiation which is ordinarily utilized for reliable broadcast reception. The shaded portions near the outer ends of this curve are

intended to indicate the appearance of variations in the signal intensity which occur at the greater distances, particularly at night, and which are known as "fading."

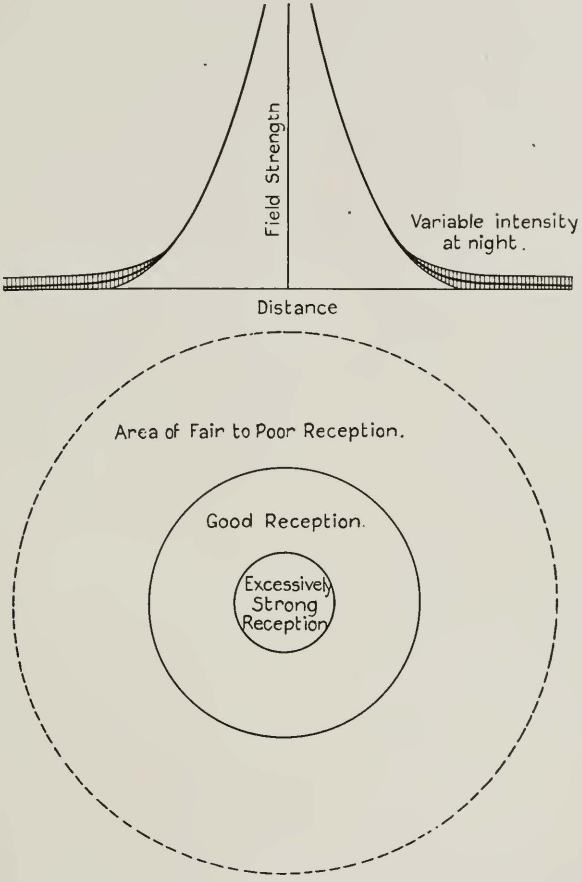


Fig. 1—The attenuation of broadcast waves in reference to the areas served

The evidence of recent researches, particularly those made at short wave-lengths, indicates that these fading variations are due to radiant energy which has left the earth's surface at the radio transmitter and has been reflected or refracted back to the earth's surface from a conducting stratum in the upper atmosphere. At broadcast frequencies the reflected wave component is observed at night but has not been noticed during the day. At locations close to the transmitting station the effect of the reflected component is negligible as compared with the strength of the directly transmitted waves. At

increasing distances the directly transmitted waves die away to very low values and the indirectly transmitted waves begin to show up and appear to become controlling at the longer distances. The fluctuations themselves appear to be due in part, if not entirely, to variations in the reflected waves themselves, resulting perhaps from fluctuations in the conditions of the upper atmosphere.

Thus, it seems clear that radio transmission involves wave components of two types: one which delivers directly to the receiving area immediately surrounding a broadcast station, a field capable of giving a reliable high grade reception; and another transmitted through the higher altitudes which permits distant reception but not with the reliability and freedom from interference required of high grade reproduction.

The effects which are actually realized in practice are indicated in a more quantitative manner by the curves of Fig. 2 which are plotted from some measurements made upon WEAF in New York and WCAP in Washington, D. C. They were made at locations in the New York and the Washington areas and at the intermediate points

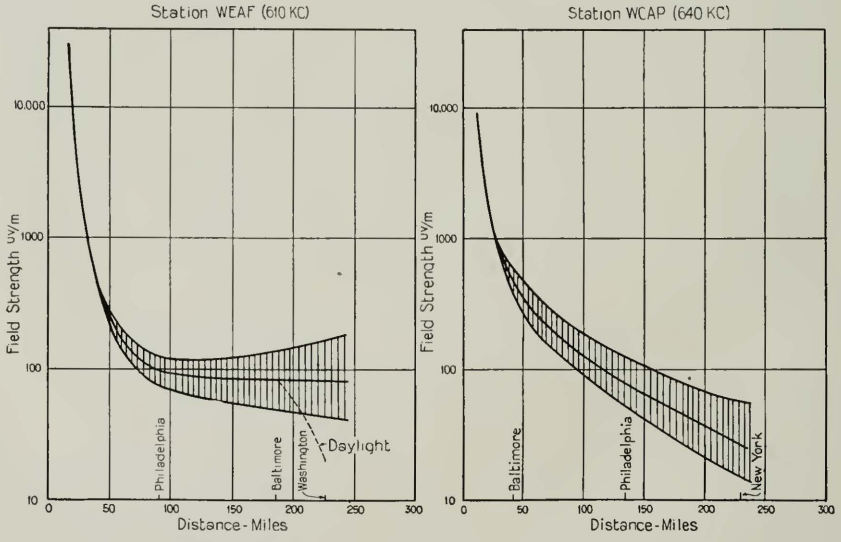


Fig. 2—Results of a few measurements upon the reduction in field strength with distance, including distances at which fading occurs

indicated on the curves. The measurements at each of these points are for one day only. They consisted in obtaining continuous graphic records of signal intensity during twenty-minute intervals out of each hour, one interval for each of the two stations. The period of time

covered for each set of measurements was that of from one hour before sundown to about three hours after sundown. The time of year was the latter part of May, 1926. The curves are plotted from an analysis of the records in terms of mean field strength. The range of variation due to fading is indicated by the shaded portions of the curves. The day and night fields were found to be roughly the same except for WEAF where there is a material drop in the daytime signal between Baltimore and Washington, shown in the WEAF curve.

Fading was observed to commence somewhere between 50 and 100 miles from the stations and the range of the fluctuations was found to increase up to the maximum distance observed. That the field of WEAF was found to be practically as strong at Baltimore as at Philadelphia is surprising. The data regarding this point are too meager, however, to enable any very definite conclusions to be drawn. The curves are useful principally in enabling the transition to be followed, in a more quantitative way than is done in Fig. 1, from field strengths capable of giving reliable reception, such, for example, as 10,000 μv./m. (microvolts per meter), to those which characterize the unreliable "distance" reception and are of the order of 100 μv./m.

A fact which is of importance to our understanding of these wave phenomena is that "fading," which ordinarily is noticed at distances of the order of 100 miles, may under some conditions become prominent at distances as short as 20 miles from the transmitting station. Such short-distance fading has been experienced in receiving WEAF in certain parts of Westchester County, New York.[3] It appears to be a case where unusually high attenuation, caused by the tall building area of Manhattan Island, has so greatly weakened the directly transmitted wave as to enable the effect of the indirect wave component to become pronounced at night.

In general, the attenuation suffered by the normal surface-transmitted waves varies over wide limits depending upon the terrain which is traversed. This is disclosed by the curves of Fig. 3, which show the drop in field strength with distance, for a 5 kw. station, for each of the following conditions:

a. No absorption, the inverse distance curve ($\alpha = 0$),
b. Sea water, for which the absorption is relatively small ($\alpha = 0.0015$),
c. Open country and suburban areas ($\alpha = 0.02$ to 0.03) as measured in the vicinity of New York and Washington, D. C.,
d. Congested urban areas ($\alpha = 0.04$ to 0.08) as measured for Manhattan Island.

[3] See "Some Studies in Radio Broadcast Transmission," by Ralph Bown, D. K. Martin and R. K. Potter; *Proceedings*, I. R. E., Feb., 1926.

The factor α will be recognized to be the absorption factor of the familiar Austen-Cohen empirical formula, which may be expressed as

$$e = 0.009 \frac{\sqrt{P}}{d} \epsilon^{-(\alpha d/\sqrt{\lambda})}$$

in which

P = radiated power in watts,
d = distance in kilometers,
λ = wave-length in kilometers,
α = absorption factor,
e = in volts per meter.

The first term represents the decrease in strength due merely to the spreading out of the waves; the second term, the decrease due to the

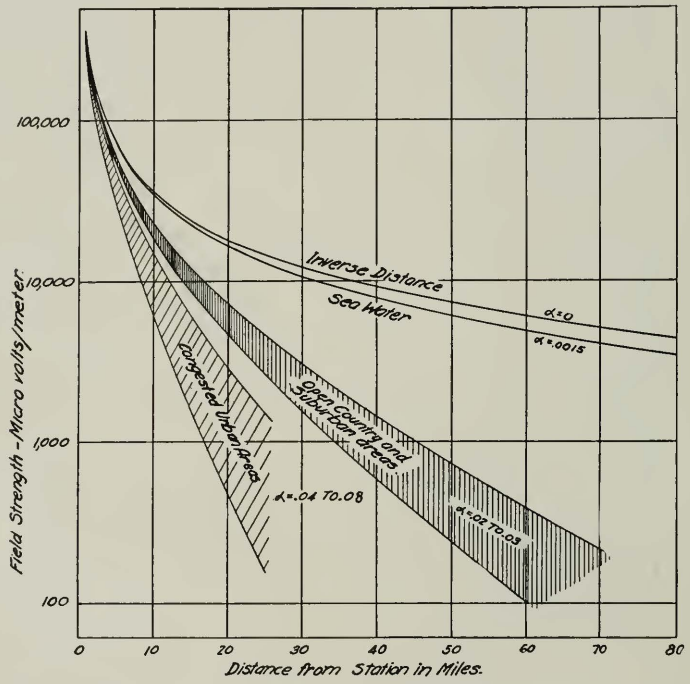


Fig. 3—Effect of the terrain in reducing the field strength of a broadcast transmitting station. 5 kilowatts in antenna, frequency 610 kilocycles, wave-length 492 meters

absorption of the wave energy by the imperfect conductivity of the earth's surface.

The curves given in Fig. 3 are derived from a considerable amount of data taken in the course of field strength surveys of the New York

City and Washington, D. C., areas. The results of some of the earlier of these surveys have already been published.[4]

Actual Distribution in New York City

Fig. 4 presents the results of a detailed survey of the field distribution effected over the New York metropolitan area by Station WEAF located at 463 West Street. The measurements upon which the plot is based were taken in the daytime during the summer of 1925. Measurements were taken at approximately one-mile intervals along each of a series of circular paths concentric with the station, the radii of which increased in steps of approximately five miles. The distribution was studied in even greater detail close to the station and in locations giving evidence of rapid change in field strength. Ferries were utilized to extend the measurements over bodies of water. Manhattan Island was circumscribed on water by measurements made upon a sight-seeing boat. The land measurements were made in all cases outside of buildings at ground level. In the built-up sections of the city they were taken in the middle of streets and street intersections, and in so far as possible in open places. The plot is based upon over 1000 measurements. While these measurements were taken over a considerable period of time, check measurements proved conditions to have remained quite stable and showed, in fact, little variation from measurements made the previous year. The type of measuring apparatus employed, together with certain of the results obtained in earlier surveys, has already been described.[5]

This plot is actually a simplification of a more detailed one. The number of contour lines has been limited to those of round figures for the sake of clarity. The line marked 10,000, for example, traces the locations at which that field strength was observed and beyond which lower values obtained.

This survey shows strikingly that the terrain over a city like New York is anything but uniform electrically; that the variations in the attenuation which the waves experience in different directions and from one area to another distort the distribution pattern from that which we might imagine from the familiar stone in the pool analogy. It is apparent that this simple analogy will have to be amended by conceiving the pool to be beset by various encumbrances causing high attenuations and reflections; and, in fact, also by the presence of

[4] "Distribution of Radio Waves from Broadcasting Stations over City Districts," by Ralph Bown and G. D. Gillett, published in the *Proceedings* of the Institute of Radio Engineers, August, 1924.

[5] See previous reference; also "Portable Receiving Sets for Measuring Field Strengths at Broadcasting Frequencies," by Axel G. Jensen; *Proceedings*, I. R. E., June, 1926.

surface ripples to represent the waves foreign to broadcasting which cause interference. Sight should not be lost, however, of the fact that the contour lines of Fig. 4 represent a two-dimensional section of



Fig. 4—Field strength contour map of distribution over the New York metropolitan area, effected by Station WEAF. 5 kilowatts in antenna, frequency 610 kilocycles, wave-length 492 meters.

a three-dimensional phenomenon. One should picture the contours as the intersections of the earth's surface with three-dimensional surfaces.

The fact previously referred to that the waves transmitted into Westchester County experience high attenuation is shown by the shape of the contour lines. The irregularity of the lines appears to be due to a splitting of the directly transmitted wave by the high building area and the filling in from the sides of wave energy transmitted along the two sides of the peninsula. Although the conditions in Westchester are quite stable during the daytime, they become unstable at night due, apparently, to the addition of the indirectly transmitted component reflected from above. An experimental study of this interference situation disclosed the fact that the bad quality

obtaining at night in certain parts of Westchester was due largely to a rapid frequency modulation of the broadcast transmitter. The frequency fluctuation of the transmitted band apparently caused the direct and indirect transmissions to slip in and out of phase rapidly. The use of a master oscillator control for insuring stability of frequency greatly improved matters, but evidence still remains of what might be called the normal night-time fading.

Another interesting effect which stands out in this map is the high attenuation of the wave-front transmitted over Long Island as compared with that which pursues the path of Long Island Sound and that of the ocean front to the south. The field over the eastern half of the island is contributed to by the water-transmitted waves from either side, giving rise to interference patterns similar to those in Westchester County.

A question which naturally arises is that of how strong a field, as measured in this way, is required for satisfactory reception. It is too early in the art to answer this question very definitely, for it depends first upon the standard of reception which is assumed, with respect to quality of reproduction and freedom from interference; and second upon the level of the interference. The interference, both static and man-made, varies widely with time and with location. It is therefore obviously impossible to give anything more than a very general interpretation of the absolute merit of field strength values. Observations made by a number of engineers over a period of several years in the New York City area, having in mind a high standard of quality and of freedom from interference, indicate the following:[6]

1. Field strengths of the order of 50,000 or 100,000 μv./m. appear to be about as strong as one should ordinarily desire. Fields much stronger than this impose a handicap upon those wishing to receive some other station.

2. Fields between 50,000 and 10,000 μv./m. represent a very desirable operating level, one which is ordinarily free from interference and which may be expected to give reliable year-round reception, except for occasional interference from nearby thunder storms.

3. From 10,000 to 1000 μv./m. the results may be said to run from good to fair and even poor at times.

4. Below 1000 μv./m. reception becomes distinctly unreliable and is generally poor in summer.

5. Fields as low as 100 μv./m. appear to be practically out of the picture as far as reliable, high quality entertainment is concerned.

[6] See also the paper by A. N. Goldsmith, "Reduction of Interference in Broadcast Reception," *Proceedings*, I. R. E., October, 1926.

Such fields, however, may be of some value for the dissemination of useful information such as market reports, where the value of the material is not dependent upon high quality reproduction.

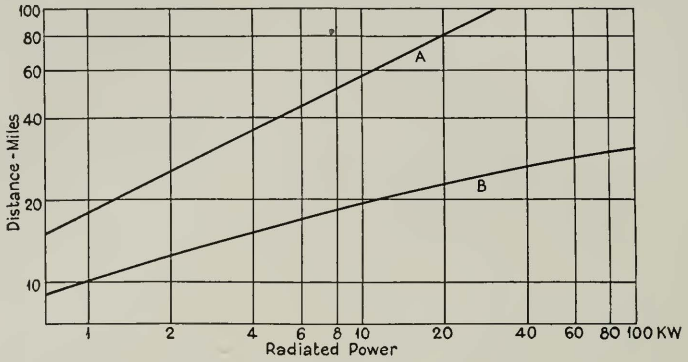


Fig. 5—Showing the increase in radiated power required to increase the range at which a field of 10,000 μv./m. is delivered. Curve *A* without absorption and curve *B* with absorption.

It is seen from the preceding three figures that a 5 kw. station may be expected to deliver a field of 10,000 microvolts some 10 to 20 miles away and a 1000 microvolt field not more than 50 miles. From this it will be evident that the reliable high quality program range of a 5 kw. station is measured in tens of miles rather than hundreds.

Higher Power Transmitting Stations Required

Rough though this interpretation of field strengths is, it indicates clearly the need which exists for the employment of higher transmitting powers. The range goes up with the increase of power disappointingly slowly. Even were no absorption present in the transmitting medium, the range in respect to overcoming interference would increase only as the square root of the increase in power. This is shown in the curve *A* of Fig. 5. It shows that a station which actually radiates five kw. of power would deliver a 10,000 μv./m. field at about 40 miles, a 20 kw. station the same field at distance 80 miles. Actually with absorption present the distances are less. This is shown by the curve *B* which gives the corresponding relations for the absorption observed for suburban and country terrain. To extend the 10,000 microvolt field from some 15 miles out to 30 miles would necessitate an increase in the radiated power from about five to 100 kw.

It is apparent from these relations that radio broadcasting is today underpowered; that the common 0.5 kw. station is entirely too small to serve large areas adequately, and that the more general use of

powers of the order of five kw. and even 50 kw. is decidedly in order. Such increases in power will be required if the broadcasting art is to be advanced to meet the higher standards of the future. The fact should be recognized that no greater interference between stations will be caused by the higher power levels, providing the increase in power is general among all stations. The interference difficulty arises in particular cases where one station suddenly makes a large increase and the others remain at their previous low power levels.

Mention should perhaps be made that the effect of raising the transmitter power in increasing the level of the *detected* signal is greater than would be inferred from the discussion above. This is because of the square-law action of the detector. In other words, the detector output reflects the increase in power of the carrier as well as the side band. In overcoming interference it is only the increase in side-band power which counts.

The ideal broadcast system from the transmission standpoint would be one in which the carrier is not transmitted from the sending station but is automatically supplied in the receiving sets themselves. This would save power, would reduce interference between stations and would reduce fading. It will be recalled that this system is being used to great advantage in the transatlantic radio telephone development. The practicability of employing it in broadcasting will depend upon receiving set development,—upon the economy with which carrier-generating receiving sets can be made and the ease with which the carrier frequency can be set and maintained with the necessary accuracy.

Transmitting Station on Tall Building

The location which naturally suggests itself for a broadcast station intended to serve a city is that of its center. Such a location might be expected to deliver the greatest strength of field to the greatest numbers because of the coincidence between high field strengths and high density of population. The other possibility, of course, is that of placing the station outside of the city, with the object of obtaining a better "get-away" condition, of covering a larger area and of laying down a more uniform, if less strong, field over the city itself. Instances of both of these types of locations readily come to mind. WEAF is a good example of a station located near the center of a large city. The results of a study which has been made upon the effect of moving the station to other possible locations are given below.

Before coming to this, however, there is another important factor to present and that is the effect of placing the transmitting station

upon a tall building. In locating a station near the center of a large city it is natural to select a tall building for the station site. This has been done for a number of stations in various cities. The operation of WEAF, when known as WBAY, was first attempted from the top of the 24 story long-distance telephone building located at 24 Walker

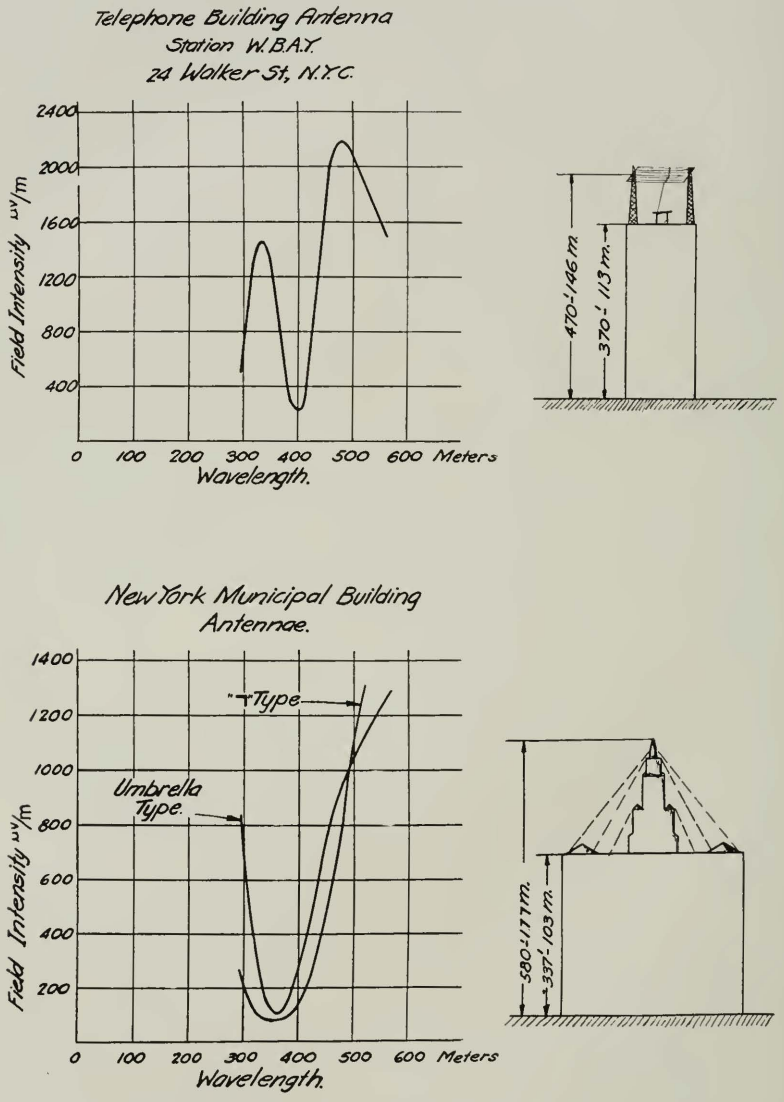


Fig. 6—The selective radiation characteristic of transmitting antennas on tall buildings

Street, New York City. It was found that with the limited wave-length range then open to broadcasting, radiation from the station was relatively poor. Measurements of the field strength delivered to a field laboratory located at Cliffwood, N. J. (on lower New York Bay), were made which gave the results shown in Fig. 6. The radiation was found to be sharply selective with respect to frequency, and to drop to a very low value at 400 meters. This happened to have been the wave-length assigned to the station at the time. When it became possible to shift the station to a longer wave-length, radiation was greatly improved, as indicated by the curve. The study made on this station was the first to disclose the fact that it is possible to have the building too high for the efficient radiation of certain frequencies.

As a result of this work it was possible to predict the probable occurrence of a similar effect in the case of a station which the City of New York desired to establish on the Municipal Building. Temporary antennas were erected and radiation from them measured at Cliffwood, N. J., using a transmitting oscillator of 100 watts. The results of these measurements are given in Fig. 6. The radiation was found to be a minimum in the vicinity of 360 meters, which was very nearly the wave-length which at that time was to have been assigned to this station. The establishment of the station at this location obviously could not be recommended until at a subsequent time when a longer wave-length was made available. The station is now operating on 526 meters, which is seen to be fairly well up on the radiation curve.

In both of these cases experiments were made with a number of different antenna arrangements and with different methods for driving the antenna and effecting the ground connection. None of the modifications, however, materially shifted the frequency of minimum radiation. This minimum occurs when approximately one quarter of the wave-length equals the height of the building. Measurements made upon buildings of lower heights have shown that for the usual broadcast wave-lengths heights of the order of 200 ft. are entirely satisfactory. The antenna of WEAF (which has been located for the past several years on the building of the Bell Telephone Laboratories, 463 West Street, New York), and that of WCAP, in Washington, are on buildings which put them at about this height above the street. They both have normal radiation characteristics.

Distribution from Suburban Locations

In order to determine the distribution over New York City which might be effected from locations outside of Manhattan Island, experi-

mental transmitting stations were established at each of several suburban locations. Use was made of an automobile truck equipped with a ½ kw. broadcast transmitter and provided with a transportable



Fig. 7—Transportable field transmitting station

mast. The experimental transmitter as set up at Secaucus, New Jersey, is shown in Fig. 7. Measurements of the field strength delivered from each of the locations chosen were made over practically the entire metropolitan area. The results of these tests are given in Fig. 8, in comparison with those of transmission from the normal location of WEAF at West Street and from the earlier location at 24 Walker Street. The measured field strengths have been adjusted to correspond to the 5 kw. transmitter of the West Street station.

The smaller irregularities in the West Street curve as compared with the others are due to the greater detail with which these measurements were made. The curves should be compared merely with respect to their major contour characteristics. The inner contour line is for 50,000 μv./m. and the outer line for 10,000 μv./m. Actually, the measurements were made in sufficient detail to enable other contour lines to be drawn, but these have been omitted for the sake of simplicity.

The radiation from Secaucus will be seen to deliver a strong field to Manhattan Island, the most densely populated section, and, in general, to encompass the rest of the city quite well within the 10,000 μv. line. The irregularity in Queens County evidently represents the shadow cast by the tall building area on Manhattan Island.

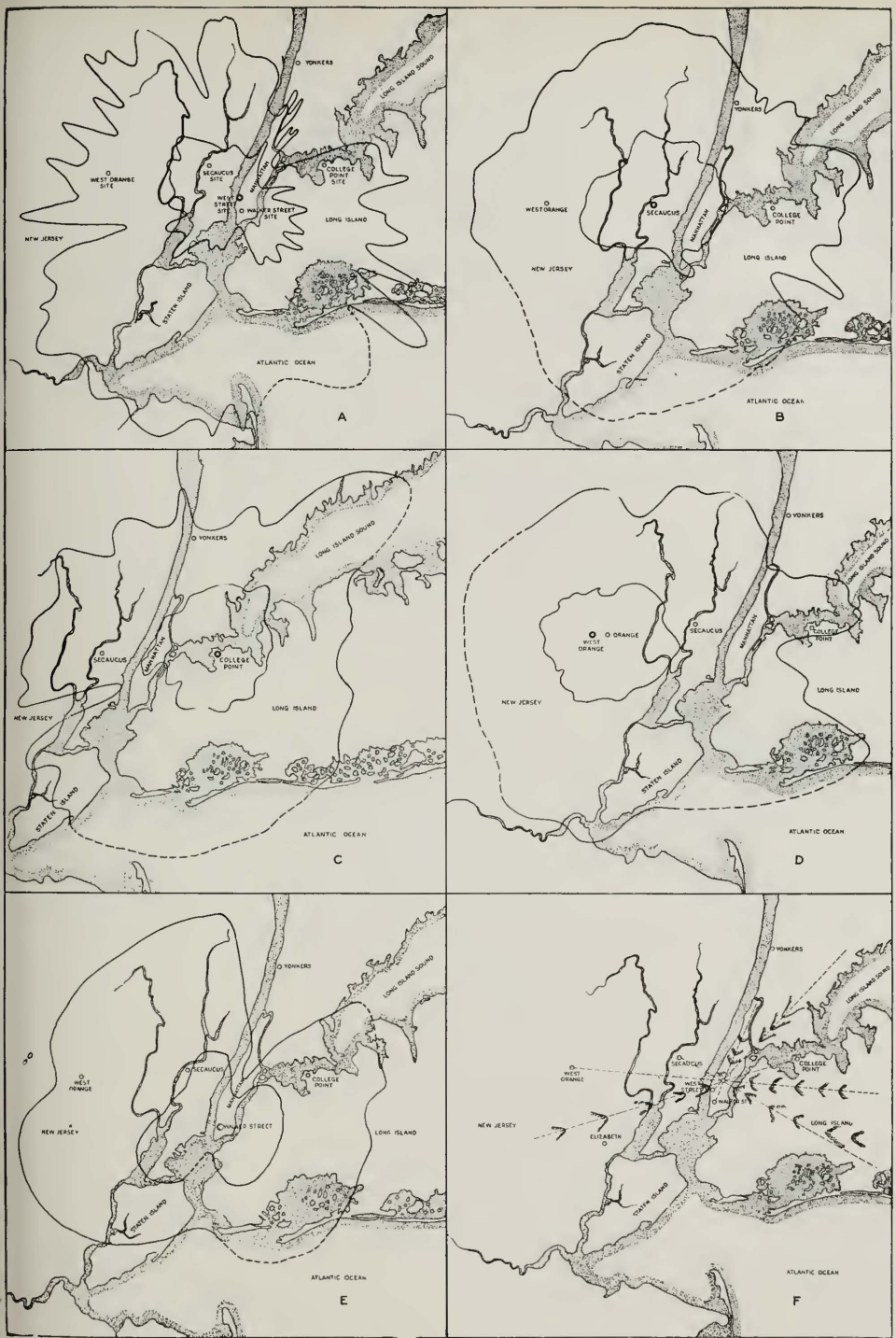


Fig. 8—Effect upon the field distribution of moving the transmitting station to suburban locations

A—463 West Street, New York City
B—Secaucus, N. J.
C—College Point, Long Island, N. Y.
D—West Orange, N. J.
E—24 Walker Street, New York City
F—Composite figure showing main shadows and center of obstacle

The distribution effected from the College Point location appears to be generally good. It does not cover the New Jersey suburbs as strongly as might be desired. The shadow cast by the Manhattan Island high buildings lies through Jersey City and lower Newark.

The distribution from the West Orange site appears to be somewhat less favorable. It is not sufficiently close in to deliver with moderate power a very strong field to the center of the population, nor is it sufficiently far out to avoid subjecting a considerable population in the immediate vicinity of the station to an excessive field were high power employed. The indent in the 10,000 μv. line in northern Queens is the shadow of the Manhattan buildings.

The distribution shown for the Walker Street location is seen to be generally similar to that of West Street. The curve presents a smoother appearance than the others because less data were taken in this one of the earlier surveys. The shadows cast to the north and to the south by the two areas of high buildings are prominent. Actually, a close examination of the contour lines reveals a noticeable angular displacement in the Westchester shadow as between Walker Street and West Street, Walker Street transmitting better up the Sound and West Street better up the Hudson. West Street turns out to be somewhat the better of the two.

The last diagram of the series brings together the shadows as determined from the several transmitting sites and shows that they project back to a common general center which locates at approximately 38th Street and Broadway, which corresponds quite well with the center of the up-town tall building area.

Relation between Wave Distribution and the Distribution of Listeners

The merit of a given distribution pattern obviously depends upon the relation which exists between it and the distribution of the receiving sets themselves. In order to study this relation more closely, the relative distribution of receiving sets was approximated by taking the distribution of residence and apartment house telephones in each of the central office districts of the metropolitan area, excluding the commercial telephones. It was assumed that the receiving set distribution is proportional to that of the telephones. For a given survey the field strength representative of each central office district is known. By assembling the figures for central office areas receiving like field strengths, and by doing this for the whole range of field strengths measured, an accumulative percentage curve may be derived which shows the percentage of the total number of receiving sets included within the contour lines of successively weaker fields.

Curves of this kind for each of the several surveys made are shown in Fig. 9. It will be seen that for field strengths of 10,000 μv./m. and better, the Secaucus and College Point transmitting sites include about 80 per cent of the receivers, that the West Street and West

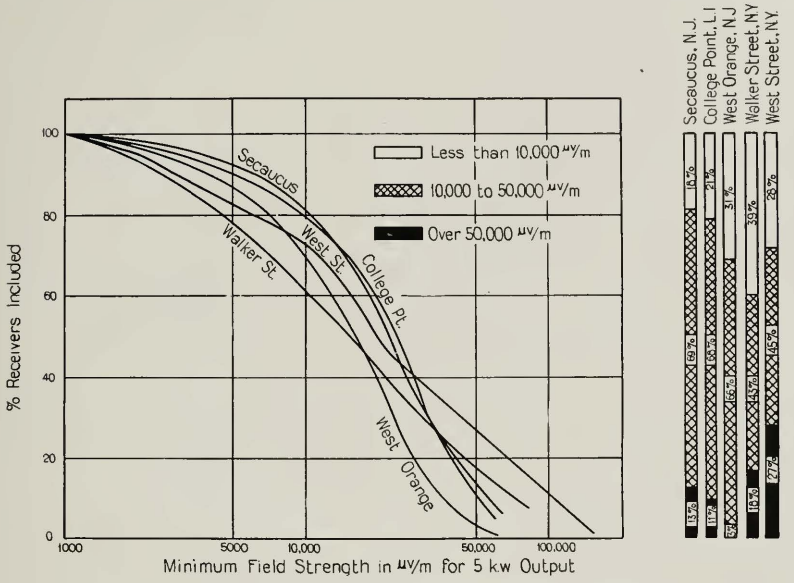


Fig. 9—Percentage of receiving sets in metropolitan area included within various strengths of field for each of the transmitting locations of Fig. 8

Orange sites include around 70 per cent and Walker Street about 60 per cent. These curves are further analyzed in the chart to the right of the figure to show in each case the proportion of the listeners which may expect to receive

a. less than 10,000 μv./m.,
b. between 10,000 and 50,000 μv./m.
c. over 50,000 μv./m.

It is seen from this that a location to the east or west of Manhattan Island would give a material improvement in uniformity of distribution as compared with a location on Manhattan Island. Had it been possible to include a station on Manhattan Island located farther north than is either West Street or Walker Street and included within the area of high steel buildings, it is probable that the corresponding curves for such a location would show the poorest distribution of the series.

The survey work described above did not go so far as to include a study of the distribution effected from a location well outside of the suburbs. The philosophy of such a location is, of course, that of attempting to encompass within the range of the station a widespread area and of so including the city within the area as to effect a more uniform distribution over it than is possible when transmitting from a location within the city. A theoretical study was made of the distribution to be effected from one such location in the general vicinity of Boonton, New Jersey, using attenuations obtained in the other surveys. Such a location would be somewhat similar to that of WJZ at Bound Brook, although the distance from Boonton to New York is less. The figures derived upon the basis of a 50 kw. broadcasting station are as follows:

Field Strength	Percentage of Receiving Sets in Metropolitan Area
Below 10,000 μv./m.	10 per cent
Between 10,000 and 50,000 μv./m.	79 per cent
More than 50,000 μv./m.	11 per cent

These figures show a good concentration in the most desirable field strength values. They should be discounted somewhat because they

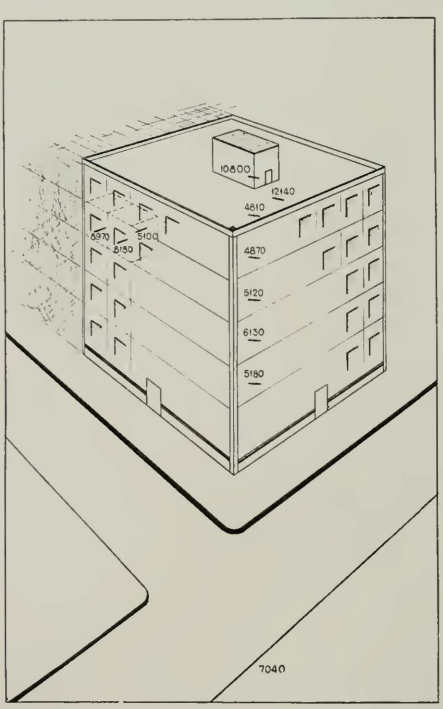


Fig. 10—The effect of non-steel apartment house building in shielding radio reception within it

are based upon symmetrical distribution and do not include the effect of irregularities, which an actual survey probably would reveal.

Receiving in Apartment Houses

The surveys described above disclose the field strength distribution as measured generally in the streets and open places. It does not disclose the details of field distribution in the immediate vicinity of a

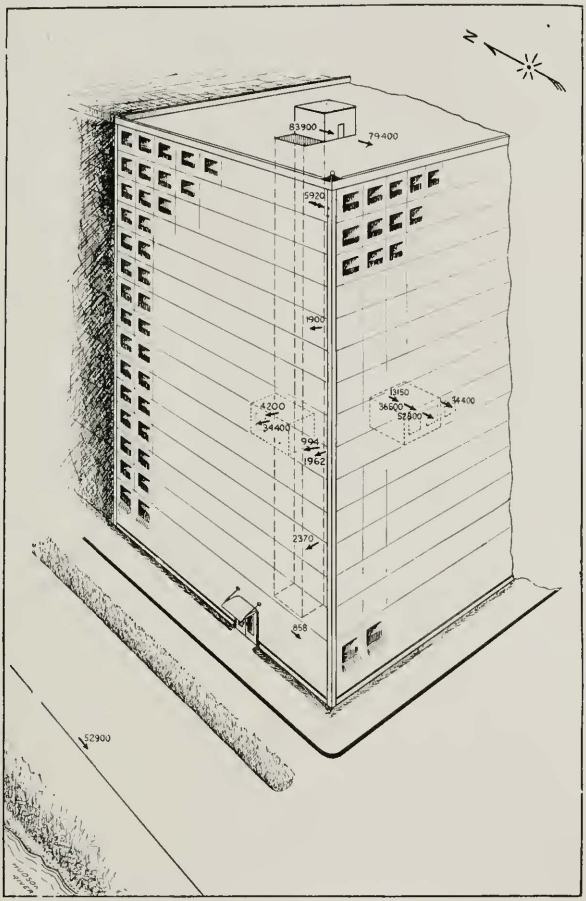


Fig. 11—The effect of steel structure apartment house building in shielding radio reception within it

receiver. Perhaps the most difficult situation is that of the large apartment house, particularly where it is desired to receive by means of an indoor antenna. Two effects are encountered: First, the reduction in signal strength by virtue of the shielding effect of the building;

and second, the existence of a relatively high noise level caused mainly by radio-frequency interference from electrical systems within the building.

The results of a few observations upon signal strength reduction within two buildings are presented in Figs. 10 and 11. Fig. 10, for a non-steel building, shows the field to be roughly halved. In the case of the steel structure building depicted in Fig. 11, the interior field is found to be reduced to as low as a few per cent of that outside the building. For outside rooms, the field strength near the window was found to be about eight times that further in the room. Such severe shielding effects obviously call for picking up the wave energy outside the building and conducting it to the receiving sets by wire circuits, preferably by shielded circuits, in order to protect against local interference.

Multi-Station Operation

The discussion given above has been directed chiefly to the relations which might be called internal to a single-channel radio broadcast system. Actually, of course, broadcasting involves the use of the common transmitting medium for a number of channels. This brings with it the problem of frequency selectivity and raises the question of the capabilities of the various types of radio receiving circuits.

In order to throw some light upon this important factor, measurements have been made upon a sample or two of each of a number of different types of radio receiving circuits. The measurements were made in the laboratory,[7] simulating as closely as possible the conditions under which the receiving sets would be used. The curves of Fig. 12 show the reduction which is to be expected in the detected audio-frequency current, were the receiving set tuned to a transmitting station on 900 kc., and the transmitting station then shifted in frequency by the amounts given along the abscissa. In this curve the reduction in current is indicated both as a ratio and in TU, which is a convenient way of indicating power ratios. The relation between TU and current ratio with a given impedance is indicated in the figure. Thus, for a carrier 40 kc. off from the one to which the set is tuned, the single-circuit, non-regenerative type of receiver

[7] The method consists in establishing a small laboratory transmitter and modulating it with a single-frequency tone. The receiving set is tuned to the modulated carrier signal as in practice. The gain or sensitivity of the receiver and its coupling with the transmitter are adjusted to produce normal load upon the detector tube. With the receiving set left at this adjustment the frequency of the radio transmitter is shifted each side of the original single frequency in 10 kc. steps throughout a range of 50 to 100 kc. For each of the offside frequencies the reduction caused in the detector output current is measured, this being an indication of the receiving set selectivity.

showed a cutoff of only 20 TU, corresponding to an audio-frequency current reduction to 0.1 that of the value at resonance. The curves will

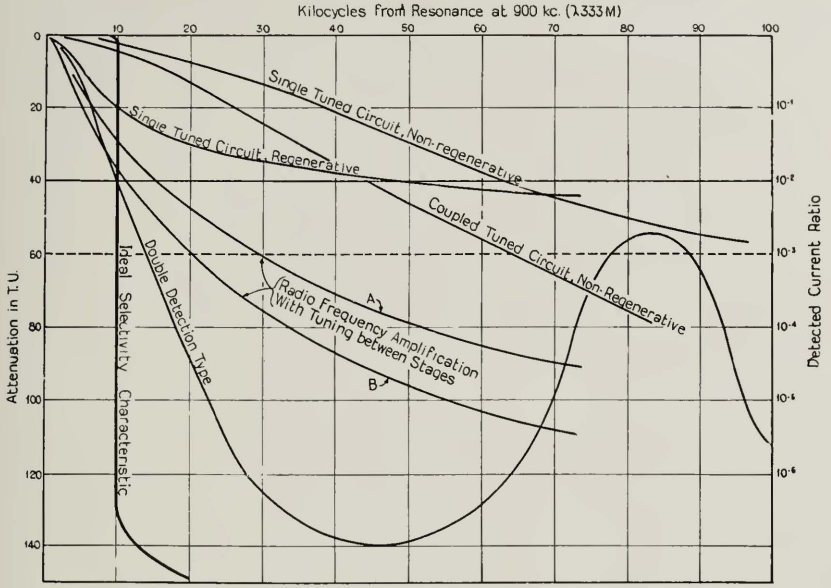


Fig. 12—Receiving set selectivity characteristics as measured from samples of receivers having different types of selective circuits

be seen to group themselves more or less into three classes in the order of their selectivity merit as follows:

1. The single-tuned circuit (non-regenerative and regenerative), and the combination of two tuned circuits coupled together.

2. Circuits employing radio-frequency amplification with tuned circuits between stages.

3. The double-detection or superheterodyne type of circuit.

The curve for the double-detection type of circuit shows a "comeback" which represents the familiar double-tuning effect. (Incidentally, the admittance of this particular set, which was not a commercial set, needs to be reduced by the use of more selectivity at the radio frequency.)

For comparison purposes there has been added to the figure the curve marked "ideal selectivity characteristic," in accordance with which the receiving set would pass without attenuation all frequencies up to 5000 or 10,000 cycles and would cut off abruptly all frequencies without this band. Attention is first called to the fact that the various circuits attenuate *within* the desired transmission band of

five or ten kilocycles. This means the higher frequency components of the side band will be reduced by the amounts indicated (after detection) with corresponding distortions of the reproduction. The distortion will be seen to be greater for the more highly selective sets. This follows from the nature of sharply tuned circuits. Selective circuits, capable of approximating the filter type of characteristic, are to be desired.

In comparing these selectivity characteristics, it is necessary to have in mind the amount of differentiation between the desired and the undesired signal which is necessary for the avoidance of interference. Each of the signals may be considered as fluctuating during the rendition of the program over a considerable range of volume which centers about some average value. The amount of differentiation required between the average values obviously depends upon the range of the fluctuations involved and upon the standard which is assumed with respect to freedom from interference. Experience with loud speaker reproduction indicates that ordinarily a level of the average of the undesired signal 40 TU lower than that of the desired signal, while not giving noticeable interference at times when the desired signal is strong, does permit the undesired signal to "show through" during times when the program rendition is weak. Reducing the undesired signal to 60 TU below the desired signal prevents this interference for the volume ranges which are now commonly transmitted. If the future art brings with it the requirement of following greater swings of volume, a further reduction in the undesired signal may be necessary. The value of 60 TU has been dotted in across the chart of Fig. 12, in order to show readily the frequency separation at which the different selective circuits give this attenuation of the undesired signal. This is upon the basis that the field strengths of the two signals are equal. Inequalities in field strength require that the 60 TU value be increased or decreased by the amount of the inequality as measured in TU.

The frequency interval which has been recommended by the National Radio Conferences for stations in the same zone is 50 kc. It is evident from the curves that sets equipped with the simpler types of tuned circuits will be subject to some interference between stations thus separated even if the receiver is so favorably situated as to receive equal field strengths from the desired and undesired stations. The selectivity of the other types of receiving circuits is seen to be sufficient to avoid interference under these conditions and allow some margin for overcoming inequalities between the fields. Such inequality becomes great where the attempt is made to receive distant

stations through the effect of local stations. Assume, for example, that the listener receives 50,000 microvolts from a local station and 500 microvolts from a distant station to which he desires to listen. There exists a 100 to 1 or 80 TU disadvantage to be overcome. When added to the 60 TU needed for crosstalk clearance, the total selectivity requirement, as measured in terms of detected audio current, becomes 140 TU, or a current reduction of the order of 10,000,000 to 1. The need for a high degree of selectivity is therefore apparent. The impracticability of receiving distant stations removed in frequency from local stations by any such narrow margin as 10 kc. is also obvious.

The effect of receiving set selectivity in increasing the area over which a station may be received without interference from a second station is illustrated in Fig. 13. The two stations are assumed to be of equal powers so that they deliver equal field strengths to receiving stations along a line midway between them. Receiving sets so located are required to have an amount of selectivity called for by the crosstalk margin itself, say 60 TU. On the desired-station side of this line the selectivity may be less; this is the region where poorly selective receivers can be employed. On the undesired-station side of the center line the selectivity requirements are greater. The noninterference area is pushed up closer and closer to the undesired station as the receiving set selectivity is improved, as is indicated by areas *A* and *B* of the figure. For example, assume that the selectivity of the receiving set is such as to give a 100 TU cutoff of an undesired station, offset by 50 kilocycles. Sixty TU of this would be required were the two signals of equal strength, so that 40 TU measures the difference by which the undesired signal may be greater than the desired one. The increased area of reception made possible by this additional 40 TU is indicated by that portion of the lower figure which is to the right of the center line and outside of the area *A*. Within the area *A* interference would be suffered. This interference area may be diminished by the use of still greater selectivity. The addition of another 20 TU of selectivity (again as measured in terms of detected audio-frequency current) would reduce the interference area to that within the small area at *B*. The extent to which the selectivity requirement of the receiving set is determined by its location, therefore, is apparent. The conditions which obtain in multistation areas, such as New York City and Chicago, obviously call for a general use of high selectivity sets.

In locating a new transmitting station it should be possible from a knowledge of the relative field strengths of other stations in the vicinity to predict approximately what the interference area will be

for the different types of receiving sets. In this connection there should be recognized the advantage from the interference standpoint which exists in grouping together the broadcast transmitting stations as far as possible in one location, and in equalizing their powers. Such

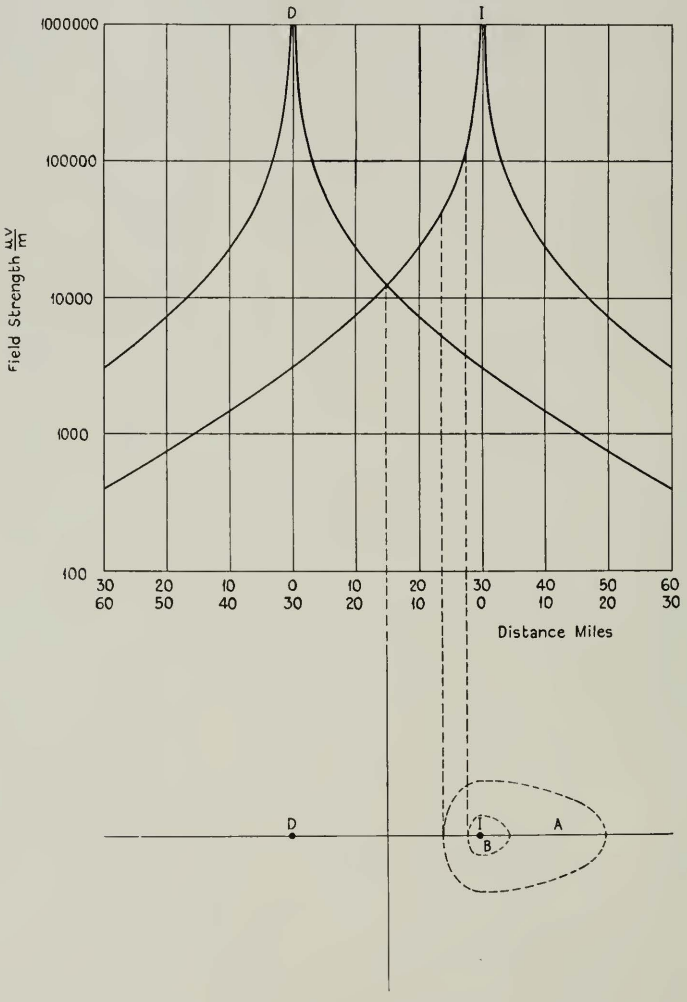


Fig. 13—Showing the greater area over which the more highly selective receiving sets may receive a desired station D and exclude an interfering station I

grouping and equalizing would enable the receivers to obtain substantially equal fields from all of the stations and would minimize the selectivity which they are required to possess. While it is im-

practicable to accomplish this result completely, it is hoped that a better understanding of the interference problem as here outlined and of the mutual advantage to be gained in reducing interference will lead naturally to a better coordination of radio broadcast stations.

Acknowledgment

The data and analyses presented in this paper are the result of the cooperative effort of a number of engineers in both the American Telephone and Telegraph Company and the Bell Telephone Laboratories, Inc. In assembling and presenting the material the writer is acting merely as the spokesman for these development groups. He wishes to acknowledge his indebtedness to his colleagues, particularly to those who have directly participated in the survey work described and assisted in the preparation of the paper, namely, to Messrs. D. K. Martin, R. K. Potter, G. D. Gillett and H. B. Coxhead, and to Messrs. S. E. Anderson and O. O. Ceccarini of the Bell Telephone Laboratories to whom is due the measurement work upon the radio receiving sets.

A Shielded Bridge for Inductive Impedance Measurements at Speech and Carrier Frequencies [1]

By W. J. SHACKELTON

Synopsis: A shielded, a-c., inductance bridge adapted to the measurement of inductive impedances at frequencies up to 50,000 cycles is described. The bridge comprises a balancing unit and associated standards of inductance and resistance. The balancing unit has resistance ratio arms specially constructed to meet the requirements imposed by the above frequency range. The reference standard makes use of inductance coils of a new type, their cores being of magnetic instead of non-magnetic material as is usually the case. The use of such cores results in coils that are smaller and hence better adapted to assembly in a multiple shielded standard.

The bridge is completely shielded so as to eliminate, to a high degree, errors due to parasitic capacitance currents. The shielding is also arranged so as to permit the correct measurement of either "grounded" or "balanced-to-ground" impedances. A series of diagrams is shown for the purpose of indicating the function of each part of the shielding system.

Equations expressing the errors resulting from any small residual capacitance unbalances in the resultant bridge network are given and calculations made of the balances required for the desired degree of measurement precision. Test data are presented illustrating a method of experimentally checking the residual shunt and series balances from which it is concluded that the bridge is capable of comparing two equal inductive impedances of large phase angle with an accuracy at the maximum frequency of 0.02 per cent for inductance and 1.0 per cent for resistance.

Introduction

THE limitations of the ordinary unshielded bridge network as a means of making precise a-c. measurements at speech frequencies were early recognized by telephone engineers. The solution of cross-talk problems arising in connection with the use of cable circuits was found to require an exact knowledge of the capacitive balances existing between such circuits at speech frequencies. For the ready and accurate determination of the capacitances defining these balances, together with their associated conductance values, G. A. Campbell devised the "shielded balance." [2] This is a bridge network having its parts individually and collectively shielded so as to define exactly the mutual electrostatic reaction of each with respect to all other parts of the electrical system affecting the balance condition.

As a means of more completely treating the cross-talk problems of cable circuits, Campbell conceived also the very valuable idea of "direct capacity" as distinguished from the "ground" and "mutual capacities" in use up to that time.[3] The shielded balance was found

[1] Presented at the New York Regional Meeting of the A. I. E. E., New York, N. Y., Nov. 11–12, 1926.

[2] G. A. Campbell: "The Shielded Balance," *Electrical World* and *Engineer*, April 2, 1904, p. 647.

[3] G. A. Campbell: "Measurement of Direct Capacities," Bell System Technical Journal, July, 1922, p. 18.

to be especially adapted to the precise measurement of direct capacities employing the substitution method devised by E. H. Colpitts.[4] Shortly thereafter, with the advent of loading for telephone lines, the same principles of shielding were extended to apply to bridge networks specially arranged for the measurement of the speech-frequency inductance and effective resistance of loading coils. As the successful commercial application of loading required the manufacture of these coils in large numbers to precise requirements, it was quite essential that testing means be available permitting a relatively unskilled tester to determine quickly whether the proper adjustment of the coils had been made. For this purpose the shielded balance has proved to be extremely valuable. More recently the employment of frequencies up to 50,000 cycles for carrier telephone and telegraph purposes has led to the need for correspondingly precise measurements at these higher frequencies. In this field the advantages of the shielded bridge are so great as to make it almost indispensable.

While the fundamental principles of the shielded balance are essentially the same for all impedance measurements, the practical application of shielding to any concrete bridge problem may vary according to the kind and range of impedances to be tested, the frequency range to be covered, and the precision required. It also presents special problems in the design and construction of several of the circuit elements. This paper describes a particular form of shielded bridge which has been developed to meet the conditions commonly encountered in the measurement of inductance at speech and carrier frequencies. The facts leading to the detailed construction are discussed and some experimental data given to illustrate the performance of the bridge.

General Features

A simple schematic diagram of the bridge circuit is shown in Fig. 1. To avoid confusion, no shielding is shown in this diagram. As will be noted, there are provided two equal non-inductive resistance ratio arms, an adjustable standard of self-inductance, an adjustable resistance standard, a thermocouple milliammeter, two transformers and two adjustable air condensers. Physically, this apparatus is grouped into three separate units, one comprising the standards of inductance, one the resistance standard, and the third, the remaining parts of the circuit. The last assembly constitutes what may be considered the balance element of the system, by means of which the unknown and standard impedances are compared. Figs. 2 and 3 show the arrange-

[4] See Note 3.

ment of the parts in this unit. Fig. 4 illustrates the appearance of the standard inductance unit and Fig. 5 shows how the units are associated when a test is being made. The thermocouple milliammeter indicates the total effective test current applied to the bridge and forms a

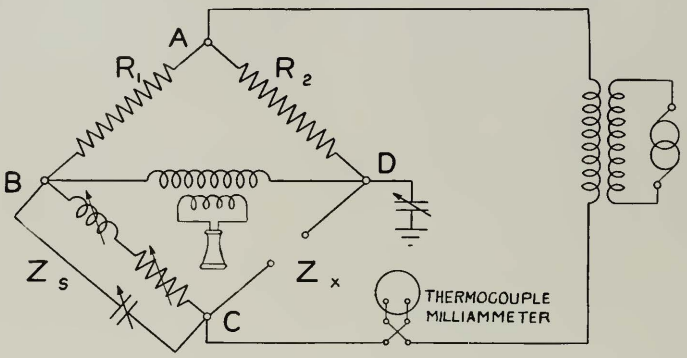


Fig. 1—Schematic diagram of bridge circuit

means of determining when this current has been adjusted to the desired value.

In operation, the air condensers are first adjusted to produce an initial or zero balance of the residual electrostatic capacitances of the apparatus. Aside from the initial balancing, the operation of the bridge follows the usual practise; that is, the standards of inductance

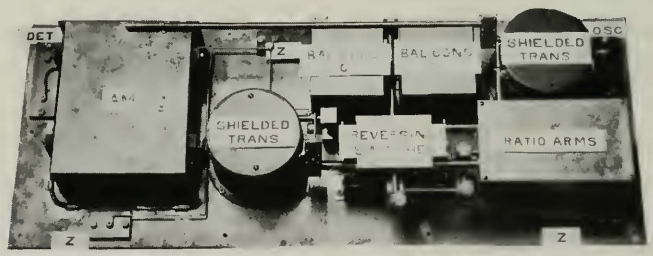


Fig. 2—Balance element of shielded bridge. Rear view of panel removed from case

and resistance are alternately adjusted until the balance detector indicates a condition of zero potential difference at every instant between the bridge points to which it is connected. The inductance and resistance values as indicated in the standard arm are then equal (within the precision limits of the bridge) to the corresponding constants of the unknown impedance.

Purpose of Shielding

The principal difficulties in attaining a satisfactory degree of precision in inductance measurements at relatively high frequencies by means of unshielded bridges are those due to the presence of residual or stray admittances existing between the bridge parts or from them to ground. All these parts have quite appreciable surface dimensions



Fig. 3—Balance element of shielded bridge. Front view of panel and case

and when exposed at the usual separations to each other or to ground, have corresponding direct and grounded admittances. Leads to the source of testing current and to the balance detector also introduce rather large admittances. In a bridge intended for rapid operation, the parts subject to manipulation must be arranged compactly and



Fig. 4—Inductance standard

conveniently to the operator. This makes it impracticable to isolate them sufficiently to make the admittance values between these parts

and between them and ground (the operator being considered to be at ground potential) negligibly small.

To make the matter more concrete, there is shown in Fig. 6 a



Fig. 5—Shielded bridge connected to vacuum tube oscillator and heterodyne detector

schematic diagram with possible positions of some of the more important of these admittances indicated as at C_1, C_2, etc. (With some exceptions, the capacitance components of these stray admittances substantially determine their full effect. In the diagrams and discussion, therefore, the conductance component will be neglected except where its effect is significantly large.) The capacitances between the two ratio arm coils, R_1 and R_2, and from each to ground, are shown as being uniformly distributed along the length of the coils symmetrically with respect to each other. If this symmetry is perfect these capacitances do not affect the bridge balance. In practise, however, they will only be approximately so, with the result that the two arms will be somewhat unbalanced to alternating currents, the effect of the unbalance increasing with the frequency. While the ratio arm capacitances can be made fairly small, others such as those indicated at C_1, C_2, C_3, and C_4 will commonly be much larger and hence of greater effect. Capacitances C_1 and C_2 are frequently comparatively large

due to the use of long distributing wires, encased in grounded conduit, for supplying the testing current. C_3 may consist chiefly of the ground capacitance of the outer layer of the detector coil winding and C_4 that of dead-end coils of the reference standard, Z_S.

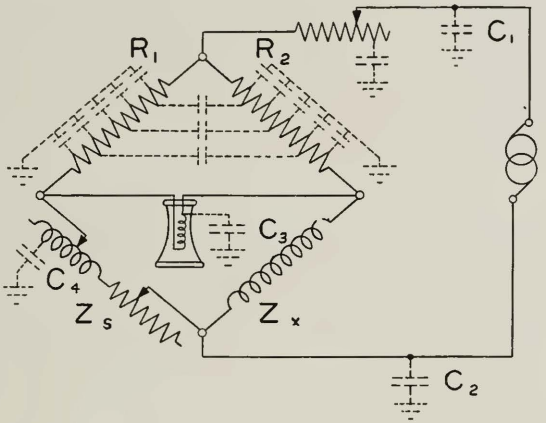


Fig. 6—Bridge circuit with stray admittances

Some of the currents flowing along the paths provided by these capacitances will complete their circuits external to the bridge network proper and will not affect the balance; for example, that through capacitances C_1 and C_2 in series. Other currents, however, will flow unsymmetrically through parts of the bridge circuit; for instance, that through C_1 and C_3 in series and the arm Z_X; also, that through C_2 and C_4 in series, returning through the ratio arm R_1. These latter currents and others of the same sort affect the potential distribution of the bridge and hence the values of the impedances required for balance. Certain of these capacitance currents in the bridge network tend to neutralize or balance the effects of others; for example, that through the arm Z_X due to the series action of capacitances C_1 and C_3 has a balancing effect with respect to that through C_1 and C_4 and the arm Z_S and would be without reaction on the bridge balance if capacitances C_3 and C_4 were exactly symmetrical with respect to the two detector terminals. Such balancing, however, is accidental in nature, seldom satisfactorily complete and, in part, not constant. Even were it made approximately complete for a particular arrangement, the substitution of another detector or the use of another source of testing current would probably destroy the balance. Variable effects would always be present; for example, those due to the changing position of the operator relative to the parts of the circuit or the

effects of parallel loads on the supply generator. The distribution and value of the ratio arm ground capacitances described above are functions of the bridge surroundings; hence they are also subject to change if the bridge is moved from place to place. In the bridge being described, however, the shielding used affords a means of definitely fixing and controlling the various inter-circuit capacitances. Consequently, such variations cannot take place, balances between the resultant capacitance currents can be made as desired, and the bridge measurements are satisfactorily precise.

Shielding System Used

It is felt that the merits of the particular shielding system adopted for this bridge can best be brought out by showing, step by step, the reasons for using each of its elements.

The first step is to simplify, for further treatment, the initial residual capacitance network of the unshielded circuit. This is done by providing individual shields for each part of the circuit that it is desired to have function as an independent unit. Such shields can be connected to one of the terminals of the part enclosed and thus there is substituted, from the standpoint of terminal-to-terminal characteristics, a definite and invariable condition in place of that which was previously a function of the relation of the part to its surroundings. For example, as shown in Fig. 7, shields would be placed

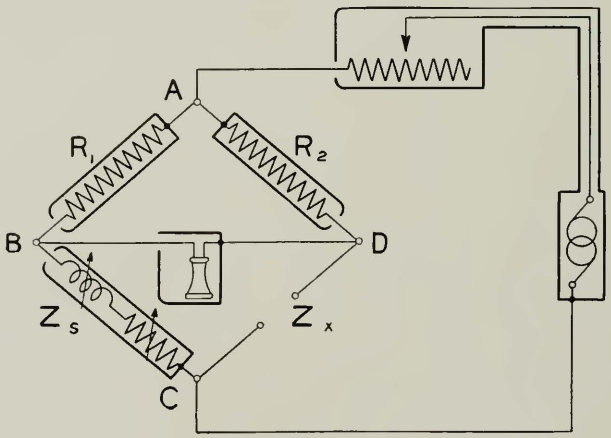


Fig. 7—Bridge circuit with local shields

around the resistance coils forming the ratio arms R_1 and R_2 and connected to the junction point A of the system, one enclosing the elements of the standard impedance Z_S and another around the source

of testing current and connected at *C*; likewise, one around the detector is connected at *D*. It will readily be seen that these shields localize the effects of the various capacitance currents. Those circulating within the shields have, of course, no effect exterior to the shields, while those flowing between the various shields directly or by way of ground enter and leave the bridge system at definite points. By themselves, these shields do little good but they are necessary in order to make the next step, the balancing of the capacitances, practicable.

Generally it will not be found convenient to shield the current supply apparatus, especially if this is a power-driven generator. Also, to promote greater flexibility in respect to testing with a wide range of frequencies, it will often be desirable to substitute one source of current for another and likewise one detector for another. The shielding of this apparatus should therefore be reduced to a minimum. This is readily effected by making both the supply and detector branches of the bridge one of the windings of a transformer. This winding can be electrostatically shielded without affecting its transformer action and then any desirable source of current supply or any type of detector can be magnetically coupled with it.[5] Introducing this change the circuit becomes as shown in Fig. 8. The

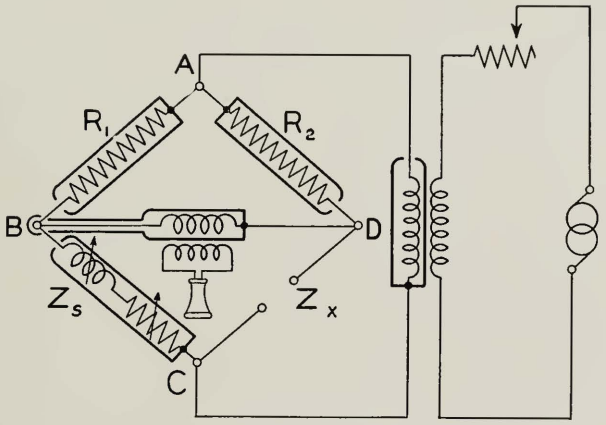


Fig. 8—Bridge circuit with shielded transformers and local shields

capacitances of the various shields to ground being still variable, the next step to correct this condition would simply be to add a ground shield around each. At this point, however, it becomes necessary to consider the ground admittance relations of the impedances to be tested.

[5] U. S. Patent No. 792248, June 13, 1905.

In general, the unknown impedance will have capacitances to ground and the effect of these will be properly included in the measurement only when certain conditions as determined by the nature of the apparatus are fulfilled. From this standpoint the impedances usually encountered are of three general classes: (1) Those having ground admittances negligibly small in comparison with the direct terminal-to-terminal admittance; (2) those having appreciably large admittances to ground approximately balanced with respect to the two test terminals; (3) those having one terminal directly grounded, the other having an appreciably large ground admittance.

In measuring apparatus of the first type it is evident that since in connecting it to the bridge circuit no additional ground admittances are introduced, the balance between those previously existing can be made without reference to the test impedance. The connection of an impedance of either of the other types will, however, introduce additional ground admittances into the bridge system, which, unless precautions have been taken, may cause the result to be something other than that which is wanted. In general, the desired test is that which gives the effective impedance applying to the apparatus as it is used. In the case of impedances having balanced admittances to ground, this is the effective value of the direct, terminal-to-terminal impedance as modified by the effect of the two ground admittances acting simply in series with each other. This condition is obtained when equal currents flow in each of these admittances, or, what is equivalent, when the electrical potentials of the terminals are balanced with respect to ground potential. To obtain this condition, when the impedance is being tested, the bridge terminals to which it is connected must likewise be balanced with respect to ground potential; that is, ground potential must be at the midpoint of the unknown impedance arm. If the only admittances to ground of the bridge system are those of the junction points (as is the case in Fig. 8), the potentials of these points with respect to ground are entirely determined by these admittances. To make any two points, such as the terminals of the unknown arm, have equal potentials to ground, it is sufficient to concentrate all of the ground admittances to these or other equipotential points and then balance the admittances from each. Referring to Fig. 9, if the testing current is applied at the points A and C, this condition is realized as shown by concentrating all ground capacitances at junction points B, C and D, and making the sum of the capacitances of junction points B and D equal to that of junction point C. This follows from the fact that when the bridge is balanced the junctions B and D are equipotential points. The mid-point of arm CD is now

at ground potential. If, however, the testing current is applied at the points B and D, the equipotential points are the junctions A and C, the sum of whose ground admittances would then be made equal

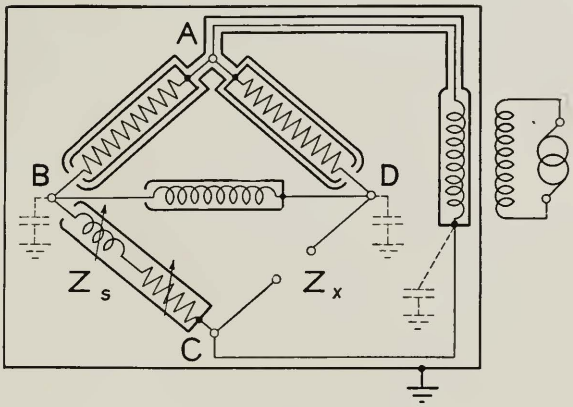


Fig. 9—Shielded bridge circuit showing location of ground admittances

to that of D and the arm CD again balanced with respect to ground potential. In this case there must be no ground admittance from junction B. To permit of testing under both conditions, point A and all connected conductors are protected with a shield which is then connected to the point C. Point B is likewise enclosed by a shield connected to point D. These two main shields then represent the junction points C and D of the bridge and are fixed with respect to capacitance to ground by a ground shield which may be common to the two.

There now exist, external to the local shields, direct capacitances only between points A and C and between B and D (which do not affect the bridge balance), and from points C and D to ground. These latter do, of course, affect the balance. Two courses are open. Their effective resultant value shunting the arm CD can be determined and allowed for by calculation. Such calculations would involve a considerable amount of labor, however, and can be avoided very simply by providing in the opposite arm an exactly equal shunt capacitance. To permit adjusting the ground capacitances of points C and D, an adjustable condenser is connected to ground from the point having the lower value. With the apparatus connected as shown this is usually point D. The shielded system then becomes as shown in Fig. 10.

When impedances, which in actual service are grounded at one terminal, are to be tested, the matter is much simpler. Then it is

necessary only to definitely ground one of the bridge terminals to which the impedance is connected and establish the proper initial capacitance balance of the bridge for this condition. This is readily done by grounding junction point *C* and adjusting the capacitance from *B* to *C* to equal the ground capacitance of *D*. The shielding system may remain the same as in Fig. 10.

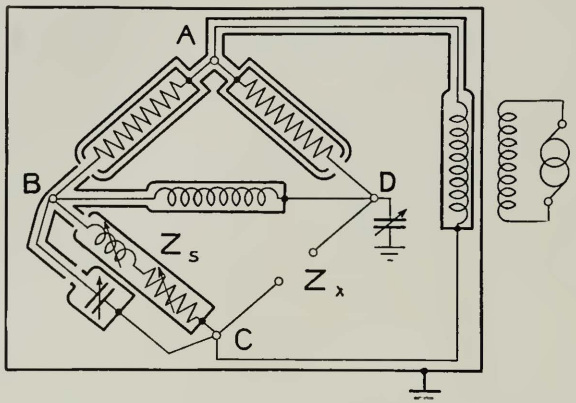


Fig. 10—Shielded bridge circuit with balancing condensers

In the case of the bridge being described it was desired to have a means of verifying by reversal the degree of balance of the ratio arms and also that of the impedance arms. The bridge is therefore equipped with reversing switches for this purpose. Due to the appreciable effect produced by a relatively small capacitance unbalance arising from factors present only when the arms are in circuit, it is quite important to be able to do this when a high degree of accuracy is desired. To effect the proper reversal, however, certain conditions must be definitely maintained. In reversing the impedance arms none of the inherent bridge admittances should be disturbed; that is, only the unknown impedance and the standard as read should be transferred. On the other hand, in reversing the ratio arms not only should the resistance element of these arms be transferred but also all associated shunt admittances. Moreover, in transferring these admittances they must be absolutely unchanged. A further requirement is that the ratio arm reversal must not occasion the shifting of any capacitances shunting the impedance arms. To accomplish these objects a suitable arrangement of shielded switches was worked out and added to the circuit of Fig. 10, the result being as shown in Fig. 11. In this arrangement all capacitances between the various parts of the switches which are subject to change due to physical movement of

the switch parts are either short-circuited or connected across opposite bridge points and hence do not affect the bridge balance. The small capacitance C_R between the switch shield and that of the ratio coil R_2 shunts this coil and is not carried with it on reversal. For this reason a corresponding capacitance C_R', shunting the coil R_1, is pro-

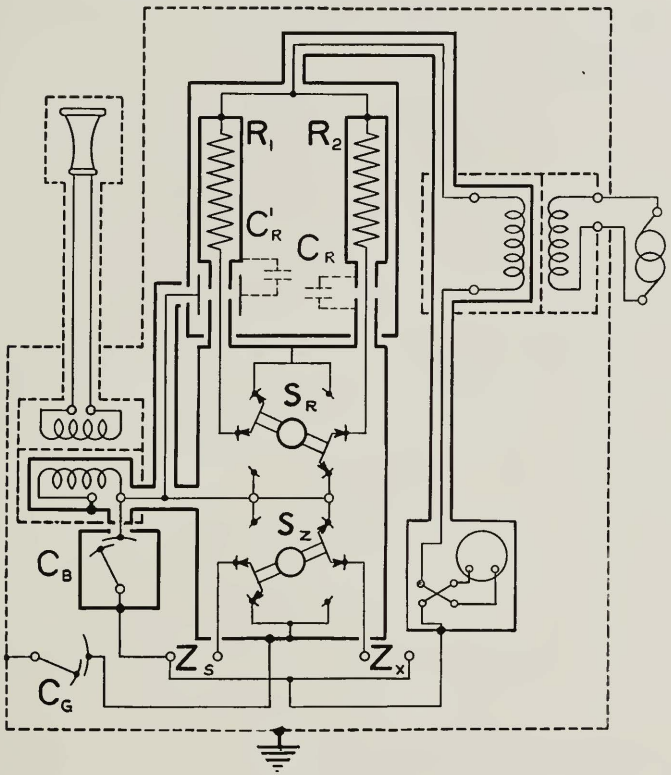


Fig. 11—Complete circuit diagram of balance unit with shielding

vided and connected to the opposite point of the switch. This is adjusted by test to equal the value of C_R. The diagram of Fig. 11 represents completely the circuit and shielding used for the balance unit of the bridge.

While, from the standpoint of the bridge balance alone, the parts comprising the standard impedance can be shielded with a local and ground shield as shown in Fig. 9, unless the standard has a very limited range, the resulting calibration is exceedingly laborious to make and use. To reduce calibration difficulties, additional shields can be used; this, of course, increasing the cost of construction. In arriving at the

proper compromise between these conflicting factors the size and impedance value of the part to be shielded must be considered. This question will therefore be taken up in more detail in the following section.

Construction

The circuit and shielding features discussed so far are of general application to impedance measurements without restriction as to the particular range of values to be tested or frequencies to be used. The physical construction is, however, dependent upon these factors. As initially stated, the bridge is intended for the measurement of audio and carrier frequency inductances. By this is meant all apparatus having reactance values nearly equal to the respective impedance values. For the purpose of the present discussion, such inductances will be more exactly defined as those having ratios of reactance to resistance of not less than 10 (minimum phase angle of 84 deg., 20 min.). The difference between the reactance and the impedance of any such inductance does not exceed ½ per cent. The impedance values range from about 100 to 10,000 ohms and testing frequencies from 500 to 50,000 cycles.

On the basis of these conditions, the following construction was developed and is used for this bridge.

Ratio Arms. It is desirable from the standpoint of sensitivity of balance to have the ratio arm impedances of approximately the same value as those of the other two arms. Considering the range of impedances to be covered and giving due weight to the values which are of most importance in telephone circuits, a ratio arm resistance of 1000 ohms was selected. The problem then was to construct two 1000-ohm resistances, balanced both as to effective resistance and effective inductance for a frequency range from 500 to 50,000 cycles when subjected to the usual temperature and humidity variations.

Curtis and Grover have discussed the factors affecting the characteristics of a-c. resistances and have suggested forms suitable for general use at frequencies up to 3000 cycles.[6] A 1000-ohm resistance, constructed according to their specifications, is made by winding with a 1/10-mm. diameter, double-silk-covered manganin resistance wire, five 200-ohm bifilar sections on a 1-in. spool of insulating material. These sections are spaced about three mm. apart on the spool and are connected in series to form the 1000-ohm coil. Such a coil, when shellacked, baked and coated with paraffin, was found to be substantially constant in resistance and to have constant phase-angle

[6] H. L. Curtis and F. W. Grover: "Resistance Coils for Alternating Current Work," *Bulletin of the Bureau of Standards*, Vol. 8, No. 3.

effects equivalent to shunting capacitances of the order of 10 to 15 mmf. for all frequencies up to 3000 cycles. Since individual coils made according to this method may differ in their effective capacitances by as much as five mmf., some adjustment of these capacitances (as well as of the resistance) is required in order to make them suitable for use as the required ratio arms. Assuming that this is done by adding to the coil having the lower value a small capacitance of suitable constancy, it may be concluded that two coils so balanced would be suitable for use at frequencies up to 3000 cycles.

In arriving at the requirements for the more extended frequency range of this bridge, the necessary phase-angle balance was first considered. Designating by L_X and R_X the inductance and effective resistance of the impedance being tested, and by L_S and R_S, the corresponding components of the standard impedance required to balance it in a bridge circuit having ratio arms of exactly equal resistances R but shunted by slightly different capacitances, C_1 and C_2, and assuming that the quantities are such that $\omega^2R^2C_1^2$ and $\omega^2R^2C_2^2$ are small in comparison with unity, the equation for balance is

$$(R_X + jwL_X)(R - jwC_1R^2) = (R_S + jwL_S)(R - jwC_2R^2),$$

which reduces to

$$R_X = R_S + w^2R(C_2L_S - C_1L_X) \tag{1}$$

and

$$L_X = L_S - R(C_2R_S - C_1R_X). \tag{2}$$

Neglecting second order effects, these can be written

$$R_X = R_S + w^2RL_X(C_2 - C_1) \tag{3}$$

and

$$L_X = L_S - RR_X(C_2 - C_1). \tag{4}$$

If the readings R_S and L_S are taken as the values of the unknown resistance and inductance, respectively, it is evident that errors as given by the last terms of these equations will be present. The percentage errors in the two cases are as follows:

$$\Delta R_X\,(\%) = 100\omega^2R(C_2 - C_1)\frac{L_X}{R_X}$$

$$= 100\omega R(C_2 - C_1)\tan\theta, \tag{5}$$

$$\Delta L_X\,(\%) = 100R(C_2 - C_1)\frac{R_X}{L_X}. \tag{6}$$

For a given capacitance unbalance of the ratio arms, it is seen that the error in inductance is inversely proportional to the time constant L/R of the impedance arm and is independent of the frequency, while the error in resistance is proportional to the frequency and to the ratio of reactance to resistance, that is, to the tangent of the phase angle. The inductance error is, therefore, maximum for the minimum time constant apparatus to be tested. Within the range previously mentioned this occurs when an impedance having the minimum reactance to resistance ratio of 10 is being measured at the minimum frequency of 500 cycles. Under this condition R/L has a value of $(2\pi \times 500)/10$ or approximately 300. The corresponding percentage error in inductance per micro-microfarad of capacitance unbalance is then $300 \times 1000 \times 10^{-10} = 3 \times 10^{-5}$ or 0.00003 per cent. Evidently a very considerable unbalance can be tolerated. In the case of the resistance component, the error is maximum when an unknown impedance having the maximum reactance to resistance ratio is being tested at the maximum frequency. A reactance to resistance ratio of 300 is very rarely exceeded. For this value, the error per micro-microfarad unbalance at a frequency of 50,000 cycles amounts to about 9.5 per cent. Hence, to limit the error from this source to the order of 1 per cent requires a balance of about 0.1 micro-microfarad. It will be appreciated that this is an extremely close balance, the maintenance of which, under the different conditions of temperature and humidity to which the bridge may be subjected, requires careful consideration of the effects of these factors.

The effective phase-angle balance, though discussed above in terms of capacitance only, is, of course, the resultant of the inherent residual inductances and capacitances of the coil windings plus the additional capacitance effects due to the coil shields. The component due to residual magnetic induction is not appreciably affected by temperature or frequency changes. The capacitance component of the winding, however, tends to vary with temperature in accordance with the temperature coefficient of capacitance of the dielectric used for insulating the wire and with frequency to the extent that the capacitance is affected by absorption.

It is common practise to employ silk-insulated wire treated with varnish or wax for purposes of protection against moisture in such coils. In order to obtain data covering the temperature and absorption effects and also the phase-angle characteristics of silk insulation, both untreated and when treated with a number of the more common materials, various samples were constructed and tested as indicated in Table I.

TABLE I

CAPACITANCE AND PHASE-ANGLE TESTS BETWEEN TWO DOUBLE-SILK-COVERED NO. 38 A. W. G. WIRES, EACH 88 IN. (224 CM.) LONG, WOUND IN BIFILAR FASHION ON A GLASS TUBE ONE IN. (2.54 CM.) IN DIAMETER AND THEN TREATED WITH VARIOUS MATERIALS. SAMPLES DRIED BEFORE TESTING.

Test sample number	Material used for treatment	Capacitance, micro-microfarads				Phase-angle tangent			
		Temp.— 20 deg. cent.		Temp.— 45 deg. cent.		Temp.— 20 deg. cent.		Temp.— 45 deg. cent.	
		1000 ∞	50,000 ∞	1000 ∞	50,000 ∞	1000 ∞	50,000 ∞	1000 ∞	50,000 ∞
1	None	193.3	188.2	188.7	184.1	0.0083	0.013	0.0076	0.011
2	Paraffin	251.1	243.2	232.1	244.3	0.0095	0.014	0.010	0.011
3	Collodion	238.9	228.0	231.4	222.5	0.016	0.021	0.015	0.018
4	Beeswax compound	258.2	248.7	273.2	265.8	0.013	0.014	0.010	0.013
5	Pyralin	253.1	241.1	258.5	246.6	0.016	0.021	0.018	0.022
6	Insulating varnish	346.4	320.1	361.5	337.0	0.027	0.036	0.030	0.033
7	Shellac	296.9	284.0	314.3	302.4	0.012	0.021	0.015	0.020

From the standpoint of percentage capacitance change (reckoning from the minimum temperature and frequency conditions as being those at which initial adjustments would be made), untreated and paraffin-treated silk insulation were found to be appreciably superior to any of the other materials. The change due to absorption effect (about three per cent for these two) was considered the more important, as normally variations in temperature would not be very large. As would be expected, the untreated silk had the lowest phase-angle effect and also the smallest capacitance. Assuming that a method of excluding moisture could be devised, it was concluded that an untreated silk-insulated winding would be the best to use, although the paraffin treatment was also considered promising. Discounting the fact that the two ratio coils would change in the same direction though not necessarily by the same amount, it was decided that a satisfactory factor of safety would be provided if it were assumed that the coils might become unbalanced by one half the observed change in one coil; that is, by about $1\frac{1}{2}$ per cent. In order that such unbalance should not exceed 0.1 mmf., the capacitance of each coil would need to be not more than about 6.0 mmf. It should be noted that this limit applies to the true inherent capacitance and not to the resultant of the coil capacitance and inductance.

Considering now the variation in resistance over the frequency range of the bridge, it can be shown, following the methods of Curtis and Grover, that the effect of a capacitance of the value noted

above on the resistance will not exceed one part in 100,000, which is quite satisfactory. The change in resistance (from the d-c. value) due to energy dissipation in the insulation is, however, somewhat larger than that due to the pure capacitance effect. This change is given to a close approximation by the expression

$$\Delta R = \frac{C\omega R^2 \tan \phi}{3},$$

where C is the total distributed capacitance between the wires of a bifilar winding and ϕ is the phase angle of the capacitance.[7] Clearly both the capacitance and its phase angle should be kept as small as practicable. An obvious and simple way of attaining the first object would be by using the very finest wire available. To do this, however, would in many cases result in excessive heating of the resistance. For bridge tests on telephone apparatus the ratio arm current will rarely exceed 25 milliamperes. The energy to be dissipated in a 1000-ohm coil is then about 0.5 watt, requiring a radiating surface of about 25 sq. cm. for a maximum temperature rise of 10 deg., which is a desirable limit. Since only the outer surface of such a coil is effective in radiating the generated heat the question of the number of layers requires consideration. Other factors being constant, it has been found that of the various possible arrangements that are easily constructed and mounted, a sectionalized, two-layer winding gives minimum capacitance. Hence one half of the winding is required to have an exposed surface of 25 sq. cm. The gauge of wire is then determined as a function of its specific resistance. A resistance alloy having a suitably low temperature coefficient (such as manganin, advance, etc.) will, on this basis, require that a wire no smaller than No. 38 A. W. G. be used. This is the size of wire used by Curtis and Grover and in the experiment covered by Table I. Using the data of this table, it was calculated that a Curtis and Grover type of coil, except for treatment, would have a change in resistance of not over one part in 50,000. The lower capacitance coils required from the phase-angle standpoint would have even smaller changes.

Summing up, then, the ratio arm coils were to be of approximately 1000-ohm resistance, wound with No. 38 A. W. G. double-silk-insulated manganin, or advance resistance wire, dried, but not impregnated with any moisture-resisting compound; the winding was to be arranged so that the true capacitances would not exceed 6.0 mmf. Besides being balanced for d-c. resistance, the resultants of their capacitance and inductance values were to be balanced to within 0.1 mmf. To

[7] See Note 6.

meet these requirements, the bridge coils were constructed as follows. The spool used is a glass cylinder $\frac{3}{4}$ in. in diameter. The winding is applied as follows: Starting at one end of the spool, a single strand of the wire is wound on until 14 inductive turns have been applied giving a resistance of approximately 50 ohms. Then the wire is tied, the direction of winding reversed, and an exactly equal number of turns wound over the first 14, but in an opposite direction. This brings the wire to the beginning. It is again tied, carried parallel to the axis of the spool over this first section and a second section wound. This is continued until ten sections have been applied. A thin sheet of mica is tied in place around the winding and the projecting ends of the wire bared of insulation. The whole is then baked to anneal the wire and dry the insulation. While hot, it is dipped several times in molten asphalt compound until a continuous coating of this moisture-proof material has been formed over the winding and surrounding mica wrapping. Adjustment for resistance balance is made by varying the length of the two wire ends.

The effective reactances of coils made as above are positive before assembly in their shields. The effect of the shield is to increase the capacitance. Table II gives data obtained on the two coils made for

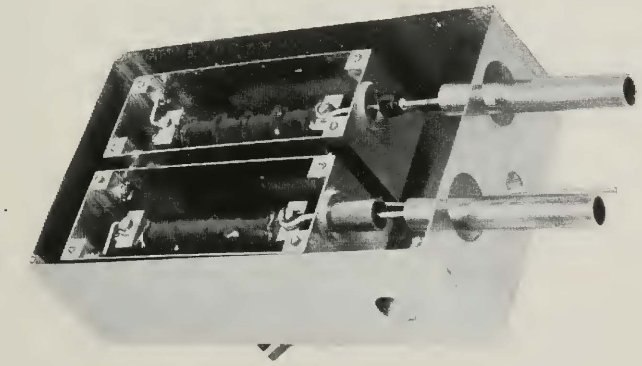


Fig. 12—Ratio arms

the bridge and shows the uniformity of phase-angle difference maintained by these coils over the operating frequency range. Final adjustment for reactance balance is made with the coils in the bridge circuit, a small amount of inductive coiling of the terminal leads sufficing for this purpose. In establishing this balance, use is made of the reversing switch described in the following section. Fig. 12 shows these coils assembled in their shields.

TABLE II

EFFECTIVE INDUCTANCE OF RATIO ARM COILS

Test Frequency	Microhenrys					
	Before Coating		After Coating		Assembled in Shield	
	Coil *A*	Coil *B*	Coil *A*	Coil *B*	Coil *A*	Coil *B*
1,000 cycles	7.4	6.9	6.7	6.3	− 1.3	− 1.0
50,000 cycles	7.4	6.9	6.7	6.2	− 1.2	− 0.8

Resistance of each coil = 1051.2 ohms

Reversing Switches. As will be noted from the diagram showing the circuit arrangement of the reversing switches, these are required to be completely enclosed in a shield which is connected to the junction point *D* of the bridge. They must also, of course, be subject to manipulation.

Obviously, then, this shield must be supported in some fashion from the outer enclosing ground shield of the bridge. The admittance between these two shields is a direct shunt on either one half or all of the impedance arm *CD*. While the capacitance component of this admittance can be readily balanced, it is more difficult to balance the conductance component, since the latter varies irregularly with frequency. Consequently, it is desired to make this factor so low that it can ordinarily be neglected. The construction adopted for this purpose is shown in the illustration, Fig. 13, which is a partially assembled view of the two reversing switches, their shield and its supporting brackets. It will be noted that the shield is supported by the brackets by means of small glass rods (four of which are shown). The low phase-angle characteristics of glass make it a favorable material to use, from this standpoint, but from the standpoint of machining into shape suitable for insulating supports, it is not so good. The construction shown, however, adapts it very well to this purpose. One other feature is worthy of special note; that is, the small change in position of any of the switch parts which occurs in effecting a reversal, the only metallic part moving being the small metal segments of the rotating disk.

Transformers. The transformers used for isolating the bridge circuit electrostatically from the source of testing current and from the detector system should have substantially zero external electromagnetic fields. This is to prevent inductive coupling to other parts

of the bridge circuit. For this purpose the transformer core is made in toroidal or ring form and the windings, both primary and secondary, are uniformly distributed about its circumference. The wound toroid is also completely enclosed in a sheet iron case.

The winding which is connected to the bridge has an electrostatic shield completely surrounding it for the purpose of concentrating all

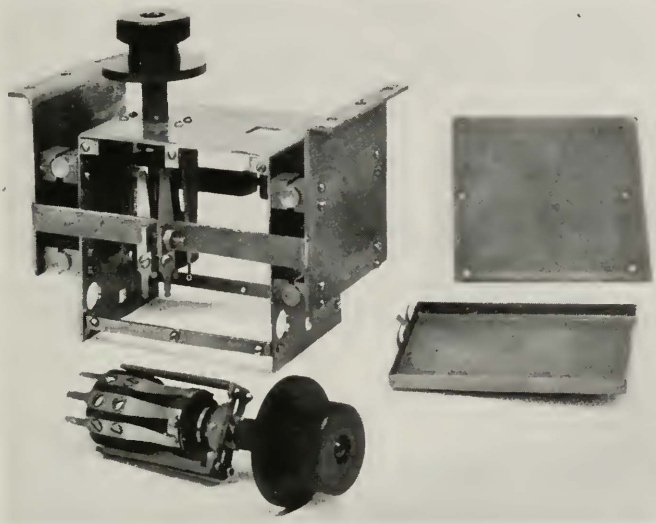


Fig. 13—Reversing switches

capacitance currents at one point. Around this localizing shield there is a second or ground shield. These two shields are made of sheet copper approximately No. 30 gauge (0.010 in.) in thickness. The inner winding terminal leads are brought out through a small brass tube leading into a terminal chamber which is an extension of the localizing shield. Since the admittance between the localizing and ground shields forms a major part of one of the balanced admittances shunting the impedance arms, it is desirable that the capacitance component be of low value and essential that it be constant. The conductance component should be negligibly small. To attain this end, the shields are separated at definite distances by means of hard rubber rings turned to fit the outer corners of the inner shield and the corresponding inner corners of the enclosing shield. These rings are made of the smallest cross-section consistent with mechanical strength requirements so as to introduce the minimum amount of solid material into the space between the shields. This minimizes the capacitance

and conductance values. These shields must not, of course, be allowed to act as short-circuited secondaries on the transformer which would be the case if they linked conductively with the windings. Each is therefore made in two parts similar to toroidal channels which upon assembly have their overlapping inner circumferences insulated from each other by means of thin mica laminations. Further details of the construction will be evident from a study of Fig. 14.

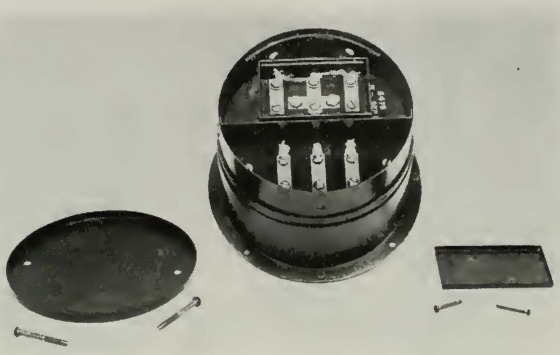


Fig. 14—Shielded transformer

The windings are, of course, proportioned so as to connect with a reasonable degree of efficiency the associated impedances. For best results two sets of transformers are used to cover the complete frequency range, one from 500 to 5000 cycles and the other from 5000 to 50,000 cycles.

Balancing Condensers and Impedance Arm Balance. It has been brought out previously that two adjustable capacitances are required, one to effect the proper adjustment of the bridge capacitances to ground and the other to balance the residual capacitances shunting one of the impedance arms. Such capacitances are provided in the form of adjustable air condensers each having a maximum value of about 500 mmf. The construction used is such as to give a high degree of stability of capacitance combined with low conductance characteristics. The arrangement of these condensers in relation to the other apparatus is shown in Fig. 3.

The effect on the accuracy of the bridge of the degree of balance of the impedance arms obtained by means of the balancing condenser C_b is determined as follows:

Capacitance Shunting the Impedance Arms. The equations giving the equivalent series inductance L' and resistance R' of a reactance of

inductance L and resistance R paralleled by a capacitance C are

$$L' = \frac{L - CR^2 - \omega^2 CL^2}{(1 - \omega^2 CL)^2 + \omega^2 C^2 R^2},$$

$$R' = \frac{R}{(1 - \omega^2 CL)^2 + \omega^2 C^2 R^2}.$$

When the bridge is balanced the equivalent series values of each component of the two impedance arms must be equal respectively to each other. If, however, the two arms have different shunting capacitances, it is evident that this equality will be obtained only by making the values of the two inductive branches of the parallel circuit somewhat different from each other. This difference represents the error introduced by the capacitance unbalance. When the values of the shunting capacitances are small these errors for the purpose of indicating their order are sufficiently closely given by the expressions

$$\Delta L_X = \omega^2 L_X^2 (C_X - C_S) \tag{7}$$

and

$$\Delta R_X = 2\omega^2 L_X R_X (C_X - C_S) \tag{8}$$

where C_X and C_S are the capacitances shunting the unknown and standard impedance arms, respectively. Reduced to percentages, these expressions become

$$\Delta L_X\ (\%) = 100\omega^2 L_X (C_X - C_S)$$

and

$$\Delta R_X\ (\%) = 200\omega^2 L_X (C_X - C_S)$$

and may also be written

$$\Delta L_X\ (\%) = 100 \frac{C_X - C_S}{C_R} \tag{9}$$

and

$$\Delta R_X\ (\%) = 200 \frac{C_X - C_S}{C_R}, \tag{10}$$

where C_R is the value of capacitance that would be required for resonance with inductance L_X at the test frequency.

These errors are thus proportional to the ratio of the capacitance unbalance to the resonating capacitance of the inductance under test. Ordinarily, values of the latter factor do not go below about 500 mmf. so that in the worst case a difference in capacitance of 0.1 mmf. corresponds to errors of 0.02 per cent and 0.04 per cent in inductance and resistance respectively.

Shields and Wiring. The shields have sufficient rigidity and are supported so as to maintain a definite and constant space relation to the part shielded and to the other shields. They are also of sufficiently high conductivity to maintain a common definite electrical potential at all points with respect to the part shielded.

The supports of shields or of bridge elements within the shields are as nearly as possible of constant specific inductive capacity, have low dissipative and leakage losses and are restricted to the minimum in number and size consistent with meeting the required rigidity of support.

Interconnecting conductors are shielded within brass tubes of approximately ½-in. diameter, the conductor which is of No. 10 gauge copper being supported at the axis of the tube by means of glass beads fitting snugly within the tubes and having holes through which the conductor passes. These beads are located longitudinally on the conductor by means of a small lump of solder placed on each side.

Standards. The impedance standards consist of adjustable self-inductance elements used in series with an adjustable non-inductive resistance. Each self-inductance element consists of a series of inductance coils and a low range inductometer of the Brooks type,[8] arranged in three decade formation and connected to dial switches by means of which any series combination of the coils can be selected. The inductometer is always in circuit and permits of balancing inductance values that fall between consecutive steps on the dials. Fig. 15 shows, schematically, the connections used for these standards and also the way in which they are shielded. It will be noted that the parts comprising each decade have a shield enclosing them and also all preceding decades of higher value. This makes a rather complicated mechanical arrangement but results in very important advantages from the standpoint of electrical performance. Due to the individual decade shields, each decade has effective values that are entirely independent of the settings of either of the other decades. Hence, once each individual setting of each dial has been calibrated, the value for the standard as a whole for any possible combination is obtained by simple addition of the separate dial values. This saves an immense amount of work in calibrating and also simplifies the reading of the standard. Without these shields the inter-coil and coil-to-ground admittances, at the higher frequencies, are sufficiently large to make the effective impedance of each decade setting depend to an appreciable extent upon the settings of the other decades. Under such conditions a calibration of

[8] H. B. Brooks and F. C. Weaver: "A Variable Self and Mutual Inductor," *Scientific Paper of the Bureau of Standards*, No. 290.

every combination would be required, and as this calibration would vary with frequency, a correction would be needed for each frequency value used.

In a standard of inductance to be used at high frequencies it is, of course, always desirable in order to minimize capacitance effects to have the inductance coils as small as possible. In the case of a completely shielded decade standard this is even more important on

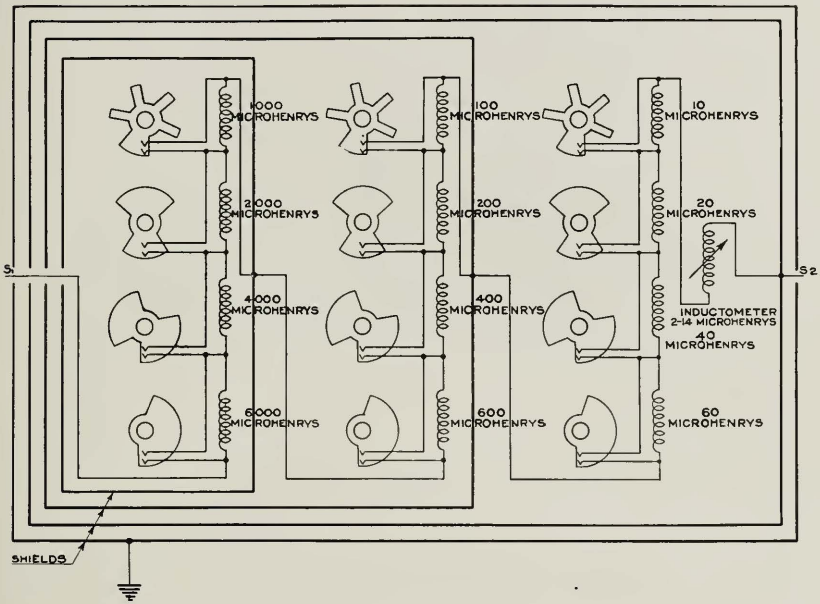


Fig. 15—Circuit diagram of shielded inductance standard

account of the capacitances added by the internesting shields. On the other hand the coil resistances should be quite small in comparison with their reactances, a requirement which tends to increase the coil dimensions. In addition to the above such coils should be highly stable in their inductance and effective resistance values with respect to the residual effects of direct and alternating currents and of temperature and humidity changes. Their values should also be of a satisfactory degree of constancy with respect to frequency and value of the testing current.

To meet these varied requirements the coils used in this bridge depart from the air core type ordinarily employed, in that they have a magnetic core of high stability and efficiency. Thus the desired inductance is obtained with a much smaller number of turns in the

winding giving a satisfactorily low resistance even in a coil only a fraction of the size of the equivalent air core coil. An adaptation of the new magnetic material, permalloy, has made this type of inductance standard possible.[9] Their cores consist of finely laminated, high

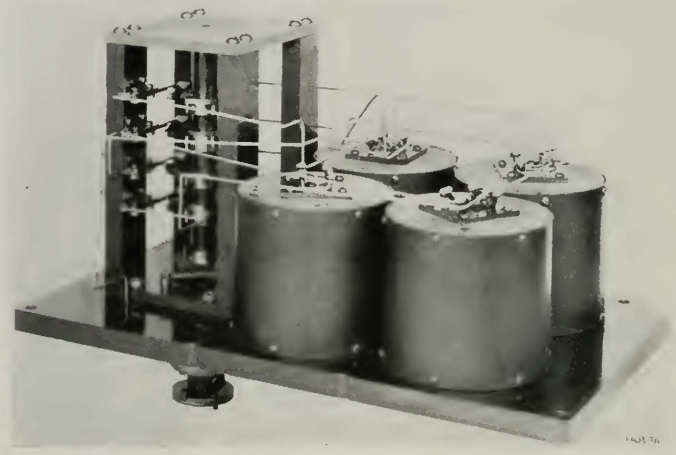


Fig. 16—Coil and dial switch assembly of typical inductance standard decade

specific resistance permalloy punchings, carefully annealed and assembled to form a toroidal structure whose effective permeability is about forty. On this is wound a sectionalized winding of insulated stranded conductor, the individual strands also being insulated from each other. The wound coils, after adjustment to the value desired, are sealed with moisture-proof compounds in phenol fiber cases. Fig. 16 shows an assembly of the four coils and switch which comprise one decade of the standard. In Table III are given data for typical coils illustrating their performance in respect to the above points.

The adjustable, non-inductive resistance is a commercial dial resistance box to which a shield has been added. It has five dials providing a range of 1000 ohms in steps of 0.01 ohm. Its shield is grounded in use, the resistance itself being connected usually between the *C* corner of the bridge and the inductance standard but in the case of an unknown impedance having a lower resistance than the standard from the *C* corner to the coil under test.

[9] H. D. Arnold and G. W. Elmen, *Franklin Institute Journal*, 195, 1923.

TABLE III

DATA ON COILS FOR INDUCTANCE STANDARDS

	Frequency Range	
	500 ∞ — 5,000 ∞	5,000 ∞ — 50,000 ∞
Overall Dimensions		
Diameter of case	6½ in.	3½ in.
Length of case	4 in.	3⅛ in.
Inductance Characteristics		
Nominal value	0.100 henry	0.010 henry
Change with frequency	+0.5% (500 ∞ — 5,000 ∞)	+2.5% (5,000 ∞ — 50,000 ∞)
Change with current	+0.01% per milliampere	+0.007% per milliampere
Temperature coefficient	−0.013% per deg. Fahr.	−0.005% per deg. Fahr.
Residual magnetization effect of one ampere d–c	Less than 0.01%	Less than 0.01%
Resistance Characteristics		
At an alternating current (effective value) of		
2.0 milliamperes	1,000 ∞ — 5.2 ohms 3,000 ∞ — 8.5 " 5,000 ∞ — 13.0 "	10,000 ∞ — 3.6 ohms 30,000 ∞ — 13.6 " 50,000 ∞ — 34.0 "
10.0 milliamperes	1,000 ∞ — 5.5 " 3,000 ∞ — 9.1 " 5,000 ∞ — 13.9 "	10,000 ∞ — 3.7 " 30,000 ∞ — 13.9 " 50,000 ∞ — 34.7 "
Temperature coefficient	−0.017% per deg. Fahr. at 3,000 ∞	− < 0.01% per deg. Fahr. at 30,000 ∞

PERFORMANCE

As was stated earlier, the operation of the bridge involves an initial balancing of its capacitances. It is then ready for impedance testing which is done by suitably connecting the unknown and standard impedances to the proper terminals and adjusting the latter until a balance of the bridge is obtained. The corresponding constants of the two impedance arms are then taken as being equal. Those of the standards being known, by calibration, it follows that those of the impedance under test can be simply derived. The degree of precision obtained depends upon two major factors, the accuracy of the calibration of the standards and the accuracy of the bridge comparison. The matter of calibration is beyond the scope of this paper and it will be assumed that a suitable calibration of the standards is available.

Due to the construction used the factors determining the accuracy of the bridge comparison of impedances are reduced to the following:

1. The resistance balance of arms AB and AD.
2. The effective shunt capacitance balance of these arms.

3. The direct capacitance balance of arms BC and CD.

4. The direct conductance balance of arms BC and CD.

5. The series inductance balance of the interior wiring to the impedance terminals of arms BC and CD.

As was explained in the foregoing two switches are provided for independently reversing the ratio arms (AB and AD) and also the outside connected impedances. These, therefore, afford a very convenient means of checking the above balances of the bridge network. By a suitable choice of the test condition under which the reversals are made, a fairly good approximation of the effect of the separate items can be made. The following series of tests indicate how this was done on one of these bridges.

The junction point C was first grounded. Then, with a telephone receiver as the detector and with a test current having a frequency of 1600 cycles, the setting of the condenser C_b was varied until a balance was obtained. The arms Z_S and Z_X were both open-circuited in this test; hence the capacitances shunting these arms alone determined the balance point. This balance was very sharp indicating that the shunting conductances were either very small or else accidentally well balanced. Leaving the condenser set at its balance point, there was then connected into one of the impedance arms a toroidal self inductance standard having a nominal inductance of 0.200 henry and an effective resistance of about 50 ohms. In the other arm there was connected a similar standard of the same nominal but of slightly lower actual value in series with a small adjustable inductance and adjustable resistance, each of sufficient range to effect a balance of the corresponding constants. The extension inductance was graduated in steps of one microhenry and the resistance in steps of 0.001 ohm. Balances for the four combination settings of the reversing switches, S_R and S_Z, were then made, only the extension elements being varied in getting these balances. Readings as given in Table IV were obtained.

TABLE IV

Switch Position		Extension Inductance	Extension Resistance
S_R	S_Z		
Right	Right	124 ± 2 microhenrys	4.00 ± 0.01 ohm
Left	Right	120 " " "	4.00 " " "
Right	Left	124 " " "	3.87 " " "
Left	Left	128 " " "	3.87 " " "

Consideration of these figures led to the following conclusions:

1. The change in inductance balance due to reversal of the ratio arms is not more than eight parts in 200,000 or 0.004 per cent. It has been shown previously that the phase-angle balance of the ratio arms is not critical with respect to inductance readings. Hence, the change in inductance may be considered to be closely indicative of the resistance unbalance of the ratio arms. From the above data it is seen that this does not exceed 0.01 per cent.

2. The change in resistance due to reversal of the ratio arms being within the limits of observational error, the phase angles of the coils themselves are, as nearly as can be determined by this test, exactly balanced.

3. Since the change of inductance balance due to reversal of the impedance arms is no more than that due to the ratio arm reversal, the capacitance balance of the impedance arms is apparently satisfactory. It should be noted, however, that this balance is not critical under these test conditions. (See eq. 9.)

4. Since the resistance balance was appreciably affected by reversal of the impedance arms, it appeared that there was an unbalancing factor present which, if affecting the ratio arms, was not reversed with them but which if present in the impedance arms was reversed. The latter might have been an unbalance of the impedance arm shunt conductances but it was assumed that this unbalance was quite small. On the other hand, as was mentioned earlier in the paper, there are two small inter-shield capacitances shunting the ratio arms which are not reversed by the ratio arm switch and it seemed likely that an unbalance of these capacitances was causing the change in resistance reading. This proved to be the case, as adjustment of the balance of these capacitances for which, as previously noted, provision had been made, resulted in identical resistance readings being obtained for both positions of the impedance arm switch. After this adjustment had been made the previous tests were repeated, resulting in readings as given in Table V.

TABLE V

Switch Position		Extension Inductance	Extension Resistance
S_R	S_Z		
Right	Right	124 ± 2 microhenrys	3.93 ± 0.01 ohm
Left	Right	120 " " "	3.93 " " "
Right	Left	124 " " "	3.93 " " "
Left	Left	126 " " "	3.93 " " "

As a further check on the performance of this unit, two inductances, each of about 0.01 henry inductance, were compared at two frequencies, 25,000 and 50,000 cycles. Table VI gives the readings obtained in these tests.

TABLE VI

Frequency Cycles	Switch Position		Extension Inductance	Extension Resistance
	S_R	S_Z		
25,000	Right	Right	123 ± 1 microhenry	5.1 ± 0.1 ohm
	Left	Right	122 " " "	5.1 " " "
	Right	Left	129 " " "	5.1 " " "
	Left	Left	128 " " "	5.1 " " "
50,000	Right	Right	38 " " "	24.2 " " "
	Left	Right	36 " " "	24.2 " " "
	Right	Left	57 " " "	23.8 " " "
	Left	Left	57 " " "	23.8 " " "

At the 25,000-cycle frequency the maximum difference in inductance from the probable correct balance does not exceed ± 5 microhenrys or 0.05 per cent. The resistance balances check to within 0.1 ohm. At 50,000 cycles, the inductance change due to ratio arm reversal is still within ± 0.01 per cent while the resistance change is within 0.1 ohm which would be just under one per cent for a coil of this reactance and a reactance to resistance ratio of 300. This is a critical test of the ratio arm phase-angle balance. Hence it may be concluded that over the entire frequency range the ratio coils meet all balance requirements. The changes in inductance occurring at the higher frequencies when the impedance arms were reversed indicated that the residual capacitance unbalance of these arms was too large. Readjustment of the balancing condenser reduced the changes to less than 0.02 per cent. The difference in resistance balance at the 50,000-cycle frequency indicates that the conductances shunting the impedance arms are not negligible at this frequency. For more accurate results these conductances would require balancing. This would be quite practicable by means of a variable high resistance shunt.

In making each series of tests outlined above, the testing potential applied to the bridge was varied by means of a resistance potentiometer from the lowest value at which a balance could be made to the maximum of the supply oscillator. This was to check the completeness of the shielding and to detect the presence of any coupling with the supply circuit. In no case was there any discernible change in balance produced.

Conclusion

A system of electrostatic shielding for a direct reading bridge for the measurement of inductive impedances at frequencies up to 50,000 cycles has been described.

The general considerations defining the balances of the various capacitances which this shielding controls have been discussed and specific requirements derived for a typical range of impedances. The physical construction of a bridge designed to meet these requirements has been described and test data given illustrating its performance. These have shown it to be capable of comparing impedances over the above frequency range with a precision which approximates that ordinarily found in routine direct current resistance measurements.

Letters to the Editor

From Mr. Arne Fisher: A Relation Between Two Coefficients in the Gram Expansion of a Function

From Dr. W. A. Shewhart: A Reply

From Mr. Fisher: A Further Note

To the Editor of the Bell System Technical Journal:

In a number of valuable and interesting contributions to this *Journal*, Dr. W. A. Shewhart has made an extended use of the infinite series of Gram. With all the controversy that at present is going on between the pure empiricists, attempting on the one hand to dragoon statistical analysis into a mere *inductio per simplicem enumerationem*, and the a priori theorists on the other hand, who claim that statistical methods so-called are nothing more than simple and evident applications of well-known principles of the probability calculus as formulated by Laplace, it has been a source of satisfaction to me to note that Dr. Shewhart apparently has given the latter methods a place of preference over the methods of the out and out empiricists.

Because of the fact that I happen to be responsible for having called the attention of English-speaking readers to the series of Gram and to have emphasized that Gram's development anteceded the less general developments by Edgeworth and the very special formula by Bowley by more than 20 years, I hope that I may be afforded an opportunity through the medium of your *Journal* to point out in brief form a few decidedly simple features of the Gram series which greatly add to its practical applications in statistical work.

Moreover, it seems that Dr. Shewhart, as well as other students in this country, have received a somewhat different idea about the nature of the Gram series than that which it was my intention to convey in my book on "The Mathematical Theory of Probabilities." This probably is my own fault. For while I have given in the above-mentioned book a description of the various methods for determining the coefficients of the individual terms of the Gram series, I did not mention the various degrees of approximations according to the number of terms as retained in the series itself. The reason for this omission is due primarily to the fact that I expect to treat this aspect in a forthcoming second volume of the book on probability in connection with the presumptive error laws of the a posteriori determined semi-invariants, which laws contain as a special case the evaluation of the standard (or probable) errors of the constants of the frequency curves.

The omission on my part to properly emphasize the close relation between the theory of sampling (i.e., the a posteriori probability theory) and the Gram series is probably also responsible for the fact that Dr. Shewhart in several of his articles has intimated that two terms in the Gram series in certain instances yield a better approximation than three or more terms. This idea has probably arisen from the mistaken notion on the part of Bowley of the generalized probability curve, which is a special example of the general Gram series. The following brief remarks should, therefore, not be taken as a criticism of Dr. Shewhart's work, but rather as a sort of amplification of some of the chapters in my own book on "The Mathematical Theory of Probabilities."

Gram's series, like the Fourier series, offers a perfectly general method for the expansion of arbitrary functions and is, contrary to the opinion of some students, not limited to frequency functions, although it there happens to be especially useful.

The underlying principles of the Gram series may be set forth briefly as follows: Let $F(x)$ be the true (or presumptive) function, which is known from either purely *a priori* considerations, or from observations, and let $G(x)$ be another function (the so-called generating function), which gives a rough approach to $F(x)$. Then according to Gram's method, we have

$$F(x) = c_0G(x) + c_1G'(x) + c_2G''(x) + \ldots + c_nG^n(x). \qquad (1)$$

The generating function $G(x)$ may assume a variety of forms. In the case of generalized frequency functions, it is customary to select as the generating function, $G(x)$, a quantity $z = h(x)$ which is normally distributed, and write $F(x)$ as [1]

$$F(x) = c_0\varphi_0(z) + c_1\varphi_1(z) + c_2\varphi_2(z) + \ldots + c_n\varphi_n(z), \qquad (2)$$

where $\varphi_0(z) = \dfrac{1}{\sqrt{2\pi}}\, e^{-z^2:2}$ is the generator and $\varphi_1(z)$, $\varphi_2(z)$ . . . $\varphi_n(z)$ its derivatives.

When viewed from the theory of elementary errors as originally introduced by Laplace in his monumental work, "Theorie des Probabilities," the Gram series takes on special significance in the way in

[1] If $z = h(x) = (x - M):\sigma$, or a linear function of x, and if the origin of the co-ordinate system is laid at M with σ as its unit, we have the *special* case, or the Charlier A series of the well-known form

$$F(x) = N[\varphi_0(z) + \beta_3\varphi_3(z) + \beta_4\varphi_4(z) + \ldots].$$

The various types of the frequency curves of Pearson may of course also be used as generators in the Gram series.

which the possible combinations of the "elementary errors" actually enter into the expansion. It can be shown that there exists a definite relationship between on the one hand the relative order of magnitude of the elementary errors and, on the other, the arrangement of the individual terms of the Gram series.[2]

This relationship was already established by Thiele. It was probably first concisely formulated by Edgeworth, and later on by Charlier and Jörgensen.

The various degrees of approximations can be expressed by the following schemata:

1st approximation $\varphi_0(z)$,
2d approximation $\varphi_0(z) + c_3\varphi_3(z)$,
3d approximation $\varphi_0(z) + c_3\varphi_3(z) + c_4\varphi_4(z) + c_6\varphi_6(z)$,
4th approximation $\varphi_0(z) + c_3\varphi_3(z) + c_4\varphi_4(z) + c_6\varphi_6(z) + c_5\varphi_5(z) + c_7\varphi_7(z) + c_9\varphi_9(z)$.

The first approximation is the usual normal curve. The second is the one which the English statistician, Bowley, erroneously thinks represents a generalized frequency function and for which Dr. Shewhart has shown a marked preference. The third approximation, except for the term involving the sixth derivative, has been used very extensively by Charlier.

Through the publication by C. V. L. Charlier in 1906 of extensive tables to four decimal places of the third and fourth derivatives, the Gram series was made available for practical statistical work in the case of frequency distributions with a moderate degree of skewness and excess (kurtosis). But although Charlier was aware of the fact that the retention of the fourth derivative—which is related to excess (kurtosis)—automatically brings about the inclusion of the sixth derivative, it was not before Jörgensen issued his large numerical tables of the first six derivatives to seven decimal places that we were able to do full justice to the third approximation of the Gram series. Incidentally it might in this connection be mentioned that it is doubtful if the much lauded test for "goodness of fit" as devised by Pearson

[2] Whenever we use the method of moments, the arrangement of the individual terms is *not* arbitrary but must be made according to "order of magnitude" of the various derivatives; and the orders of magnitudes do not correspond to the indices of the derivatives. The generic term "order of magnitude" has in this instance only reference to the formation of the "elementary errors"; if taken in any other sense it is meaningless. The fourth and sixth derivatives are of the same order of magnitude; while the fifth, seventh and ninth all are of the next order following the fourth and sixth. The concept of the different orders of magnitude of the elementary errors is due to Poisson who already in 1832 arrived at the second approximation of the Gram series.

really is able to test the graduating ability of the Gram series as adequately as the more powerful, although far more complicated, "error critique" of Thiele. From Pearson's derivation it appears that his test is not able to take care of elementary errors beyond the first or second order, while it is necessary to consider the formation of elementary errors of the third order in the third approximation of the Gram series. In some work I have been doing in the way of construction of compound mortality curves, I have at least found that the Pearson test is inadequate, if actually not misleading, because it apparently fails to measure the effect of the elementary errors of higher order which enter into the formation of such compound mortality curves.

There exists, however, a very simple relationship between the coefficients c_3 and c_6 in the third approximation. We have, namely, with a fair approach to exactitude, the simple relation: $c_6 = \frac{1}{2}c_3^2$. It is therefore not necessary to calculate the semi-invariants or moments of higher orders than those of the fourth order, since we shall have

$$F(x) = c_0\varphi_0(z) + c_3\varphi_3(z) + c_4\varphi_4(z) + \tfrac{1}{2}c_3^2\varphi_6(z)$$

as a third approximation.

As an illustration of the above formula, we may select the expansion of the point binomial $(0.1 + 0.9)^{100}$. We have here, according to the formulas on pages 263–264 of my "Mathematical Theory of Probabilities":

$$s = 100, \qquad p = 0.1, \qquad q = 0.9$$

and

$$\lambda_1 = M = sp = 10, \qquad \sigma = \sqrt{spq} = 3, \qquad c_3 = -0.0444, \qquad c_4 = 0.0021$$

and

$$c_6 = \tfrac{1}{2}c_3^2 = 0.0010,$$

or

$$(0.1 + 0.9)^{100} = \tfrac{1}{3}[\varphi_0(z) - 0.0445\varphi_3(z) + 0.0021\varphi_4(z) + 0.0010\varphi_6(z)],$$

where

$$\varphi_0(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2:2}$$

and

$$z = (x - 10) : 3.$$

A comparison between the above approximation and the true expansion of the point binomial $(0.1 + 0.9)^{100}$ to 4 decimals is given in the following table.

x = No. of Successes	Gram Series	True Value	No. of Successes	Gram Series	True Value
0	.0000	.0000	13	.0744	.0743
1	.0003	.0003	14	.0515	.0513
2	.0016	.0016	15	.0327	.0327
3	.0060	.0059	16	.0192	.0193
4	.0160	.0159	17	.0105	.0106
5	.0338	.0339	18	.0054	.0054
6	.0594	.0596	19	.0026	.0026
7	.0888	.0889	20	.0012	.0012
8	.1149	.1148	21	.0005	.0005
9	.1305	.1304	22	.0002	.0002
10	.1318	.1319	23	.0001	.0001
11	.1198	.1199	24	.0000	.0000
12	.0988	.0988			

The approximation is in this case well nigh perfect and comes much closer to the true values of the point binomial than any of the six approximations as given in Dr. Shewhart's article in the January 1924 number of this *Journal*. It also shows that with exactly the same amount of computation as that involved in the so-called Charlier A series, we can reach greatly improved results through the inclusion of the sixth derivative in the series. This arises from the important fact that once we have computed the coefficients c_3 and c_4, it is not necessary to calculate c_6 since $c_6 = \frac{1}{2}c_3^2$ approximately. Moreover, since extensive tables, notably those of Jörgensen, now are available for the normal function and its first six derivatives, there seems no good reason why we should not use the more exact approximation than the inexact formula by Bowley.

In conclusion, it might be well to emphasize the fact that while it is important to consider the relative order of magnitudes of the separate terms in the Gram series when we use the methods of semi-invariants or of moments, such restrictions are not necessary if we use the method of least squares in conjunction with properly determined weights.

ARNE FISHER.

December 10, 1926.

To the Editor of the Bell System Technical Journal:

I have read Mr. Fisher's communication with considerable interest. We who do not read the Scandinavian language owe much to him for his very able amplification and interpretation of many important contributions of the Scandinavian school of mathematical statisticians and this debt has been increased by the above communication insofar as it brings to light a very interesting relationship (the discovery of

which is attributed to Thiele), namely, that in the notation of the communication the constant c_6 is approximately equal to $\frac{c_3^2}{2}$.

Mr. Fisher definitely states that no criticism of my work is intended, but incidental to bringing out the above relationship he makes certain statements upon which I should like to comment briefly.

He states that the omission on his part to properly emphasize a close relation between the theory of sampling and the Gram series is probably responsible for the fact that I have intimated that two terms of the Gram series in certain instances yield a better approximation than three or more terms. To my knowledge this is not the case.

The special form of the Gram series used in my published articles in this *Journal* is that represented by his Equation 2.[1] The validity of this expansion rests upon the Lebedeff theorem.[2] So far as I am aware I have not intimated that two terms of the series yield a better approximation than three or more terms in the sense that

$$| F(z) - [c_0\varphi_0(z) + c_3\varphi_3(z)] |$$

should be less than

$$| F(z) - [c_0\varphi_0(z) + c_3\varphi_3(z) + \ldots + c_n\varphi_n(z)] |$$

irrespective of n, although it is in this sense that Mr. Fisher discusses his example of the graduation of $(.9 + .1)^{100}$. To have done so would have been an obvious blunder because, assuming the Lebedeff theorem to be true, the absolute value of the difference ϵ between the function $F(z)$ and the sum of the first n terms of the series can be made as small as we please by taking n sufficiently large.[3]

I did say, however, in my article in the October issue of this *Journal:* "Carrying out steps 1 and 2, we conclude that the best theoretical equation representing the data in Fig. 1 is either the Gram-Charlier series (2 terms) or the Pearson curve of Type IV for both of which the estimates of the parameters may be expressed in terms of the first four moments μ_1, μ_2, μ_3 and μ_4 of Fig. 3." Of course the first two terms of the Gram-Charlier series requires only μ_1, μ_2 and μ_3. "Best" as used here obviously is in the sense of probability of fit which is entirely different from saying that the first two terms is the best approximation in the sense discussed by Mr. Fisher at least as illustrated by his

[1] It is of course understood that, in practice, transformations are made so that c_1 and c_2 are both equal to zero. In what follows, therefore, the second term of the series will be $c_3\varphi_3(z)$.

[2] Fisher, Arne, "Mathematical Theory of Probabilities," 2d edition, 1922, p. 203.

[3] It can be seen from my published work, however, that the sum of two terms is sometimes better than the sum of three.

example. In this case I found that the probability of fit for two terms was greater than that for three. Now, I find that it is as good as for Mr. Fisher's third approximation. It may be of interest also to know that statistical distributions sometimes arise where the first three terms give as good a fit as Mr. Fisher's third approximation involving 4 terms. This is particularly true when the universe from which the sample is drawn is nearly symmetrical. My action in this connection can be justified both upon theoretical and practical grounds but we need not do more than mention this point to make sure that the reader will not confuse my statement quoted above with what Mr. Fisher is talking about in his communication.

Having thus dismissed the questions which may arise in connection with published work in this *Journal*, I should like to add a word or two of caution to the reader of Mr. Fisher's letter where it reads: "Moreover, since extensive tables, notably those of Jörgensen, now are available for the normal function and its first six derivatives, there seems no good reason why we should not use the more exact approximation than the inexact formula by Bowley."

We have made far more use of the Gram series in connection with our inspection work than indicated in the published papers. In this work we have found that it is theoretically not necessary in certain instances and in many more instances it is not practical to follow Mr. Fisher's suggestion. I shall limit my remarks to the application of the series which we have made in expanding a known function in terms of an infinite series in which the generating function is the normal law. In this connection the outstanding practical question is: Given the known function $F(x)$, what number n of terms of the infinite series must we take in order that the absolute magnitude of the difference between the function $F(x)$ and the sum of the n terms will be less than a given preassigned quantity ϵ? I am sorry that Mr. Fisher does not answer this question. Instead he proposes a grouping of terms upon the basis suggested in a footnote to his article. Now, it may easily be shown in the particular case cited by Mr. Fisher, i.e., the graduation of the point binomial $(.9+.1)^{100}$, that the sequence of signs depends upon the value of z, that for certain values of z his second approximation is just as good as his third, and that in many instances the difference between the second approximation and the third is not sufficiently great to be of any practical importance. Whether we should use the second, third, or higher approximation in a given case is one for special consideration.

In closing let me say that I have not made the above remarks with any intention of discrediting the applications of this series but rather

to indicate to the casual reader that there are certain technical questions involved in its application which must be given due consideration even beyond the stage outlined in Mr. Fisher's communication. I have found that this series often has many advantages over competing methods of analyzing data although not all of these advantages are referred to in the literature of the subject.

W. A. SHEWHART.

December 28, 1926.

To the Editor of the Bell System Technical Journal:

The question raised by Dr. Shewhart as to the measure of the absolute magnitude of the difference between a known function, $F(x)$, and the first n terms of the series has been treated by Gram in his original article on "*Rækkeudviklinger bestemte ved Hjælp af de mindste Kvadraters Metode.*" (On Development of Series by means of the Method of Least Squares.) In this article Gram also discusses at length the decidedly practical question of arriving at an estimate of the remainders (or residuary terms), which invariably occur in practice where we, of course, are forced to deal with a finite number of terms.

It would, however, be beyond the limits of the present communication to enter into this aspect of the question, which necessarily is somewhat complicated. In passing it, I wish merely to state that Gram's original method of determining the coefficients in the series on the basis of the principle of least squares is decidedly easier to apply than the relatively cumbersome method of moments in arriving at a reliable measure of the remainder of the series after, say, the n^{th} term.

Dr. Shewhart's further contention that two terms of the Gram series sometimes give as good a fit as three or even four terms, and that three terms in the case of nearly symmetrical distributions serves as well as four terms, seems to me to be almost self-evident from a simple consideration of the way in which the coefficients c actually enter into the series.

All the terms containing uneven indices tend to produce skewness, and all the terms with even indices produce excess (kurtosis). If the coefficient c_3 is not too large, and if c_4 is small as compared with c_3, it is evident that

$$F(x) = c_0\varphi_0(z) + c_3\varphi_3(z)$$

will give about as good an approximation as

$$F(x) = c_0\varphi_0(z) + c_3\varphi_3(z) + c_4\varphi_4(z) + \tfrac{1}{2}c_3^2\varphi_6(z).$$

On the other hand, in nearly symmetrical distributions with a pro-

nounced excess (kurtosis), where c_4 is large as compared with c_3, it seems also reasonable that

$$F(x) = c_0\varphi_0(z) + c_4\varphi_4(z)$$

might in certain instances give as good a fit as

$$F(x) = c_0\varphi_0(z) + c_3\varphi_3(z) + c_4\varphi_4(z) + \tfrac{1}{2}c_3^2\varphi_6(z).$$

These aspects of the series have been discussed by Thiele.

ARNE FISHER.

January 10, 1927.

Abstracts of Recent Technical Papers from Bell System Sources

Loading for Telephone Cable Circuits.[1] D. W. Whitney. This paper summarizes the principal characteristics of the loaded telephone line and discusses the major improvements in loading. Up to 1900 there was a general avoidance of the use of cable in the toll telephone plant, due to the high attenuation and distortion of speech currents not experienced in open wire lines. By the use of the loading coil and the telephone repeater, a network of toll cables has grown very rapidly which now connects the large population centers of the Atlantic seaboard and the upper Mississippi Valley region.

Electric Annealing of Magnetic Materials for Telephone Apparatus.[2] W. A. Timm. This paper briefly describes the annealing equipment used by the Western Electric Company prior to 1909 and some years afterward at the Hawthorne plant, in contrast to the more recently installed electrical equipment.

The Problem of Secondary Metals in World Affairs.[3] F. W. Willard. Preliminary to the discussion of the secondary metal industry, the author gives a brief statistical survey of the rate of depletion of the world's primary metal resources along with a study of metal production in the United States.

The rapid rate of exhaustion of known resources is emphasized. Graphs also show the relation of prices and production of copper and lead in the United States over a period of 40 years.

Each important economic metal is discussed briefly showing how its present commercial uses affect its return through the secondary market. Attention is directed to the degradation of metals in use, forever eliminating them for re-use. Platinum, though not generally classified among economic metals, is treated briefly because of its key importance in certain industries. The seriousness of the present trend of using platinum in jewelry is indicated.

The growing importance of the secondary metal industry is shown graphically and briefly discussed, leading to an emphasis of the need

[1] To be published in the *Proceedings* of the Telephone and Telegraph Section of the American Railway Association.

[2] *Trans. American Soc. Steel Treat.*, p. 782, Nov. 1926.

[3] *Industrial and Eng. Chemistry*, p. 1178, Nov. 1926. Presented at the Round Table Conference on the rôle of chemistry in the world's affairs at Williamstown, Mass., Aug. 1926.

for encouragement of scientific research to increase metal recovery. The United States is favored above most nations in being largely self-contained with respect to original sources of economic metals, yet it has no source for tin and platinum and inadequate sources for manganese, chromium and other less common metals essential to present-day steels.

The work of a joint commission of the American Institute of Mining and Metallurgical Engineers and the Mining and Metallurgical Society of America is outlined, giving briefly their recommendations for international control of minerals.

The paper is concluded by emphasis of the importance of secondary metal recovery to national existence in times of stress when the nation may be thrown entirely upon its own resources. The conclusion also makes an appeal for the consideration of international conventions on economic metal exchanges and suggests that competent technical men be taken more into political councils in the treatment of a problem of this kind.

Tone Reproduction in the "Halftone" Photo-Engraving Process.[4] HERBERT E. IVES. The "halftone" photo-engraving process was invented, and its technique developed, prior to the days of what is commonly called "photographic sensitometry." The necessary conditions and the appropriate operations for securing "highlight" and "shadow detail" were found by empirical studies guided by the appearance of the result, as appraised by the unaided eye. No comprehensive sensitometric study of the halftone reproduction process appears to have been made, at any rate none have been published. The work described is a rough survey of the problem, but, due largely to the use of accurate photometric measurements, and the correlation of these measurements with other sensitometric data, rather decisive conclusions have been possible as to the essential characteristics of the process in question, and on the procedures necessary for its complete success.

Frequency Measurements with the Cathode Ray Oscillograph.[5] FREDERICK J. RASMUSSEN. The cathode ray oscillograph frequency measurement circuit described, differs from previous circuits in the use of by-pass condensers and plate leaks which permit the connection of the oscillograph to a.-c. circuits having large d.-c. components and which permit the use of biasing controls for shifting the position of patterns on the screen.

[4] *J. O. S. A. and R. S. I.*, March, 1926, p. 537.

[5] Presented before the *A. I. E. E.*, New York, N. Y., Nov. 1926.

Reference oscillators are used in conjunction with the frequency standards. They are of a type chosen for their high stability.

The well-known properties of Lissajous' figures are reviewed briefly and then are developed more fully for the cases in which only one term of their ratios may be determined from the oscillograph pattern. Following a general discussion of the accuracy of syntonization, there is discussed a detailed method of calibrating oscillators. The patterns used may be interpreted from one term of their ratio.

Several special circuits are described for use in frequency measurement work with the cathode ray oscillograph.

The methods and apparatus described are suitable not only for the technical measurements of a development and research nature but are equally adaptable for routine commercial work. The advantages which particularly commend themselves are the rapidity with which such work may be done and the ease with which the average man can learn the work.

Contributors to this Issue

JOHN R. CARSON, B.S., Princeton, 1907; E.E., 1909; M.S., 1912; Research Department, Westinghouse Electric and Manufacturing Company, 1910–12; instructor of physics and electrical engineering, Princeton, 1912–14; American Telephone and Telegraph Company, Engineering Department, 1914–15; Patent Department, 1916–17; Engineering Department, 1918; Department of Development and Research, 1919–. Mr. Carson's work has been along theoretical lines and he has published several papers on theory of electric circuits and electric wave propagation.

JOHN DAVIDSON, JR., B.E., Tulane University, 1906; E.E., Cornell University, 1909. Testing Department, General Electric Company, 1906–07; Construction Department, Allis Chalmers Company, 1907–1908; Electrical Department, Carnegie Steel Company, 1909–11; American Telephone and Telegraph Company Engineering Department, 1911–19, and the Department of Development and Research, 1919–. Since 1911 Mr. Davidson has been engaged in equipment development work and is now engaged on toll switchboard and signaling arrangements.

P. G. EDWARDS, B.E.E., Ohio State University, 1924; Western Union Telegraph Company, Traffic Department, 1919–21, Plant Department, 1921–22; American Telephone and Telegraph Company, Long Lines Plant Department, 1922–24; Department of Development and Research, 1924–. Mr. Edwards has been engaged in the development of toll test board equipment and toll test board methods.

H. W. HERRINGTON, Washburn College, 1918–21; B.S. in E.E., University of Kansas, 1923; E.E., University of Kansas, 1926; American Telephone and Telegraph Company, Department of Development and Research, 1923–. Mr. Herrington has been engaged in the development of toll test board equipment and toll test board methods.

KARL K. DARROW, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

I. B. Crandall, A.B., Wisconsin, 1909; A.M., Princeton, 1910; Ph.D., 1916; Professor of Physics and Chemistry, Chekiang Provincial College, 1911–12; Engineering Department, Western Electric Company, 1913–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Crandall has published papers on infra-red spectroscopy, the condenser transmitter, thermophone, etc. More recently he has been associated with studies on the nature and analysis of speech which have been in progress in the Laboratory.

Lloyd Espenschied, Pratt Institute, 1909; United Wireless Telegraph Company as radio operator, summers, 1907–08; Telefunken Wireless Telegraph Company of America, assistant engineer, 1909–10; American Telephone and Telegraph Company, Engineering Department and Department of Development and Research, 1910–. Took part in long distance radio telephone experiments from Washington to Hawaii and Paris, 1915; since then his work has been connected with the development of radio and carrier systems.

William J. Shackelton, B.S. in E.E., University of Michigan, 1909; Western Electric Company, Manufacturing and Installation Department, 1909–10; Engineering Department, 1910–24; Bell Telephone Laboratories, 1925–. Mr. Shackelton's principal activities have been in connection with the design of loading coils and the development of methods of high frequency measurement.

The Bell System Technical Journal

April, 1927

Developments in the Manufacture of Copper Wire[1]

By JOHN R. SHEA and SAMUEL McMULLAN

SYNOPSIS: This paper describes a new copper rod and wire mill located at the Western Electric Company's Plant at Chicago. It includes a brief survey of the copper rolling and wire drawing art at the time the investigation was started; a summary of tests made in varying the practice in rod rolling and wire drawing; and an outline of the work done by the Western Electric Company engineers in developing and designing new types of wire drawing machinery.

The rod mill is converting 225 pound wire bars into ¼" rod in fourteen instead of the usual eighteen passes. This is accomplished by making heavier reductions in the first four passes while the copper is hot. The new wire mill incorporates many novel features, and the wire drawing machines are more compact in design and of considerably higher speed than those in general use.

The design of the wire mill was undertaken following a comprehensive survey of wire drawing processes and equipment used in this country and abroad. Part of this survey consisted of a study of the manufacture of diamond dies, it having been found that dies suitable for high speed drawing required a differently shaped "approach," a better polish, and a shorter "land," than those which were available for low speed work. The economies in floor space and plant investment due to the use of more compact and higher speed machinery are outlined. Some of the outstanding features in plant arrangement which contribute to more efficient operation are discussed in the concluding pages.

RAPID growth in the various branches of electrical communication accompanied by widespread research are constantly leading to the more efficient and economical meeting of the increasing demands for service. In this connection, one of the more recent and very interesting investigations indicated the possibility of effecting substantial improvement in the process of manufacturing copper wire. Accordingly, a comprehensive study of all the factors concerned was undertaken which resulted in the construction of a rod and wire mill at Chicago embodying many unique and improved features. A schematic layout is given in Fig. 1.

At the outset the sources of copper and its transportation were studied and it was found more economical to ship wire bars to Chicago for conversion into wire than to locate a wire mill near some of the large refineries and ship wire to the factory. It was also considered that this plan would reduce the investment in wire during the process of manufacturing cable and telephone apparatus.

[1] Read before the Midwinter Convention of the A. I. E. E., New York City, Feb. 8, 1927.

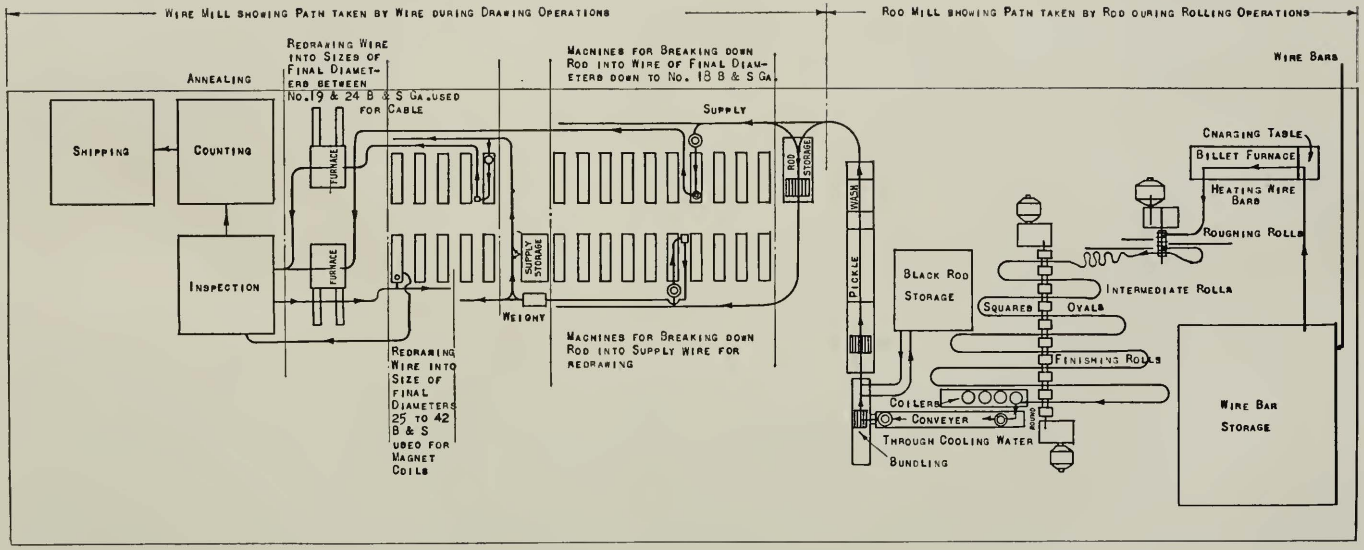


Fig. 1—Schematic layout of Western Electric Company's copper rod and wire mill at Chicago, Ill.

Rod Rolling Mill

The Rod Rolling Mill equipment consists of a billet heating furnace, a roughing mill, an intermediate mill, a finishing mill, coilers, conveyors, and pickling tubs. The mills are water-cooled and equipped with a down-draft exhaust which carries the fumes produced during the rolling operation to an air washer where the copper dust is recovered before the air is discharged.

The 225 pound wire bars as received in cars from the refineries are unloaded onto skids in the train shed and transported by an electric truck to the charging end of the billet heating furnace. Here they are transferred in groups of six by a hoist to the charging table, where a compressed air-pusher moves them along through the furnace which holds 120 bars. The bars are brought up to the required temperature for rolling as they move through the furnace, which is heated by fuel oil. When the bars reach the opposite end of the furnace they are withdrawn at about 1600° F. with a pair of tongs through the discharge door and pushed into the roughing mill one at a time. These tongs operate on a trolley suspended from a beam, which is in line with the first groove of the mill.

The roughing mill consists of three motor-driven rolls, one above the other. The bar, after passing through the first groove between the top and middle roll, drops upon feed rolls set in the floor and is returned through the second groove, between the middle and bottom roll; then raised into position and passed through the third groove, which is in the same rolls as the first pass. Five passes are made in this manner until its cross-section is reduced sufficiently for it to enter the intermediate mill. As the bar enters the roughing mill it is 54 inches long and about 4 inches square. When it leaves this mill it has been rolled into an oval cross-section and is about 124 feet in length. Formerly the last pass on this mill was handled manually, and recently a mechanical repeater has been added as illustrated by Fig. 2.

From the roughing mill the bar goes to the intermediate mill and is passed through the first pair of rolls. As it emerges an operator catches the end with a pair of tongs and passes it back through the next pair of rolls. The increased length between each pass at the intermediate and finishing mills is allowed to run out in a loop on a sloping iron covered floor on each side of the rolls. This catching and returning is repeated at each set of rolls until the original copper bar finally emerges a round, quarter-inch rod about 1200 feet long. This last pass goes through a guide pipe into a coiler, Fig. 3. The reductions in cross-section are illustrated in Fig. 4.



Fig. 2—View of roughing mill showing repeater on last pass



Fig. 3—View of intermediate and finishing mills and coilers

The coils are automatically unloaded from the coilers onto a conveyer, which carries them through cooling water in a tank underneath the floor. An appreciable amount of copper oxide scale is carried off

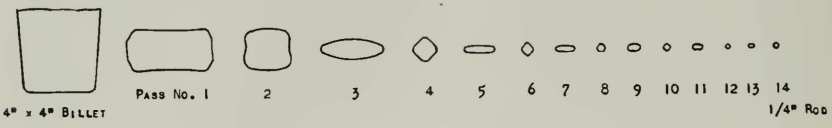


Fig. 4—Rod mill reductions—4 in. by 4 in. billet to ¼ in. rod

with the cooling water, and deposited in a reservoir from which it is later salvaged. Eighty-two seconds after entering the roughing mill the bar is a coil of ¼ in. rod ready to proceed on its way to the pickling tanks. The mill has a capacity of 70,000,000 pounds annually on a 48 hour per week basis.

While the diagram and photograph of the intermediate and finishing mills indicate for simplicity that the rod follows only a single path, in actual operation sufficient material is kept in the mill to practically

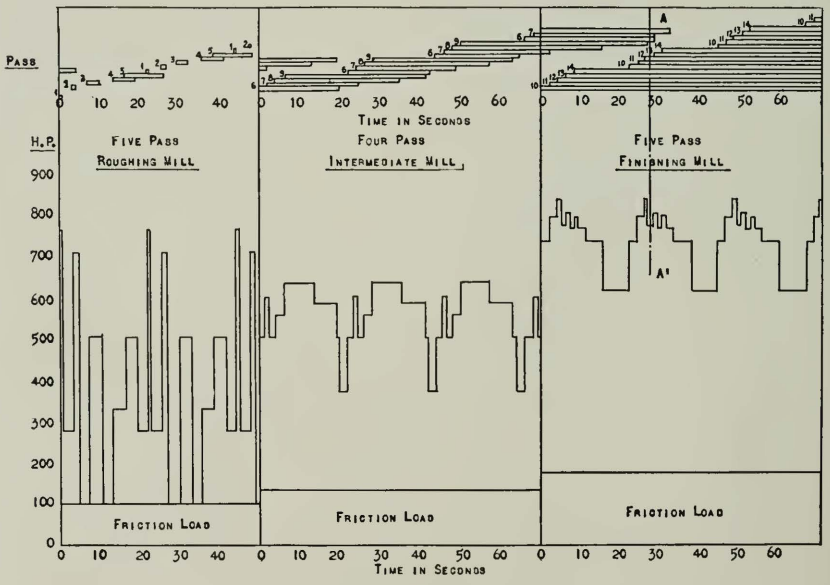


Fig. 5

maintain at least two rods in the finishing mill. This is illustrated graphically by that part of Fig. 5 which covers the finishing mill. Referring to line (A–A'), 11 reductions are being made in this mill at the

same time, two for each of the first four pair of rolls and three on the final rolls. At this period in the cycle of operation 800 H.P. is required.

When the rod mill was started eighteen passes were in use by several of the most modern mills. A sixteen pass arrangement was adopted for the new mill, in which the metal was subjected to a greater amount of working in the earlier passes when it was hot. Later, as a result of further study, fourteen passes were adopted. Fig. 6 illustrates graphi-

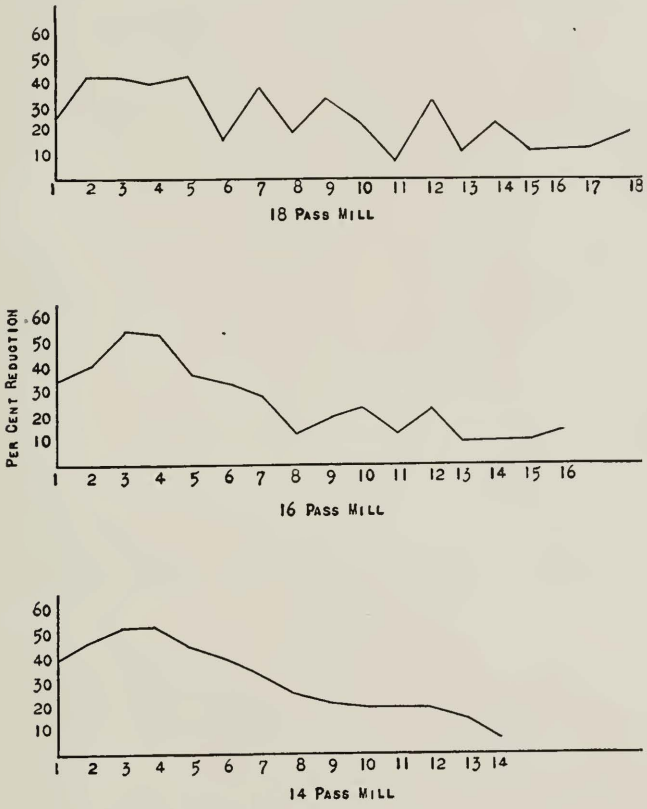


Fig. 6—Per cent reduction

cally the per cent reduction effected at each of the above passes. The reductions plotted as the abscissa are in terms of reduced area in cross-section at each pass and the passes reading from left to right are plotted as ordinates.

It is obvious that careful planning must be done in changing the number of passes in a mill in order not to exceed the safe working capacity of the mill rolls and stands. Such calculations have been made

using rolling mill formulæ.[1] Based on the design of the mill using the eighteen pass arrangement the first four passes would operate at about 62, 100, 105, and 90 per cent of the safe working load of the mill. These same passes calculated on the basis of the sixteen and fourteen pass arrangement of the more sturdy mill at the Chicago plant operate at 86, 87, 90, 85 and 96, 96, 90, 90 respectively. This indicates that a further reduction may be made in the number of passes in the mill provided roll adjustment is improved.

Relation between Working and Physical Properties

It has been often stated that the more passes (i.e. the more gradual working) given the copper, the better the physical qualities of the rod. Actual tests (see Table I) made on representative lots of ¼ in. rod fail to confirm this impression.

TABLE I

Lot	Number of Passes	Elongation	Tensile Strength Lbs. per Square Inch
1	18	35.8%	33,752
2	18	40.0%	31,445
3	16	37.1%	32,468
4	16	41.0%	32,160
5	14	42.0%	32,391
	Average of 5 lots	39.5%	32,243

The averages indicate that fourteen pass rod is superior in elongation, and better than the total average in tensile strength.

Cleaning of Rod

When the coils emerge from the tank through which the rod coiler apron conveyor passes, they are cool enough to handle and after being tied with wire, several are lifted together by a monorail crane, and placed for thirty minutes in a pickling tank containing from 5 to 10 per cent free sulphuric acid, in order to remove the black oxide caused by oxidation of the hot copper in the air during rolling. The solution is maintained at approximately 120° F., and the copper content varies from 1 to 3 grams per 100 c.c. Experiments have shown a difference of less than 10 per cent in pickling time between the minimum and maximum acid used, the greater solubility being obtained from the weak solution. Actual results obtained were checked with Sidell's

[1] "Pass Limitation in Rolling Mill Practice," *Machinery*, July, 1918. "The Theory and Practice of Rolling Steel," Wilhelm Tafel.

Table of Solubilities (see Fig. 7). While a variation from the minimum to maximum acid concentration does not materially affect the pickling

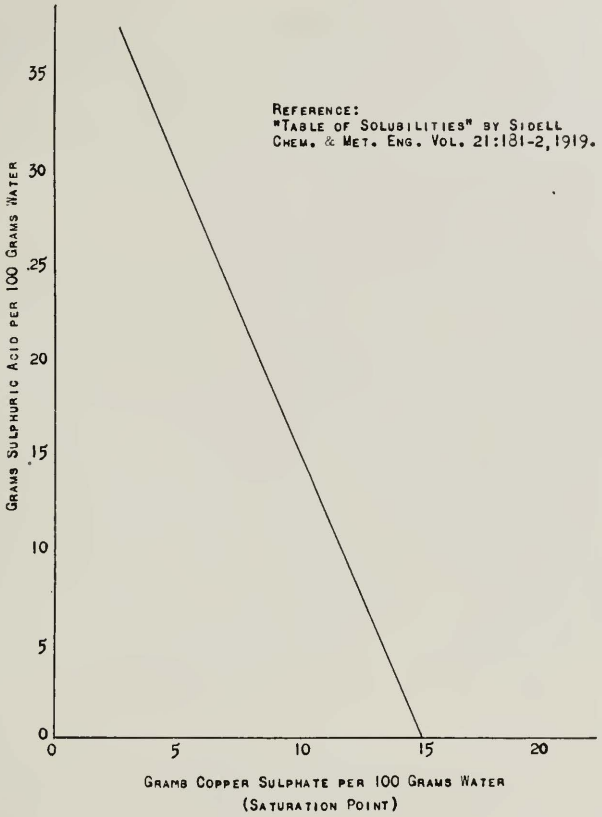


Fig. 7—Solubility curve of copper sulphate in a sulphuric acid solution (Temp. 25° C.)

time, a variation in temperature has a decided effect as may be seen from Fig. 8.

Electrolytic Plant

Figure 9 shows a plant in which the copper is reclaimed from the pickling bath at about the same rate as it is absorbed. This is accomplished by electrolytic deposition according to principles worked out and practiced in the large refineries which produce electrolytic copper.

The electrolytic system operates best with a minimum content of about 1 per cent copper and 5 per cent acid and a maximum of 3 per cent copper and 10 per cent acid. The copper and acid contents are

kept as low as practicable to minimize "carrying out losses" [2] during the pickling operation. About 775 pounds of acid and 430 pounds of copper are recovered per day from the electrolyte. The anodes are operated at a current density of 5 amperes per square foot with a rate of deposition of about .00261 pound of copper per ampere hour.

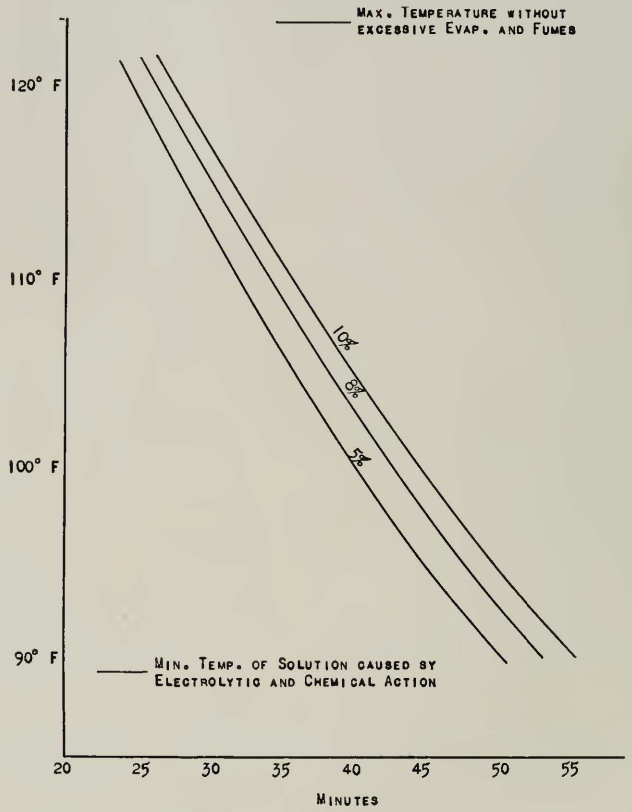


Fig. 8—Rate of pickling at practical acid concentrations

The heat generated in the plating tanks under normal operating conditions maintains a minimum temperature of about 90° F., throughout the acid system, and the maximum temperature is obtained through steam heating coils in the pickle tanks. Faster pickling would result from the use of higher temperatures but experience has shown that the additional steam and gas released above 120° F. results in unsatisfactory operating conditions.

[2] Pickling solution carried out when coils are removed from tank.

The coils of rod after pickling are thoroughly washed with lake water [3] at a pressure of about 70 pounds per square inch to remove loose



Fig. 9—Electrolytic recovery of copper from rod mill pickling solution

copper dust and acid, and then immersed in an alkaline fat solution to neutralize any trace of acid and to provide a protective coating against oxidation until converted into wire.

Wire Mill

The coils, after being pickled and washed, are carried by monorail cranes to the wire drawing machines where they are converted into wire of the desired size. The dies used in the heavy wire drawing machines are pulled into place at the starting end of the coil of rod on a die stringing machine (Fig. 10). The coil, with dies strung into position, is then placed in a heavy wire drawing machine.

The heavy gauges of wire, such as line wire, are drawn with one set-up on this machine; medium sizes, used in lead covered cable, are made by taking the wire as it comes from the heavy machine and re-

[3] Lake water is relatively free from mineral salts which would corrode the rod and affect the wire drawing compound.

drawing it on the intermediate machine; and finer sizes, commonly known as magnet wire, are produced by redrawing intermediate sizes.

The present capacity of the wire mill is approximately 42,000,000 pounds annually, and the sizes range from .165 in. line wire to 42 B. & S. (.00247 in.) gauge magnet wire. Provisions have been made in the construction of the building and its foundations so that the mill may be expanded in capacity when needed.



Fig. 10—Heavy wire die stringer

The No. 1 or heavy wire drawing machine shown by Figs. 11 and 12 draws line wire, heavy toll cable sizes, and supply wire for the loop cable wire machines. This ten die machine with its auxiliary equipment and operating area occupies a floor space of 270 square feet and



Fig. 11—Battery of No. 1 wire drawing machines

runs at 1500 to 2000 feet per minute as compared with 470 square feet for a commercial nine die machine running about 1000 feet per minute.

A battery of these machines costs much less than an installation of commercial machines of the same capacity, and in addition effects a considerable economy in floor space.



Fig. 12—Close-up of No. 1 machine

The commercial types of ten die intermediate machines for drawing cable wire require about 130 square feet of floor space as compared to 90 square feet for a twelve die multiple head machine. The former is a single unit machine and the latter a four unit machine operating at twice the speed and capable of producing about five times the output of the commercial equipment. This new multiple unit machine, Fig. 12A, costs more than regular equipment, but considering the four units, the cost is materially less per unit, and very much less on an output basis.

The magnet wire drawing machine is a high speed twelve die multiple head machine of eight wire drawing units occupying 90 square feet of floor space including the operating area. A close-up view of two heads of this machine is shown by Fig. 13. Fifty-one square feet of



FIG. 12A.

floor space are required for a single unit (one head) commercial machine of the same die capacity. The saving in investment in this case is even greater than for the heavy and intermediate types of machines. The use of these compact machines and overhead monorail equipment for transporting material instead of using trucks with large aisles has permitted the installation of the wire drawing mill in less than one fourth of the building area which would have been required if commercial equipment had been purchased.

GENERAL PLANT FEATURES

The present connected load of the motors in the Rod and Wire Mill is about 6000 horse power for which it was necessary to enlarge the main power plant. A 700 foot tunnel connects the power plant with the Rod and Wire Mill in which are laid pipes for carrying hot and cold water, steam, gas, and air and lead covered power cables.

The basement under the Rod Mill houses the electrolytic equipment, control boards for the roughing and intermediate mills, pumps for cooling water, and exhaust fans connected with an air washer for removing the fumes from the Rod Mill. A tunnel which passes beneath the intermediate and finishing mills connects with a room which houses the drives for the four rod coilers, the coiler control boards, the finishing mill control board, and the main power panel. In the wire mill base-



Fig. 13—Close-up view of units of No. 3 wire drawing machine

ment are six large tanks which hold the compound used to lubricate and cool the wire drawing dies. This compound is supplied under pressure to the wire drawing machines on the floor above and returns by gravity.

All the wire drawing machines are controlled by push buttons mounted on the machines, which connect with compensators in the basement. The 100 horse power motors driving the large wire drawing machines are mounted in a tunnel and are connected to the machines above by chain drive.

This arrangement permits accessibility for maintenance of the electrical equipment with a minimum of interference to production, prevents the wire drawing operators from having access to the electrical equipment, and reduces accident hazard to a minimum.

Developments in Wire Drawing Equipment and Methods

The Rod and Wire Mill just described was designed following a comprehensive survey of wire drawing processes and equipment used in this country and abroad. In connection with these studies, extensive laboratory investigations were undertaken relative to the characteristics of different types of commercial machines especially from

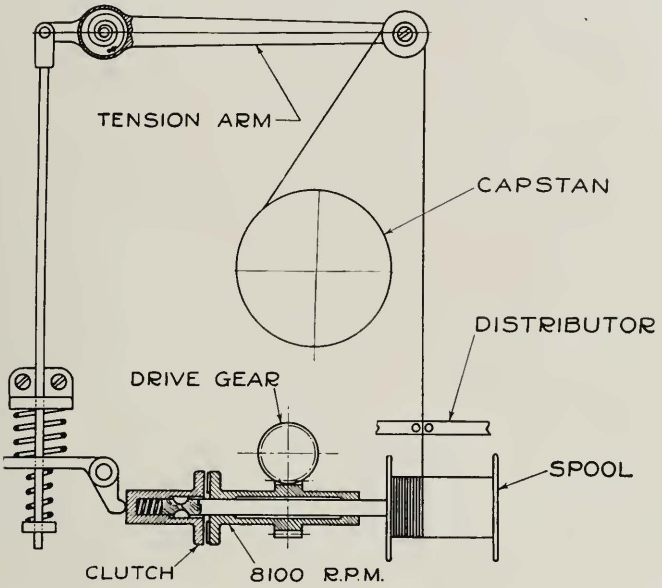


Fig. 14—Automatic tension mechanism—No. 3 wire drawing machine

the standpoint of operating efficiency, investment, and floor space requirements. As a result of these investigations, it developed that marked improvements could be effected if wire could be produced commercially at higher machine speeds and with more compact machine equipment.

While the design of the drawing mechanism in the new machine was very important, it was also deemed essential that the finished wire be taken up on spools instead of coils. After considerable experimental work, a sensitive take-up device was developed to permit spooling at a constant drawing speed.

This spooling mechanism is illustrated by Fig. 14 in which the spool spindle is driven by a slipping clutch member controlled through a tension arm, on which an idler pulley is located over which the wire passes on its way from the drawing capstan to the take-up spool. The

take-up mechanism rotates the core of an empty spool at a speed synchronous with the speed of the wire as it leaves the drawing capstan. As the spool fills and the speed tends to increase, the wire on the tension arm tightens and compresses the tension arm against a spring adjusted for the proper gauge of wire. This in turn reduces the pressure of the clutch driving the take-up spindle, permitting the spool of wire to readjust its speed.

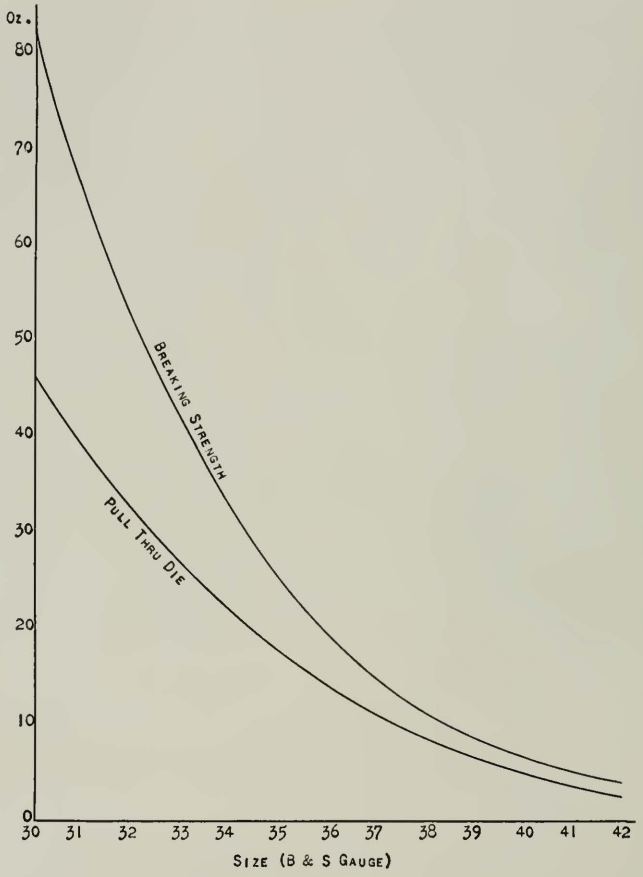


Fig. 15

This device is extremely sensitive as illustrated in the drawing of No. 42 B. & S. wire at 2000 feet per minute, in which case the control arm must be adjusted to operate between 90 and 150 grams, since the pull required is 87 grams and the breaking strength of the wire is 170 grams. This device is so flexible that it can be adjusted to a drawing

tension of from 9 pounds for No. 25 wire to 3 ounces for No. 42 wire. Fig. 15 illustrates its operating range on wire sizes No. 30 to No. 42, showing the gradual narrowing of the limits as the sizes decrease. A larger machine used for drawing loop cable wire from No. 18 to No. 30 B. & S. gauges contains a similar mechanism.

The use of this sensitive device and a clutch which would slip without overheating as the spool filled, together with improvements in the wire drawing compound and the shape and quality of the diamond dies later described, permitted the drawing of wire at speeds ranging from 2000 to 3000 feet per minute.

Wire Drawing Compound

At low speeds it was discovered that the compound for lubricating wire drawing dies required little attention but as the speeds were increased the necessity for close analytical control was evident. The compound consists of an emulsion of soap, tallow, and water, the percentage of the soap and tallow being varied depending upon the size of wire and type of machine on which it is used.

It is important that the degree of emulsification [4] be carried far enough to break the tallow into particles about one micron in diameter, so that the material will stay in suspension in the water. If the tallow content is increased beyond a certain point, it holds in suspension in the solution a large amount of the copper dust which flakes off in a very fine state during the wire drawing operation and this clogs the dies and causes breakage during the wire drawing. Ordinarily this copper dust settles out of the solution while in the large cooling tanks and a considerable amount is salvaged in this manner.

Effect of Drawing on Copper

Tests were made to determine if the drawing of the smaller cable and all magnet wire sizes [5] in Brown and Sharpe (A.W.G.) steps was yielding the maximum reduction possible per die. These tests showed it was feasible to make much heavier than A.W.G. reductions at the first draft when annealed wire or soft copper rod was being drawn. It also showed that the elongation [6] of the rod or annealed wire was rapidly reduced to the drawing minimum after the first pass, and remained at that point throughout the process.

[4] "The Theory of Emulsions and Emulsifications," W. Clayton.

[5] A.W.G. ("American Wire" or "Brown and Sharpe" gauge) reductions are not used in converting the rod to line wire; these are generally specified in B.W.G. and N.B.S. gauges.

[6] See Figs. 16, 17, 18, and 19 showing the elongation of the rod or wire dropping to about 1½ per cent at the first die reduction and remaining practically constant.

Figure 16 illustrates the effect of a five die reduction on elongation and tensile strength. It may be seen that the elongation drops very rapidly at the first die when a reduction in area of about 42½ per cent is made, and the tensile strength increases rapidly because of the cold working of the metal.

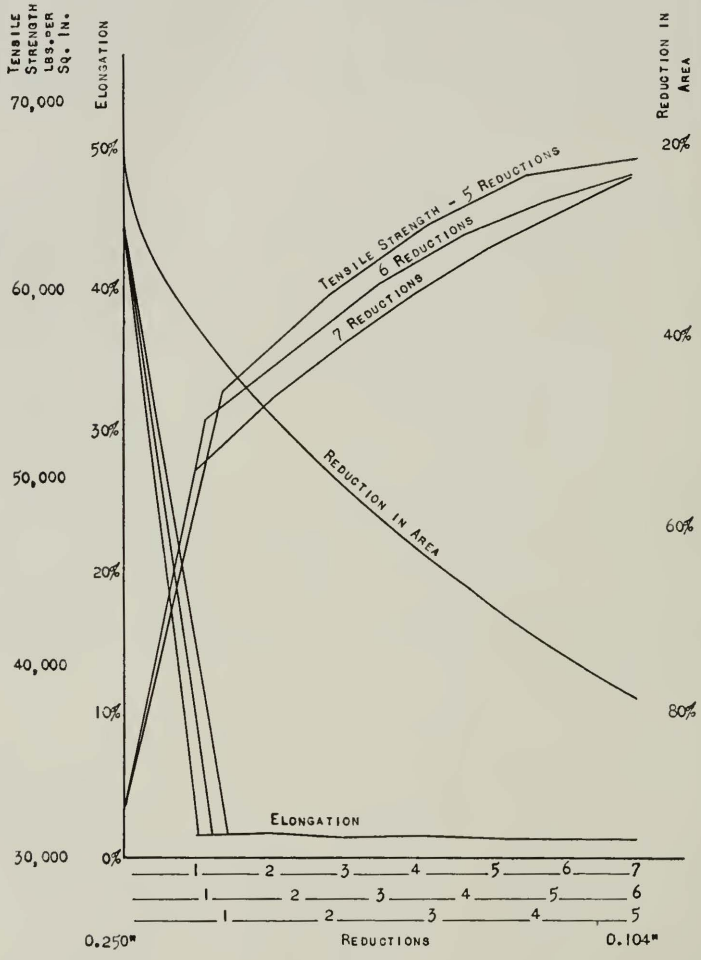


Fig. 16

This same figure shows the tensile strengths obtained when five, six, and seven die reductions are used to produce line wire of .104 diameter from the same supply. Here the elongation loss is about the same in each case, but the tensile strength is greater with the heavier

reductions. The five die arrangement is satisfactory according to the results shown on the curve, but the heavy reduction at the first die often results in rough or slivered wire. The six die arrangement, therefore, gives the greatest factor of safety. The seven die arrangement is less satisfactory since the elongation and tensile strength in the finished wire are so close to the requirements.

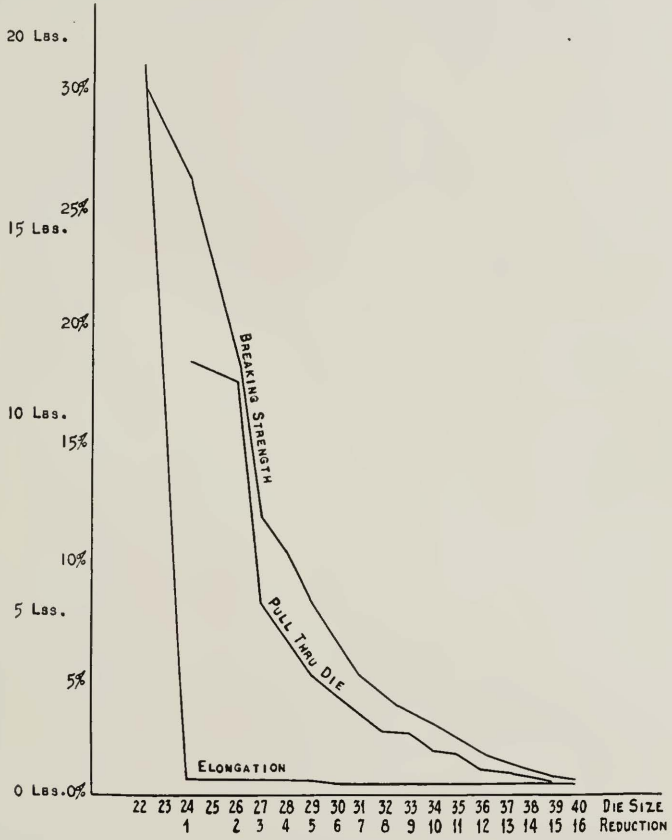


Fig. 17

The use of A.W.G. reductions for the finer sizes of cable and magnet wire provides flexibility since a change in the size of wire can be accomplished simply by increasing or reducing the number of dies used. Tests were conducted to determine the gain by using heavier reductions and annealing the wire before redrawing, and Fig. 17 shows the increased reduction possible at the first die when the metal is plastic. In this test, an annealed No. 22 gauge wire of 31 per cent elongation

was reduced to No. 24, two gauges, in one draw. The soft copper permitted a double reduction at the first die, but the elongation dropped during the operation to less than 1 per cent; the second reduction on this test was from No. 24 to No. 26 gauge and the pull required for this pass practically coincides with the breaking strength of the wire. Wire drawing under such conditions is impractical because the annealing operation is much more expensive than drawing hard wire from No. 22 to No. 24 in two passes.

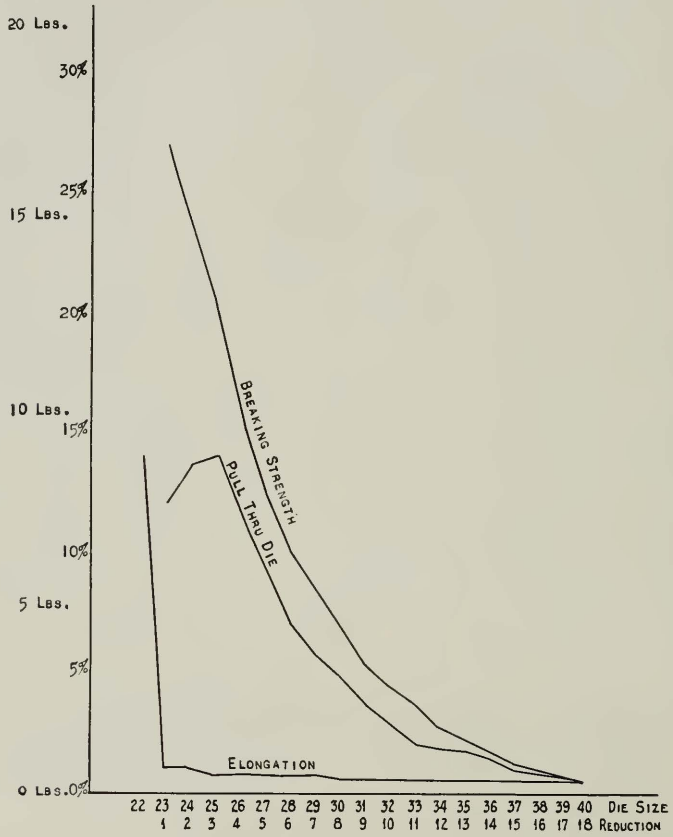


Fig. 18

Figure 18 illustrates the results obtained when drawing annealed wire with A.W.G. reductions. The large margin of safety between the pull required and the breaking strength of the material again disappears after two reductions. Fig. 19 [7] illustrates practical drawing

[7] Slight irregularities in the curves are due to variations from the mean in the diameters of the dies used during the test.

conditions adopted for drawing wire to finished sizes without annealing during the process.

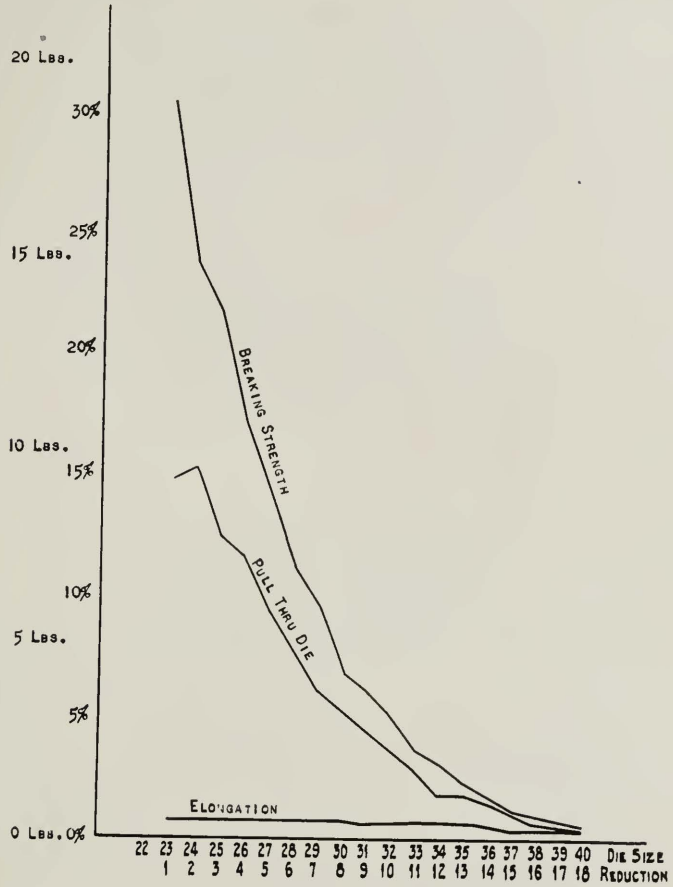


Fig. 19

Chilled Iron Dies

The dies used for drawing heavy wire are cast with a tapered hole from chilled cast iron and reamed to the desired size. When the die wears too large for a particular size of wire, it is reamed to a larger size and used in that manner until the die goes above the maximum size used. These dies, shown in Fig. 20, are used for drawing line and heavy gauge wire for which the cost of diamond dies would be excessive. Many alloy steel dies have been tested as substitutes for chilled iron dies for copper wire drawing, but so far have failed to replace them, due

to excessive cost. For the wire sizes smaller than No. 16 down to as fine as No. 42 B. & S., diamond dies as described below are used.

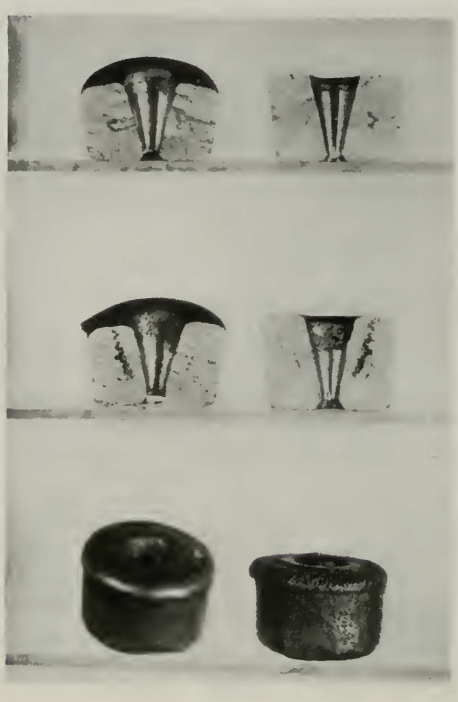


Fig. 20

Diamond Die Study

It was necessary to make an extensive study of the manufacture of diamond dies because dies through which wire could be satisfactorily drawn at low speeds failed to draw to gauge and without excessive breakage of the wire as the speeds were increased. At this time practically all commercial diamond drilling was done in Europe, Belgium being the hub of the diamond cutting industry, and the art was new to this country. The diamonds generally used for wire drawing dies are obtained from South Africa,[8] Australia, and Brazil, and made into diamond dies in Europe.

[8] The South African and Australian diamonds are the more suitable for wire drawing. There are two types of the former, the smooth brown premier which is not suitable for dies because of its tendency to crack and split, the other commonly known as the Jager, a product of the Jagerfontein mines. These stones, very irregular in contour and light gray to black in color, are most suitable for dies. The Australian diamonds are gray to brown to almost black in color and can be distinguished from the Jager. Many of the Brazilian diamonds are a dark gray similar to graphite in color and not being translucent are difficult to inspect for seams, cracks or inclusions.

In view of the difficulty of obtaining dies for drawing wire at high speeds and the large investment in dies required for the proposed wire mill, it was decided to undertake a laboratory investigation of the

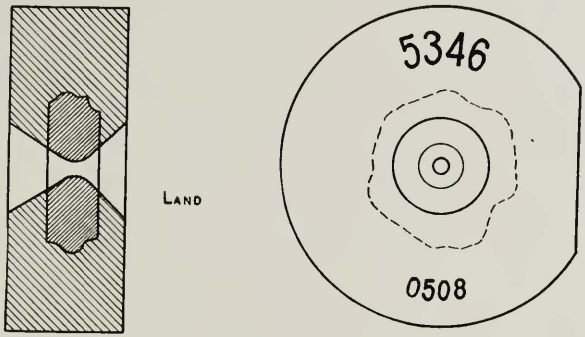


Fig. 21—Diamond wire drawing die (outline sketch)

manufacture of diamond dies suitable for drawing cable and magnet wire.

It was found that the dies suitable for high speed wire drawing required a differently shaped approach, a better polish, and a shorter land [9] than used for low speed drawing. In addition, the origin of the

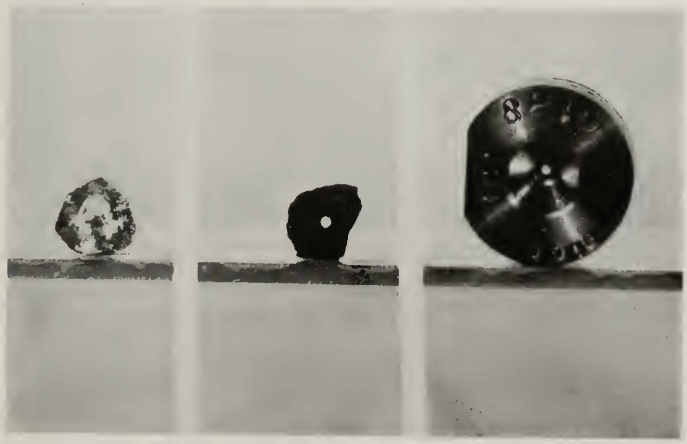


Fig. 22

stone, the shape of the diamond, and its setting are all very important because of the internal strain to which the die is subjected during the drawing operation.

[9] See Fig. 21.

It has not been possible to definitely establish any quantitative relationship as to the effect of high speed drawing on the wear of dies except that about the same number of million feet of wire may be expected from a properly lubricated die irrespective of the drawing speed. Under such conditions, the high speed die naturally runs a shorter time, but length of life is not the important factor; tonnage of a satisfactory quality with a minimum plant and labor investment is the prime consideration.

Figure 22 shows a diamond before drilling, a stone drilled and lapped, ready for mounting, and a die in the final mounting ready for use.

Figure 21 gives an outline of the shape of the working surfaces of a wire drawing die.

Annealing

Hard copper wire is obtained by using the wire as it comes from the wire drawing machine. This same wire may be softened by annealing, or medium-hard wire can be produced by annealing hard wire at such a point in the drawing operations that the final draws will give the desired degree of hardness.[10]

In a recent commercial type of annealing furnace, Fig. 23, wire may be bright annealed, but it requires a drying operation in order to remove the water through which it passes in leaving the furnace. The retorts of these furnaces are water-sealed and filled with steam to exclude the outside atmosphere, which would discolor hot copper. To obtain bright wire, it is passed under water into the retort to exclude the air and is generally taken out and cooled under water or in an atmosphere of steam or gas, which excludes oxygen until the wire is relatively cool.

A special steam seal annealing furnace for small spools of wire was developed on an experimental basis from which the wire was obtained bright annealed and free from moisture. In this furnace the spools were submerged in water to displace the air, raised into the charging end which was under water, thence to the muffle to be heated, and then along a cooling tube to the discharge opening. Air was excluded from the retort and cooling chamber at the discharge end by means of a steam jet.

The success of the small furnace led to the construction of a larger machine (Fig. 24) for annealing cable wire on spools. The spools are placed in perforated metal baskets which are charged into the furnace at a specified time interval, pushing each other through the retort and along the cooling tube to the discharge end.

[10] "Experiments in the Working and Annealing of Copper," F. Johnson, *Journal Institute of Metals*, Volume XXVI, No. 2, 1921.

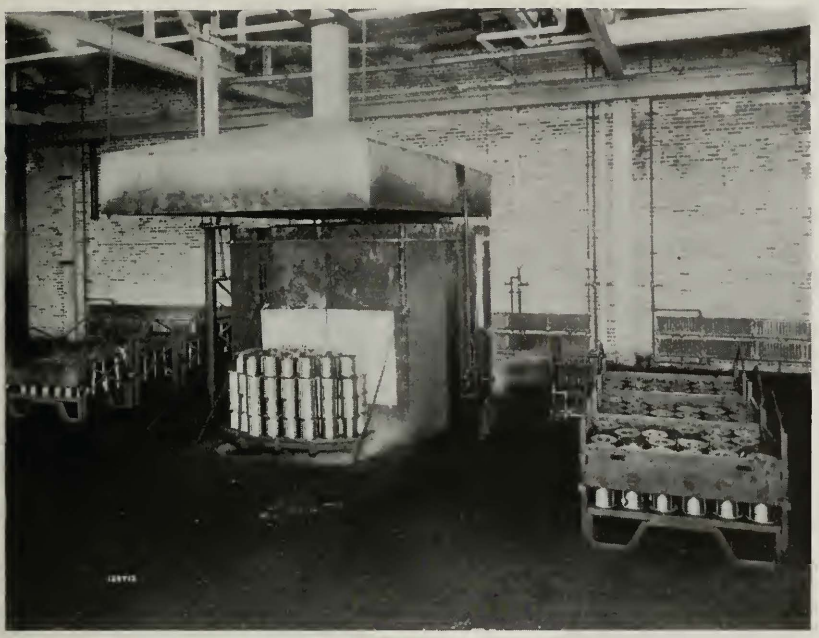


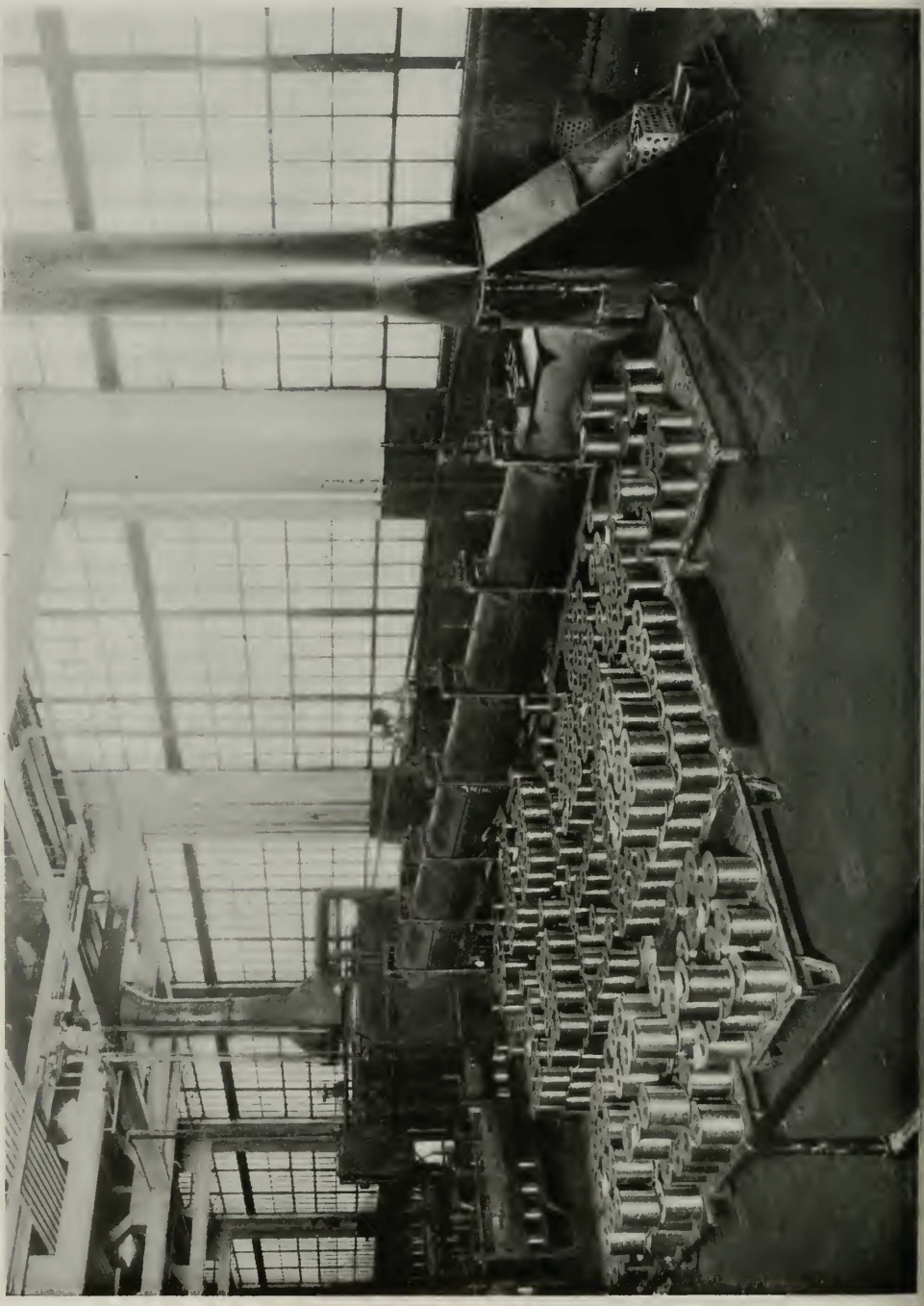
Fig. 23—Water seal annealing furnace

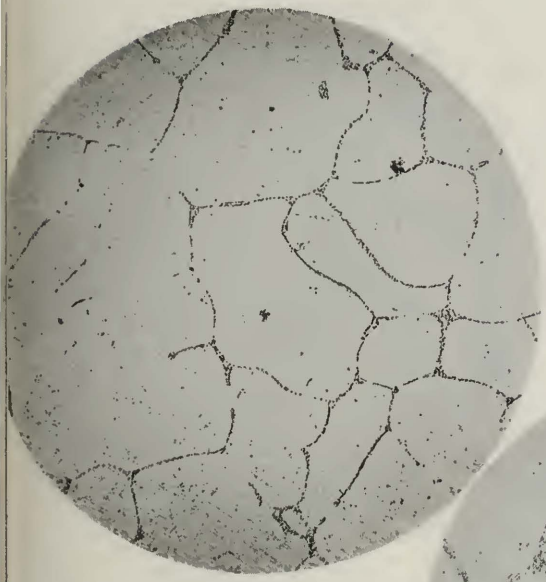
Inspection of Raw Material and Finished Product

Wire bar made from electrolytic refined copper is used as a material in the manufacture of wire. This material is practically free from silver and other elements which ordinarily exist in the ore, and which have a detrimental effect on the electrical or physical properties of the finished product. A small percentage of silver [11] seriously affects the annealing qualities of the wire. Traces of other impurities have a very detrimental effect on the wire drawing properties. During the refining process, the molten bath is oxidized in order to carry off the foreign material in the form of slag, and it is very important that the oxygen content be later reduced to a very small point if bars of proper set are desired. Fig. 25 shows three photomicrographs of wire bar containing varying amounts of cuprous oxide.[12] Ordinarily the surface condition on top of the bar is a good index of the oxygen content. If the bar is level set or slightly convex on top, it is usually a satisfactory material. If it is low set or concave, it usually contains a large amount of copper

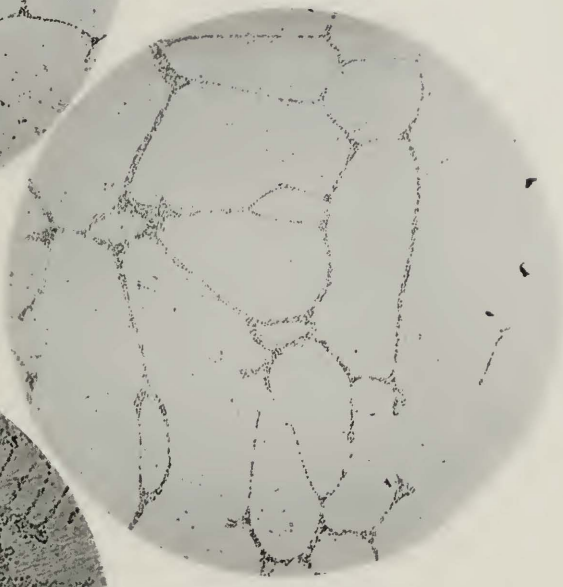
[11] "Effects of Silver on the Recrystallization Temperature of Copper," Caesar and Gerner, *A. S. M. E.*, Volume 38, 1916.

[12] "Microscopic Structure of Copper," H. P. Pulsifer, *Mining and Metallurgy*, January, 1926.

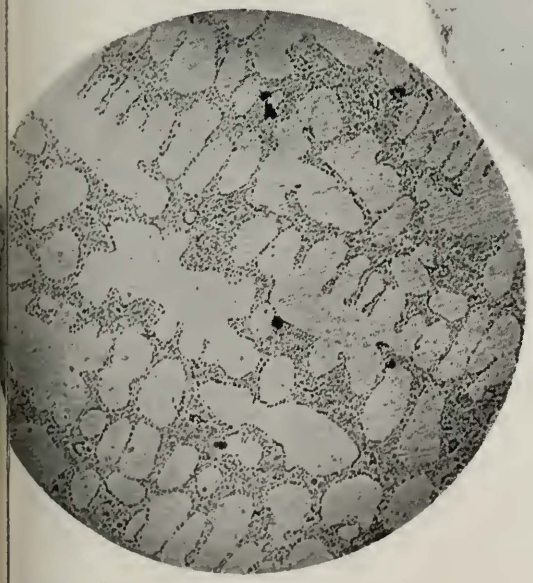




A. High Set—Oxygen .035%



B. Level Set—Oxygen .050%



C. Low Set—Oxygen .12%

Fig. 25—Photomicrographs of wire bar (magnification × 100)

oxide, which caused the metal to shrink in solidifying.[13] When excessive shrinkage occurs it has an adverse effect during the rolling operation.

The finished wire is inspected for dimensional limits, tensile strength, elongation, and surface condition. The limits for 42 B. & S. gauge wire (.002475 in.) are .00245 in. minimum and .0025 in. maximum.

Conclusion

The establishment of this industry as a part of the plant at Chicago represents the combined effort of a large number of inventors, engineers, designers, and mechanics. While the actual plant was built within a comparatively short period, the advances which have been made in the art represent several years' effort. Briefly, the development of com-

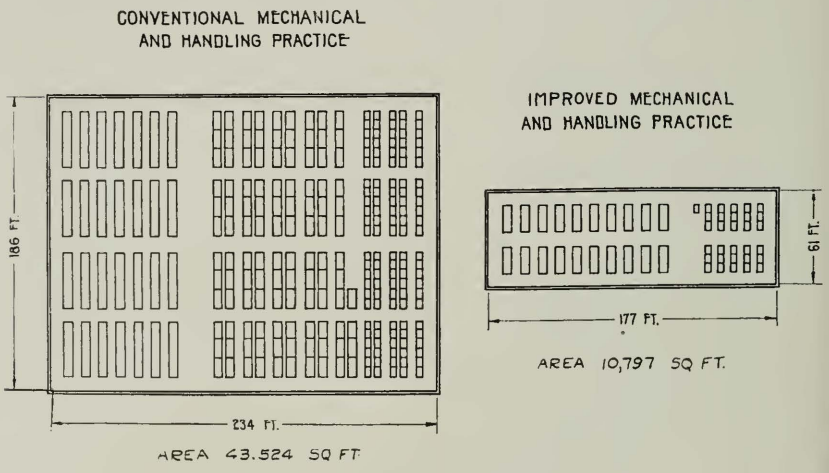


Fig. 26—Wire drawing plant

pact and high speed wire drawing machines has required a much smaller investment in buildings and equipment as compared with a plant of the same capacity using commercial equipment. A comparison of the relative floor area, based upon the conventional and the improved types of wire drawing equipment, is illustrated by Fig. 26. The supervisory force in charge of the operation of this new mill must be given a considerable share of the credit for its successful operation.

[13] "Copper Refining," Lawrence Addicks. "Metallurgy of Copper," H. O. Hofman.

An Analyzer for the Voice Frequency Range

By C. R. MOORE and A. S. CURTIS

[Editorial Note: The frequency analyzers described in this paper and in the paper immediately following, demonstrate in an unusual manner how a single fundamental principle may be employed to accomplish quite dissimilar results. The analyzers described in both papers employ a resonating element of fixed frequency and translate the wave components under study to this frequency by heterodyning them with the output of variable frequency oscillators. In the analyzer described in the first paper, the wave components under study are translated to a higher frequency while in that described in the second paper the translation is downward to a lower frequency. In view of these differences in design it is desirable to call particular attention to the reasons which have led to the working out of the two designs.

The analyzer discussed by Moore and Curtis has been so designed as to sweep through the voice frequency range to as high as 5,000 cycles by the manipulation of a single control. To accomplish this end, it was found desirable to heterodyne upward by employing a variable frequency oscillator of considerably higher frequency than 5,000 cycles. The frequency of this oscillator can be varied continuously throughout the range from about 11,000 cycles to 16,000 cycles, and the fixed frequency resonating element is tuned to about 11,000 cycles. As translation of the wave under study to a higher frequency range reduces the percentage separation of the various components, it was necessary to choose a very sharply tuned resonating element. This takes the form of a steel rod which is loosely coupled magnetically to a driving circuit at one end and a registering circuit at the other. As the modulator used to accomplish the heterodyning process produces many frequencies other than the first of the "sum" and "difference" terms, it has been necessary to choose the frequency ranges such that all undesired frequencies which can not be made extremely small will be well removed from the single difference frequency under observation.

The analyzer described in the paper by Landeen is capable of working over the range from about 3,000 cycles to 100,000 cycles. A requirement of this design was that very high resolution be obtained. To assist in accomplishing this end, the frequencies under study are translated downward in the frequency scale to the resonating element which consists of a circuit tuned to 800 cycles. This downward translation increases the percentage difference of frequency separation of the components under study. Because of the great range of frequencies covered by the analyzer it is not possible to have sum and difference terms other than those of the second order fall outside of the range of sensitivity of the resonator. The modulator has therefore been so designed as to preclude formation in the higher order terms. To increase its discrimination, the analyzer makes use of two tuned circuits and amplifiers arranged in tandem and placed before the modulator. The frequency to which these circuits are tuned must of course be variable and is set to coincide with the component under study.]

THE present analyzer was designed to aid in the solution of certain problems arising in the study and development of commercial telephone transmitters. These problems require high discrimination and the accurate measurement of frequency components in the presence of much larger components.

The present analyzer differs fundamentally from that described by R. L. Wegel and one of the present authors about two years ago.[1]

[1] "An Electrical Frequency Analyzer," by C. R. Moore and R. L. Wegel, *A. I. E. E. Journal*, September 1924; *Bell System Technical Journal*, October 1924.

It is of the type employing a single resonating element of fixed frequency, the component waves under study being translated in frequency to it by heterodyning with the output of a variable frequency oscillator. It was deemed essential that the analyzer be so designed as to cover the entire voice range up to 5,000 cycles by the manipulation of a single control. In the presenta analyzer this control is the variable condenser of the heterodyning oscillator. Two of these analyzers have been built and are in use in the Bell Telephone Laboratories.

The development of this new form of analyzing device was undertaken only after a careful review of existing types. Such a study led to the conclusion that none of the available forms was applicable to the solution of our type of problem. This problem which had to do

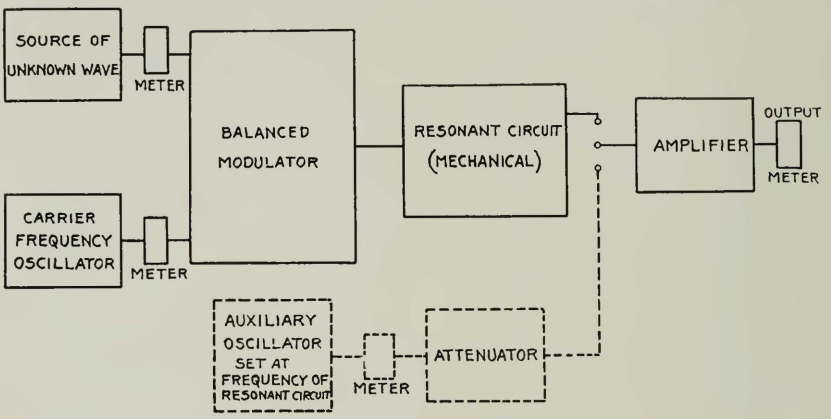


Fig. 1—Functional Diagram

with the study of commercial telephone transmitters did not require the analysis of a complex wave containing a fundamental frequency and its associated harmonics. What we were interested in was the measurement of the transmitter output at a particular frequency in the presence of a much greater output at other frequencies. For example, the measured component may be only one per cent of the magnitude of the disturbing component and separated from it by only 30 to 40 c.p.s. It is evident that if the frequency of the component which we wish to measure is close to that of one of the disturbing components and much smaller in magnitude a very high degree of frequency discrimination on the part of the analyzer is necessary.

A functional diagram of the analyzer is shown in Fig. 1. The wave to be analyzed is impressed at the "voice" input terminals of a balanced modulator. The variable carrier frequency is also supplied to the

modulator and in conjunction with the wave to be analyzed produces the familiar upper and lower sideband frequencies in the modulator output.[2] The modulator output is connected to a resonant element, the natural frequency of which is left unchanged during the analysis, and the response of the resonant element is measured by a suitable amplifier and a meter. The tuned element of the analyzer makes use of longitudinal vibrations in a steel bar which in the apparatus herein described has a natural frequency of approximately 11,000 c.p.s. The frequency range of the carrier oscillator is from the natural frequency of the resonant element to 5,000 c.p.s. above. It will readily be seen that if, for instance, there is a 1,000 c.p.s component in the unknown wave and the carrier frequency oscillator is set at a frequency of 1,000 c.p.s. above the natural frequency of the resonant element, the lower sideband or difference component from the modulator will be at the frequency of this resonant element. The process of analysis is then to vary the frequency of the carrier oscillator gradually and to determine the output from the resonant element at each desired frequency. Inasmuch as the frequency range of the carrier oscillator is relatively small, the entire variation of frequency can be accomplished by a single air condenser. Therefore, the frequency setting of the analyzer may be varied continuously instead of in discrete steps. The frequency calibration chart of the carrier oscillator is arranged to show the input frequencies of the unknown wave to which a given frequency of this oscillator will correspond rather than to indicate the frequency of the oscillator itself.

It has been found convenient, for comparing the magnitudes of the various frequency components in the unknown wave, to use an auxiliary oscillator supplying current to a potential attenuator. The frequency of this oscillator is maintained at the frequency of the tuned element. The input terminals of the amplifier may be connected either to the output of the resonant circuit or to the output of the potential attenuator. The procedure is to note the deflection on the output meter of the amplifier produced by the resonant element and then switch the amplifier to the attenuator and to reproduce this deflection. The frequency components can readily be compared in this manner and their relative magnitudes determined directly in TU by reference to the attenuator dial setting.

In order to give a clearer idea of the operation of the analyzer the pertinent theory of the vacuum tube modulator will be discussed in the Appendix.

[2] *A. I. E. E. Journal*, April 1921—"Carrier Current Telephony and Telegraphy," by E. H. Colpitts and O. B. Blackwell.

The modulator used employs two vacuum tubes and its circuit is arranged to suppress the carrier frequency together with certain higher order modulation components in the modulator output. These together with other higher order modulation components which are not eliminated in this type of balanced modulator could produce false indications of frequency components in the wave to be analyzed, but it will be shown in the Appendix that these errors may be reduced to any desired extent by keeping the magnitude of the wave to be analyzed as low as is consistent with securing satisfactory meter readings.

The suppression of the carrier wave is desirable in that it makes it possible to carry the analysis to lower frequencies than could be done if the carrier frequence were present. When analyzing low frequencies, the frequencies of the carrier wave and the lower sideband approach each other and if the relatively large carrier were present in the output of the modulator it would tend to obscure the results.

It will readily be seen that since the resonant frequency of the tuned circuit is 11,000 c.p.s. and since the frequency discrimination at low frequencies depends upon the sharpness of resonance of this circuit, extremely sharp tuning is necessary. The considerations here differ somewhat from the ordinary considerations in tuned circuits where the effect of the resonance depends upon a percentage departure from the resonant frequency. The reactance-resistance ratio, Q, which is in common use in the treatment of electrical circuits, gives a measure of what we may call the percentage sharpness of tuning of a circuit, that is, with a given value of Q, a given percentage departure from the resonant frequency will cause the same loss independent of the resonant frequency. While it is possible at frequencies from 10,000 to 20,000 c.p.s. to obtain higher values of Q than at frequencies from 100 to 1,000 c.p.s., it is not feasible to secure the same loss with a given departure in cycles from the resonant frequency at high frequencies as it is at low frequencies. It is evident that in this method of analysis we are not concerned with a percentage departure in frequency from the resonant frequency of the tuned circuit but are concerned with the loss in transmission through this circuit per cycle departure from the resonant frequency. Therefore, inasmuch as we would desire to have good discrimination between a frequency of 100 c.p.s. and one of 110 c.p.s., the requirements of the 11,000 c.p.s. resonant circuit are extremely rigid.

Some consideration was first given to the design of an electrical network which would give sufficiently sharp tuning. At best, such a network required a considerable number of coils and condensers and these coils would require higher values of Q than could be obtained economi-

cally. Moreover, inasmuch as this selective circuit would consist of a number of highly resonant elements, it would be rather questionable whether these elements would all be affected alike by ordinary variations in room temperature. Some experiments were then made using mechanical resonance and these have given a very satisfactory solution of the problem.

The resonant element now in use consists of a steel rod clamped at the center having the magnetic element of a telephone receiver at each end with its poles separated a few mils from the end of the rod as shown in Fig. 2. One of the receiver units is connected to the output of the modulator and is used for driving the bar while the other receiver unit

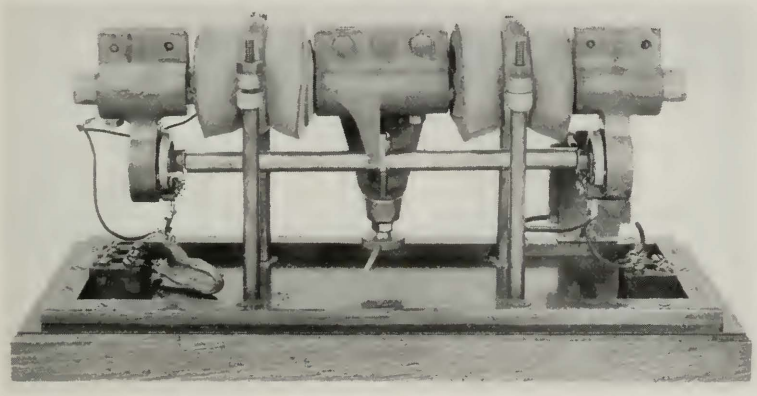


Fig. 2—Mechanical resonant circuit

is connected across the input of a suitable amplifier having a thermocouple and meter at its output. In order to minimize the effect of other extraneous frequency components the amplifier is tuned to have its maximum efficiency at the resonant frequency of the bar. The bar is approximately 9 inches long and resonates to longitudinal vibrations at 11,350 c.p.s. A frequency response curve of the bar, showing variation of the output of the driven receiver with frequency, is shown in Fig. 3. It will be seen that a departure of 10 c.p.s. from the resonant frequency gives a loss of over 25 *TU* corresponding to a voltage ratio of approximately one to twenty. Therefore, even when frequencies in the unknown wave are as low as 50–100 c.p.s. the frequency discrimination is quite satisfactory. It may be of interest to note that the value of the reactance-resistance ratio Q, calculated from the curve, is about 15,000 whereas the construction of an electrical inductance to operate at this frequency having a value of Q over 200 would be difficult and expensive.

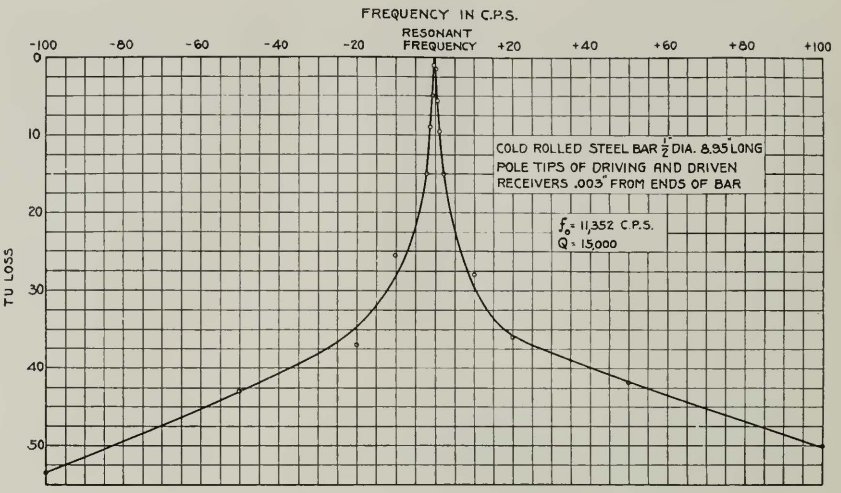


Fig. 3—Response-frequency characteristics of resonant bar

One of the two heterodyne frequency analyzers is built as a self-contained unit mounted on wheels and contains its own "A" and "B" batteries for the vacuum tubes. Fig. 4 shows the top view of the panel where the necessary controls and meters are situated and Fig. 5 shows the complete assembly with some of the compartments open. A schematic diagram of the complete circuit is shown in Fig. 6.



Fig. 4—Top view of panel

In addition to its use for determining the relative magnitudes of the components in an electrical wave, the analyzer is also useful in making acoustic measurements. It was primarily developed for the study of what we may term the frequency distortion of transmitters; it may also be used in the measurements of non linear distortion. As may be

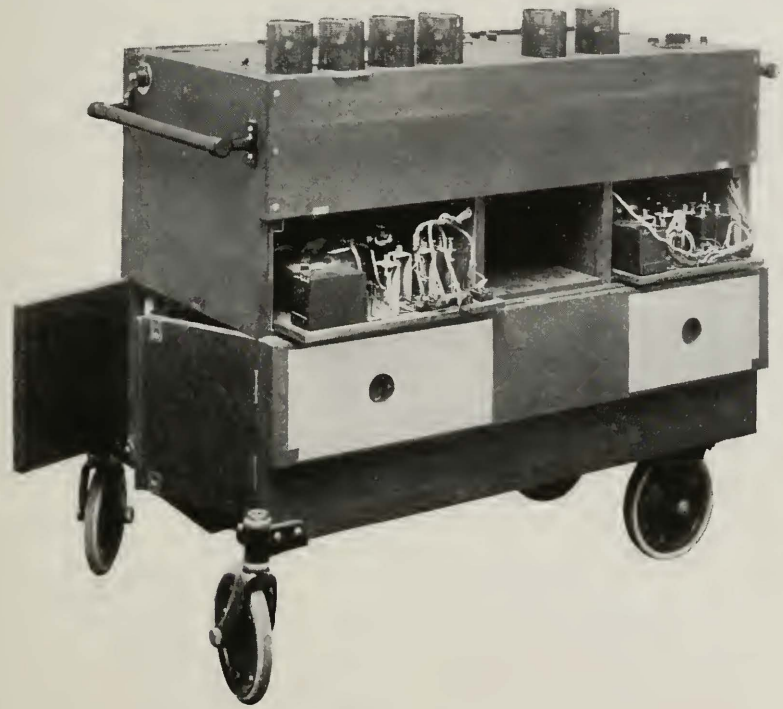


Fig. 5—Portable analyzer unit

inferred, frequency distortion is a departure of the wave form of the electrical output from that of the acoustic input due to the fact that the transmitter does not respond equally to acoustic forces of the same magnitude over the frequency range under consideration. Non-linear distortion is the distortion produced by the fact that the voltage across the transmitter at any particular frequency is not a linear function of the magnitude of the impressed acoustic force. This type of distortion usually manifests itself by the production of frequency components in the electrical output which are not present in the acoustic input. The analyzer may be used for quantitative determinations of this non-linear distortion and in such measurements high frequency discrimination and ability to measure frequency components of widely different magnitudes are very valuable.

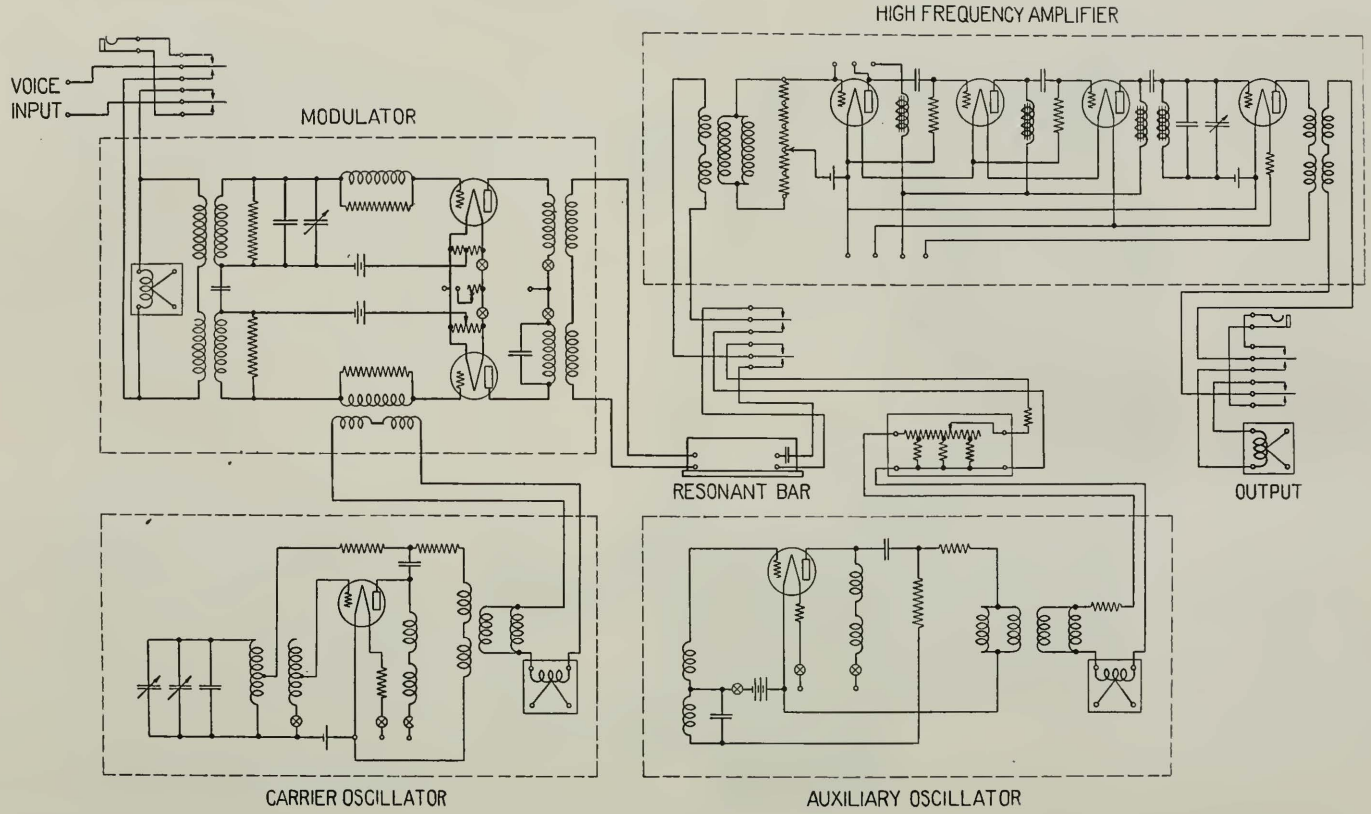


Fig. 6—Schematic diagram of analyzer circuit

A number of other uses of the analyzer in acoustic measurements might be cited; for instance, the output of a transmitter at one particular frequency may be measured in the presence of room noises or other disturbing sounds. In general its high selectivity, convenience of operation and portability make the analyzer an extremely valuable instrument in a wide variety of acoustic and electrical measurements.

Appendix

Inasmuch as the incorrect operation of the modulator could give false indications of frequency components in the wave to be analyzed, it may be of interest to take up some of the more important aspects of the theory of the vacuum tube plate current modulator as applied to this analyzer.[3] In general, the output of a vacuum tube is of the form

$$I_1 = A_0 + A_1E + A_2E^2 + A_3E^3 + A_4E^4 + \text{etc.}, \tag{1}$$

where E is the voltage applied between the cathode and grid of the tube and the coefficients A_0, A_1, A_2, A_3, etc., depend upon the average potential of the grid, the constants of the tube itself and the total impedance of the output circuit to the various frequency components appearing in the output. If we apply two voltages of the form $P\cos(pt-\theta)$ and $Q\cos(qt-\phi)$ simultaneously to the input of a vacuum tube, the above general expression will have the form

$$\begin{aligned} I_1 = A_0 &+ A_1[P\cos(pt-\theta) + Q\cos(qt-\phi)] \\ &+ A_2[P\cos(pt-\theta) + Q\cos(qt-\phi)]^2 \\ &+ A_3[P\cos(pt-\theta) + Q\cos(qt-\phi)]^3 \\ &+ A_4[P\cos(pt-\theta) + Q\cos(qt-\phi)]^4, \text{ etc.} \end{aligned} \tag{2}$$

For simplicity let $a = P\cos(pt-\theta)$ and $b = Q\cos(qt-\phi)$ and then expanding algebraically we obtain

$$\begin{aligned} I_1 = A_0 + A_1(a+b) + A_2(a^2+2ab+b^2) + A_3(a^3+3a^2b+3ab^2+b^3) \\ + A_4(a^4+4a^3b+6a^2b^2+4ab^3+b^4) + \text{etc.} \end{aligned} \tag{3}$$

Now substituting $P\cos(pt-\theta)$ and $Q\cos(qt-\phi)$ for a and b respectively and simplifying trigonometrically, the frequencies appearing in the output of the tube are shown in Table 1.

If we employ a balanced modulator of the type shown in Fig. 7, certain frequency components shown in the table are eliminated. In this modulator the voice input transformer is so connected as to

[3] For a more complete discussion of the vacuum tube modulator see *Proceedings I. R. E.*, April 1919; "A Theoretical Study of the Three Element Vacuum Tube," by J. R. Carson.

TABLE 1

TERMS IN THE FIRST FOUR ORDERS OF MODULATION

$P \cos (pt - \theta)$ AND $Q \cos (qt - \phi)$

General Expression for Current Output

$I_1 = A_0 + A_1E + A_2E^2 + A_3E^3 + \text{etc.}$

Frequency $\omega = 2\pi f$	Coefficients			
	A_1	A_2	A_3	A_4
0		$1/2\ P^2$, $1/2\ Q^2$		$3/8\ P^4$, $3/2\ P^2Q^2$, $3/8\ Q^4$
$(pt - \theta)$	P		$3/4\ P^3$, $3/2\ PQ^2$	
$(Qt - \phi)$	Q		$3/4\ Q^3$, $3/2\ P^2Q$	
$(2pt - 2\theta)$		$1/2\ P^2$		$1/2\ P^4$, $3/2\ P^2Q^2$
$(2qt - 2\phi)$		$1/2\ Q^2$		$1/2\ Q^4$, $3/2\ P^2Q^2$
$(pt - \theta) \pm (qt - \phi)$		PQ		$3/2\ P^3Q$, $3/2\ PQ^3$
$(3pt - 3\theta)$			$1/4\ P^3$	
$(3qt - 3\phi)$			$1/4\ Q^3$	
$(2pt - 2\theta) \pm (qt - \phi)$			$3/4\ P^2Q$	
$(pt - \theta) \pm (2qt - 2\phi)$			$3/4\ PQ^2$	
$(4pt - 4\theta)$				$1/8\ P^4$
$(4qt - 4\phi)$				$1/8\ Q^4$
$(3pt - 3\theta) \pm (qt - \phi)$				$1/2\ P^3Q$
$(2pt - 2\theta) \pm (2qt - 2\phi)$				$3/4\ P^2Q^2$
$(pt - \theta) \pm (3qt - 3\phi)$				$1/2\ PQ^3$

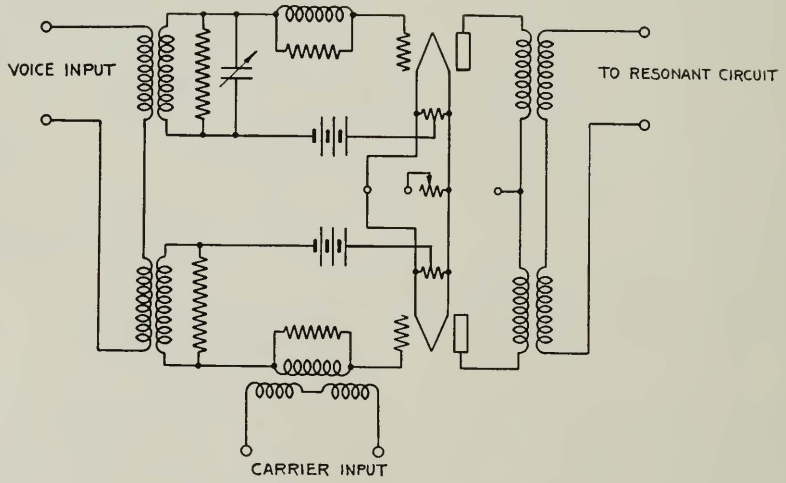


Fig. 7—Schematic diagram of modulator

impress instantaneous voltages of opposite signs on the two tubes, while the carrier windings impress instantaneous voltages of the same sign on the tubes. The output transformer is connected so that the algebraic difference of the two plate currents appears in the output

of the modulator. Equation (3) shows the output of one of the tubes, but the output for the other tube in which the voice input is reversed will be

$$I_2 = A_0 + A_1(a-b) + A_2(a^2 - 2ab + b^2) + A_3(a^3 - 3a^2b + 3ab^2 - b^3) + A_4(a^4 - 4a^3b + 6a^2b^2 - 4ab^3 + b^4) + \text{ etc.} \quad (4)$$

If the circuits of the two tubes are exactly balanced, then since the output of the modulator is so connected that only the algebraic difference of the two plate currents will appear, all of the frequencies arising from the terms of like sign in equations (3) and (4) will be suppressed or, subtracting equation (4) from equation (3), the resultant current will be

$$I_1 - I_2 = +2A_1b + A_2(+4ab) + A_3(+6a^2b + 2b^3) + A_4(+8a^3b + 8ab^3). \quad (5)$$

The frequency components arising from the various terms of equation (5) are shown in Table 2.

TABLE 2

FREQUENCY COMPONENTS ARISING FROM ALGEBRAIC DIFFERENCE OF $I_1 - I_2$

$2\ A_1Q \cos (qt - \phi)$
$2\ A_2PQ \cos [(pt - \theta) \pm (qt - \phi)]$
$3\ A_3P^2Q \cos (qt - \phi)$
$3/2\ A_3P^2Q \cos [(2pt - 2\theta) \pm (qt - \phi)]$
$1/2\ A_3Q^3 \cos [3qt - 3\phi]$
$3/2\ A_3Q^3 \cos (qt - \phi)$
$A_4P^3Q \cos [(3pt - 3\theta) \pm (qt - \phi)]$
$3\ A_4P^3Q \cos [(pt - \theta) \pm (qt - \phi)]$
$3\ A_4PQ^3 \cos [(pt - \theta) \pm (qt - \phi)]$
$A_4PQ^3 \cos [(pt - \theta) \pm (3qt - 3\phi)]$

In the analyzer the only useful order of modulation is the second, that is, the frequencies arising from the A_2 term of equation (1). The component of interest here is the term $2A_2PQ \cos [(pt - \theta) - (qt - \phi)]$. As can be seen, this frequency component is proportional to both P and Q and for a given value of P is proportional to Q. In other words the input into the tuned circuit is a linear function of the magnitude of the particular frequency component under consideration in the wave to be analyzed. As will be seen from Table 2, there are a number of other frequency components due to the various orders of modulation in the modulator output. A little consideration, however, will show that a number of these components cannot appear in the output of a resonant circuit tuned to approximately 11,000 c.p.s. where the upper frequency limit of the voice input is 5,000 c.p.s. and the range of the carrier oscillator is limited from the resonant frequency of the tuned circuit to 5,000 c.p.s. above. The only components of Table 2 which can be passed by the tuned circuit are:

$$2A_2PQ \cos [(pt - \theta) - (qt - \varphi)],$$
$$3A_4P^3Q \cos [(pt - \theta) - (qt - \varphi)],$$
$$3A_4PQ^3 \cos [(pt - \theta) - (qt - \varphi)],$$
$$1/2\ A_3Q^3 \cos (3qt - 3\varphi),$$
$$A_4PQ^3 \cos [(pt - \theta) - (3qt - 3\varphi)].$$

In addition to the desired term in the second order modulation of frequency $(pt - qt)/2\pi$ there are two other terms of the same frequency in the fourth order modulation. These have respectively the coefficients $3A_4P^3Q$ and $3A_4PQ^3$. With a given value of carrier input P, the first of these is proportional to Q and it will add to the second order term but will cause no serious trouble. The second term, however, is proportional to Q^3 and, therefore, would cause the input to the tuned circuit to depart from the desired linear relationship with respect to Q. However, the ratio of the cofficient of this fourth order term to that of the second order term is $3A_4Q^2/2A_2$ which is proportional to Q^2 and, therefore, the effect of the fourth order term will fall off rapidly as Q is reduced. In the third order modulation, there is a component of frequency $3qt/2\pi$ which would be passed by the tuned circuit if there were a component of 1/3 the resonant frequency of this circuit in the wave to be analyzed. Considering the extreme sharpness of the tuned circuit, it is rather improbable that such a condition will occur. Moreover, this component is proportional to Q^3 and, therefore, will fall off rapidly as Q is reduced. In the fourth order modulation, there is also a component $A_4PQ^3 \cos [(pt-\theta) - (3qt-3\phi)]$ of frequency $(pt-3qt)/2\pi$. This component would indicate a third harmonic in the unknown wave although it actually contained no other frequency than that of $qt/2\pi$. However, the ratio of its coefficient to the desired second order term is $A_4Q^2/3A_2$ and, therefore, the false indications of a third harmonic can be reduced to any desired extent by reducing Q.

It is evident, therefore, that in an analyzer of this type it is desirable to keep the magnitude of the input of the unknown wave as low as is consistent with obtaining satisfactory meter readings. In the two analyzers now in operation, false indications by the introduction of extraneous frequency components due to the third, fourth and higher orders of modulation are negligibly small. Measurements on an essentially pure frequency within the frequency limits of the apparatus show harmonics less than 0.1 per cent of the fundamental. With an indicated harmonic of not more than this magnitude, it is difficult to tell whether such a harmonic is actually present in the wave or a false indication. At any rate, the harmonic is small enough as to be of no significance. It is, of course, difficult to build the modulator so that it

will be exactly balanced over the frequency range covered by the carrier oscillator and, therefore, frequency components such as $(pt-2qt)/2\pi$ appearing in the third order modulation as shown in Table 1 are not totally eliminated. However, the effect of unbalance is of no serious consequence in the practical operation of the analyzer. There is, of course, a possibility of false indications due to higher orders of modulation than the fourth, but the coefficients A_3, A_4, etc., are usually small in comparison with A_2 and in general become successively smaller. Moreover, it will be evident that these false indications may be reduced to any desired extent by reducing the magnitude of the unknown wave.

Analyzer for Complex Electric Waves

By A. G. LANDEEN

IN problems concerned with the electrical transmission of intelligence it is necessary to have means for studying complex electric waves. In certain steady state conditions these complex waves become periodic, and, although not sinusoidal as a whole, may be resolved into a number of sinusoidal components. It is particularly important to be able to measure these components individually.

In studies on systems employing carrier currents which may be transmitted over wire lines it is often necessary to measure a signal wave component which may lie anywhere in the frequency range between 100 and 100,000 cycles per second. The most important range at the present time is, however, below 40,000 cycles per second. In addition to covering a wide range of frequencies these components may also vary considerably in amplitude, both as to absolute value and as to value relative to other components in the signal wave.

For several years there has been in use in the Bell Telephone Laboratories special apparatus by means of which a single component of a complex periodic current wave may be selected from the remaining components and its amplitude determined. The sensitivity and selectivity of this apparatus are such that components of small amplitude may be accurately measured even in the presence of other components of several hundred times the amplitude and differing but little in frequency. With the latest improved form it is now possible to measure current components having amplitudes as low as 10^{-7} amperes with a possible error of 10 per cent. For such minute currents this is within the error which might be introduced by the external apparatus such as attenuators and thermocouples together with their calibration charts.

Though the apparatus was primarily designed for use in current wave analysis work, it may also be readily adapted to voltage analysis. Suitably calibrated, it can be used also as a frequency meter of extremely high precision.

Introduction

The method of analysis here described had its origin in a circuit built by J. W. Horton in 1917. This had a resistance coupled tuned circuit responsive to the component desired. Following the tuned circuit two stages of amplification were used to magnify the selected current. This current was then passed on to a third unit where it was rectified and measured by a D.C. meter. It was evaluated directly by

noting the meter deflection and referring to calibrations of the analyzer which had been made with known input currents.

This elementary form of measuring circuit was developed during the World War for the analysis of the sound waves encountered in listening devices used for the detection and location of submarines and torpedoes. It covered the range of audible frequencies and had sufficient sensitivity for its original purpose.

It will be remembered that the first commercial application of multiplex transmission by means of carrier currents came almost simultaneously with the Armistice. The continued study of carrier systems found a useful tool in the current analyzer but placed considerably more rigorous requirements on its performance. These were met by the addition of a second tuned circuit and amplifier system, working from the output of the first, thus giving far greater selectivity than is obtainable in a single circuit. The presence of the multi-stage amplifier between the selective circuits facilitates tuning by avoiding interactions between the circuits. A second modification was the use of a substitution method for evaluating the amplitude of the selected components, as with the considerable increase in the ranges of amplitude and frequency covered, the calibration method for measuring the current became impracticable. To evaluate the current, the output from a sine wave oscillator, which was tuned to the same frequency as the component being measured, was substituted at the input to the analyzer and the amplitude adjusted until it gave the same meter deflection as the unknown component. Since the current from the oscillator is of the same frequency and amplitude as the original component, we can determine the magnitude of the latter by measuring the oscillator output. A convenient means for doing this is to interpose between the oscillator and the analyzer a variable attenuator. It is then possible to fix the oscillator output current at some convenient value, such as 1.0 milliampere or 10 milliamperes and to adjust the input to the analyzer by means of the attenuator. The current can then be read directly from the attenuation tables, it being only necessary to know the location of the decimal point.

The development of the analyzer in this form was carried to the limit of its practicability by F. Mohr. With an analyzer containing three units it was possible to carry through an extensive study of the modulation introduced into the Key West-Havana cable due to the non-linearity of the characteristics of the iron used for loading.

The Heterodyne Method

With the advance in carrier communication, greater refinement in measurement became necessary, calling for still higher selectivity in the analyzer. The best means for accomplishing this appeared to be to heterodyne the wave under investigation in such a manner as to move it to a lower position on the frequency scale. Then with a fixed tuned circuit which would pass only the low frequency current corresponding to the desired component, much greater selectivity might be obtained because of the relatively greater spacing.

To heterodyne the desired component there is required a separate oscillator and a modulator in which the current to be measured and the separately generated current are combined to produce a current of lower frequency. This in effect translates the current under investigation from a high frequency to one of much lower frequency; retaining, however, the relative amplitudes of the components. Since the amount of this translation is determined by the frequency of the local oscillator, a particular component can always be given a certain predetermined value by adjusting the oscillator. This permits the use of a fixed tuned circuit which is highly selective to the difference frequency in the modulator output. By choosing a low value for this frequency it is possible to make the percentage difference between this and interfering frequencies much larger than between the corresponding high frequencies. In the present analyzer 800 cycles per second has been chosen as a suitable value. If, then, the current to be measured had a frequency of 20,000 cycles per second, the local oscillator would be set at 20,800 cycles. It could of course also be set at 19,200 cycles if desired and produce the same difference frequency. If there were also present another current of say 20,500 cycles, the interval in the original wave would only be 2.5 per cent; after heterodyning, however, it would appear as a 300 cycle current, if heterodyned by the 20,800 cycle current. The interval thus becomes nearly 40 per cent of the frequency for which the tuned circuit is adjusted. If these currents are heterodyned directly in a simple modulator, there is also the possibility of modulation between components in the original complex wave. This would result, in the case chosen, in a current having a frequency of 500 cycles, but the percentage difference between the 800 and either the 500 or the 300 cycle currents is many times greater than that between the original high frequency currents so that the fixed tuned circuit would have a very high discrimination to the interfering current.

In some cases the intermodulation between components of the original wave may coincide with the component being measured, or may

interfere with the measurement in some other manner. This difficulty is minimized in two ways: first, the amplitude of undesired components is reduced relative to the component under observation, by using selectivity of the type previously described, before impressing the wave on the modulator; second, the modulator is made of a balanced type which permits efficient heterodyne action between the selected component and the heterodyning current, but which gives very little intermodulation between such components as remain after the initial selection. The partial elimination of undesired components before modulation is of further advantage in preventing unnecessary loading of the modulator, which might tend to give it different efficiencies with the complex current and with the sine wave calibrating current.

Description of Apparatus

A schematic diagram showing the several functional elements is given in Fig. 1. The units have been arranged in the order in which

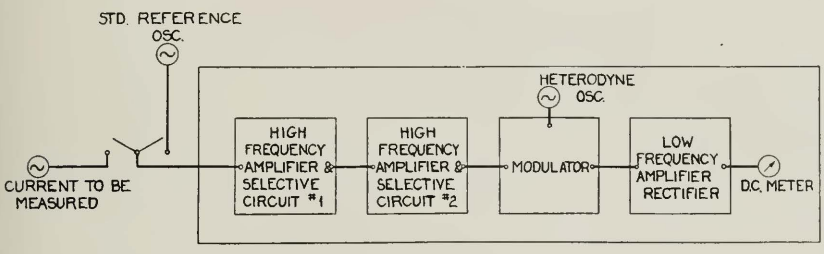


Fig. 1—Schematic diagram of heterodyne current analyzer

the measuring current would proceed, which is also the order of assembly of the completed instrument. In the following, a more detailed description will be given of the individual units.

High Frequency Amplifiers. The circuit for the first unit is shown in Fig. 2 and for the second unit in Fig. 3. These circuits differ mainly in the output terminations of the second tube. The first high frequency unit consists of a simple series tuned circuit together with two amplifier tubes. In the tuned circuit is also included the coupling resistance, R, for controlling the input to the amplifier. The sharpness of resonance will therefore depend to some extent upon the value of this resistance, but under most operating conditions it is relatively small in comparison to the total effective resistance which includes that of one tuning coil and the condensers.

In both of these circuits the A.C. input voltage to the first tubes is obtained from the drop across the condensers. If it is desired to dis-

criminate against a component of higher frequency than the one being measured, it is advantageous to use the voltage across the condenser, whereas if the interfering component is of lower frequency, the voltage across the inductance should be used. It is therefore desirable to make

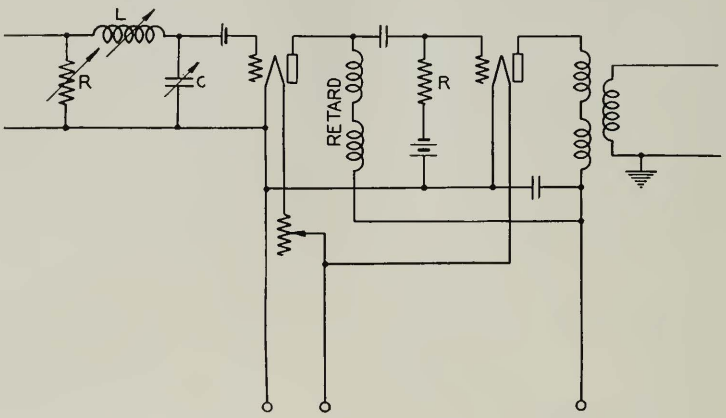


Fig. 2—High frequency amplifier unit No. 1 of heterodyne current analyzer

provision by means of a switching arrangement whereby the voltage may be applied from either the inductance or the capacity depending upon the discrimination desired. With the coils and condensers used the voltage across the condenser at resonance is over a hundred times that across the coupling resistance.

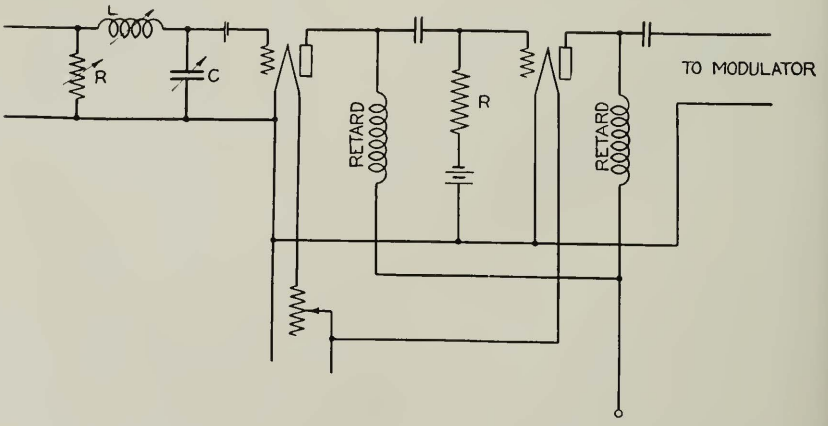


Fig. 3—High frequency amplifier unit No. 2 of heterodyne current analyzer

The first tube of each unit serves as a voltage amplifier. This works into a high resistance between the grid and filament of the power tube.

Because of the wide range of frequency over which the amplifiers will be operated, resistance, rather than transformer coupling, was chosen as the most reliable form of interstage connection. Due to the large step-down in impedance necessary between the first amplifier and the coupling resistance of the second selective circuit, a transformer having

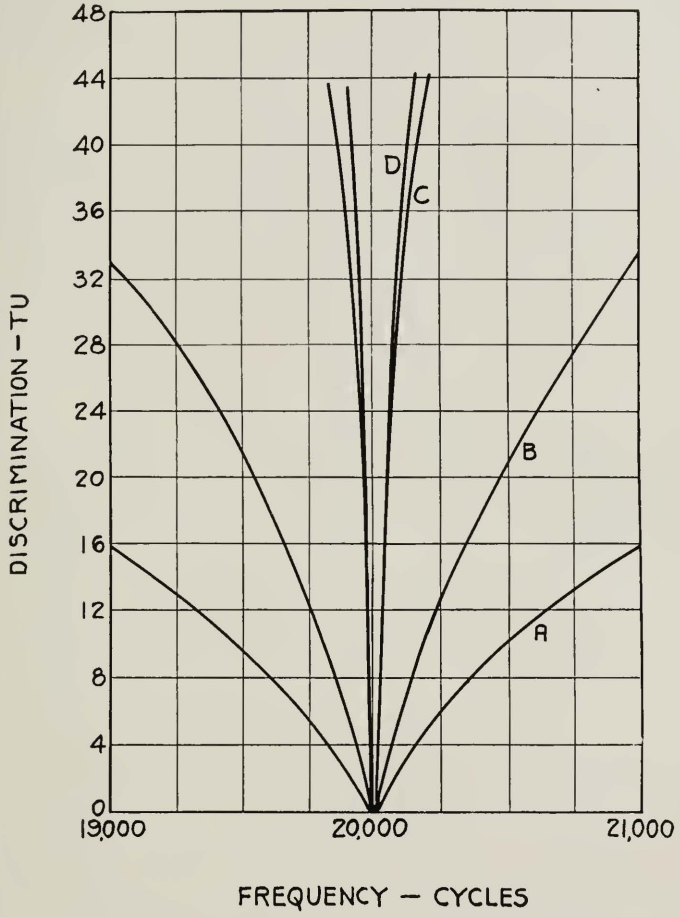


Fig. 4—Discrimination curves of heterodyne current analyzer

a high step-down ratio was used. The output of the second amplifier also works into a transformer through a fixed resistance and potentiometer as shown in Fig. 5. The purpose of the fixed resistance is to maintain a more uniform output impedance as the potentiometer is varied. Though the two amplifiers are almost identical in their circuits, they differ in the plate voltage. The first section is operated at

240 volts since its tubes are subjected to the heaviest input, being preceded by only one selective circuit which can but partially eliminate large interfering currents. This seems a most unusual arrangement until it is remembered that, although the amplitude of a particular part of the current may be increased, the total load on the first stage may well be greater than that on the final stage. This is the reverse of the situation in cascade amplification where the first tubes handle only a small current and the succeeding ones a proportionately larger current.

The discrimination of one of the tuned circuits when a coupling resistance of 1 ohm is used is given by curve *A* of Fig. 4. Curve *B* shows the effect of adding the second tuned circuit. If regenerative amplification were employed, both the selectivity and amplification could of course be greatly increased, but this has not been used because of the necessity for high stability and measurement precision.

Modulator. As previously mentioned, the second amplifier unit works into a modulator, the circuit of which is shown in Fig. 5. This is

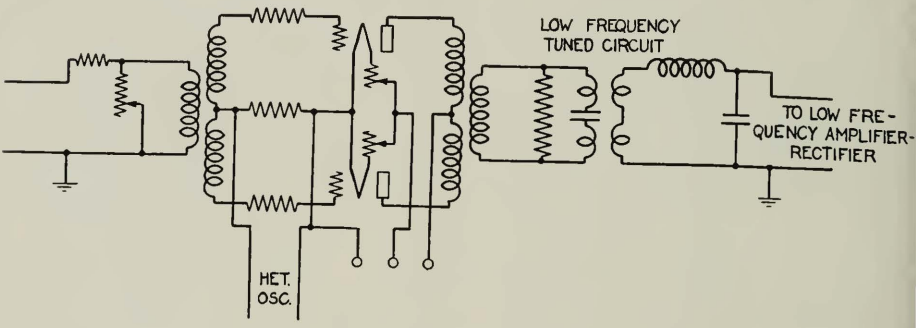


Fig. 5—Modulator for heterodyne current analyzer

of a two tube balanced type in which modulation, or frequency transformation, takes place in the grid circuit. The heterodyning frequency is applied in the common input lead across a suitable resistance. The input from the amplifier is applied through a transformer across the grids of the two tubes in series with a high resistance in each side. No biasing potential is applied on the grids. A modulator operated in this manner has the property of giving a modulation output proportional to the smaller of the two input currents and independent of the larger. The amplitude of this output may, therefore, be determined entirely by the amplitude of the component being measured. Another desirable characteristic of this type of modulator is that its efficiency is not affected by interference, hence it will show a fixed relation between

the low frequency output and a given input component regardless of the presence of other interfering currents in the input side. Through the use of a balanced circuit intermodulation is reduced considerably below the limit possible with a single tube modulator.

The output of the modulator is connected directly to a double tuned circuit which selects the low frequency modulation product corresponding to the component of the complex wave being examined. The frequency to which this circuit is adjusted is, as already mentioned, 800 cycles.

Low Frequency Amplifier-Rectifier. The 800 cycle output from the modulator is generally too small to measure on a meter of the usual type without first being amplified. For this reason a low frequency amplifier has been added and the output of this rectified so that all measurements could be made on a sensitive D.C. meter, having a full scale deflection of 1 or 2 milliamperes.

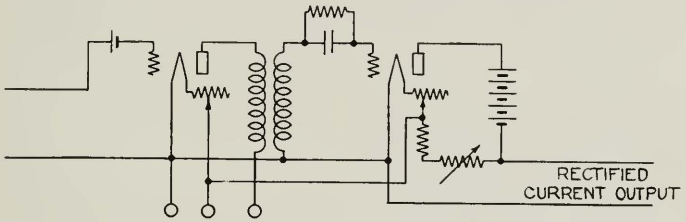


Fig. 6—Low frequency amplifier-rectifier for heterodyne current analyzer

The combined low frequency amplifier-rectifier circuit is shown in Fig. 6. A step-up transformer is used between the amplifier tube and the rectifier in order to increase the amplification so that, for a given A.C. output, a smaller input to the first amplifier might be used. Since this circuit is always to be operated at the one frequency (800 cycles), its overall frequency range characteristic is not of particular interest and its performance was studied only at the one frequency. The rectifier is of the grid leak and condenser type. Its performance differs from the usual type of rectifier since it is operated over that portion of its characteristic which gives a linear relation between input voltage and direct current output. By suppressing the space current corresponding to zero input the accuracy with which data can be taken is greatly increased. As shown by the circuit arrangement in the output side, part of the "A" battery current is used to oppose the space current through the meter. This permits using a meter of high sensitivity, having a full scale deflection of one or two milliamperes, on which 1/100th part of a milliampere can easily be read.

Heterodyne Oscillator. The major requirement to consider in the design of the heterodyne oscillator was that of frequency stability. As the only function of the oscillator was to furnish a current for heterodyning the one being measured the output requirements were moderate, 10 milliamperes into 600 ohms being ample, but it was important that the frequency remain constant during a series of measurements as even slight variations, of a fraction of a per cent, would change the attenuation of the 800 cycle tuned circuit to the sideband current. Stability in this case depended mainly upon the "A" and "B" battery voltages, since the output load consisted of a pure resistance and there was no reaction back to the oscillator due to a variable output impedance.

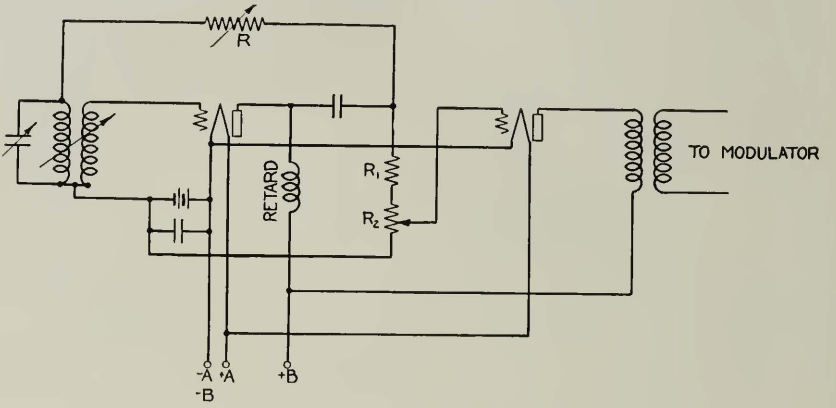


Fig. 7—Oscillator for heterodyne current analyzer

The oscillator circuit is shown in Fig. 7. This shows two tubes, one as an oscillator and one as an amplifier. The coupling consists of a 20,000-ohm resistance used as a potentiometer, which is placed in series with a 100,000-ohm resistance and the two used as the oscillator load. This makes the coupling impedance only one sixth of the total oscillator output impedance and therefore reduces the effect which the amplifier tube might have on the frequency. The change in frequency due to the "A" and "B" voltage can also be controlled by inserting a high resistance in the feed-back path between the plate and oscillation circuits. This should be several times that of the tube impedance so that any change in the latter would then be a proportionately smaller part of the total impedance and hence have a less effect upon the frequency.

The selection of tuning coils for various frequency ranges is made by keys which at the same time select the proper feed-back resistance. Only three coils are used to cover the frequency range between 3,000

and 50,000 cycles. The output of the amplifier tube works into the resistance in the mid-branch of the modulator input circuit.

Mechanical Construction

Figs. 8 and 9 show the front view of the complete analyzer as in present use and also the interior view of an individual unit (the heterodyne oscillator), to indicate the method of construction.

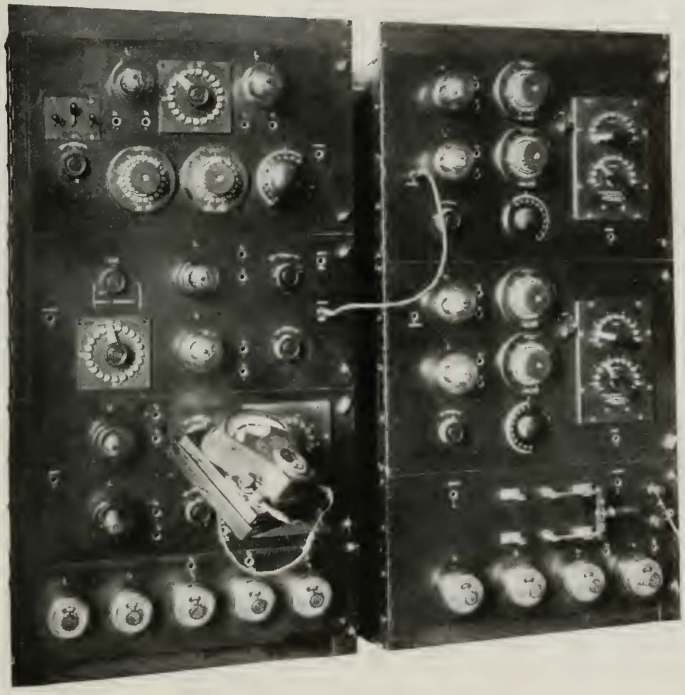


Fig. 8

Each section is completely enclosed in metal, by having shielded cases in the rear and heavy metal panels in front. Perhaps the feature of most interest in the mechanical construction is that of the hinged front panels. These were made of 1/4 in. aluminum to insure ample strength and rigidity with a minimum of strain on the supporting hinge. Each panel is provided with two thumb screws on the edge opposite the hinge so that the units may be readily inspected. This feature of accessibility is particularly desirable in making periodic inspections and in renewing the "C" batteries, which have been mounted on the panels. The heavy material such as tuning coils has been mounted inside the case. The flexible leads, as shown in the lower corner of the

opened unit, connect the panel apparatus with the "A" and "B" batteries, and with the modulator. All of the panels are provided with a metal strip, such as shown along the top and bottom edge, which fits into a groove in the case and thereby provides better shielding between

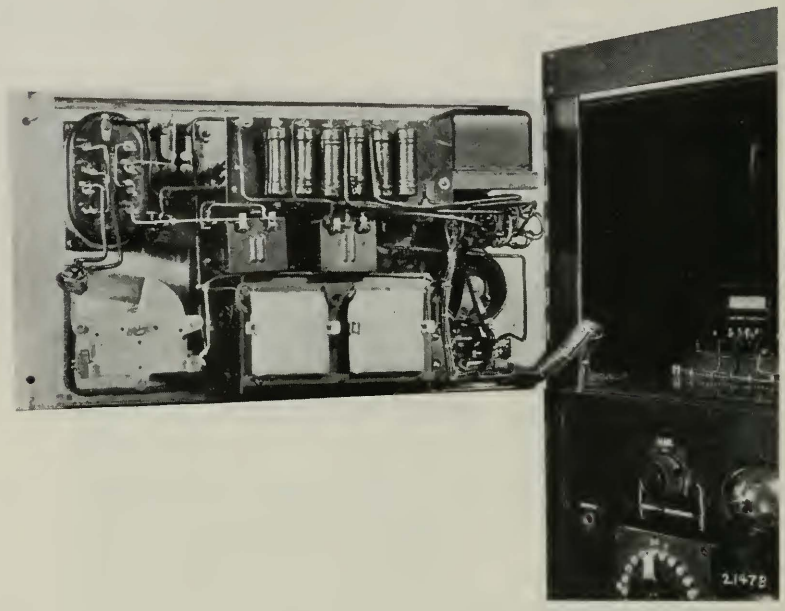


Fig. 9

each of the units. Since each unit of the analyzer is provided with a separate shielding case, two sheet iron walls are interposed between any pair, forming a double shielding to electric fields occurring within the analyzer. The complete unit of two sections is very economical in space as the overall height, width and depth of each section are only 36, 19 and 12 1/2 inches respectively.

Characteristics of Complete Current Analyzer

The successful operation of the heterodyne current analyzer depends, of course, upon knowing its limitations and its reliability. Under limitations may be grouped sensitivity, selectivity and modulation in the analyzer itself; and under reliability, the limits within which readings can be repeated.

Sensitivity. The sensitivity depends upon the coupling resistance used in the two amplifier units, and increases with added resistance so long as this is only a few ohms and small as compared to the total

effective resistance in the selective circuits. When these are made 1 ohm each, it requires only 1 microampere to give one milliampere of rectified current. With 10 ohms coupling resistance only 10^{-8} amperes input would be required. This, however, is a larger coupling than it is desired to use, since the analyzer becomes too sensitive and susceptible to mechanical vibrations as well as to electrical interference from outside sources.

One desirable feature is that the sensitivity characteristic of the complete current analyzer is a straight line so that doubling the input will give twice the deflection in the meter reading the rectified current. This is an advantage since, if the deflections and input amplitudes do not change by the same ratio, the presence of interfering currents is indicated.

Selectivity. The selectivity of the current analyzer depends upon the time constants of both the high and low frequency tuned circuits, and to some extent upon the coupling resistances, but the latter are usually relatively small and do not have an appreciable effect. The discrimination obtained by the use of the heterodyne method and fixed low frequency selective circuit is shown by curve *C* of Fig. 4. This curve may be compared with curve *A*, which shows what can be done with high grade elements in a single tuned circuit. The discrimination of the complete analyzer, including the initial stages and the heterodyne stage, is given by curve *D*. Tests with two frequencies show that if one is 250 times as large as the other the smaller may be measured without appreciable error if the difference in frequency is not less than 1 per cent; if the ratio of amplitudes is 1,000 to 1, the frequency difference need not be less than 2 per cent.

Modulation. As to modulation in the current analyzer there are two sources which contribute, the vacuum tubes and the tuning coils and transformers. Of these the tubes are the most troublesome, since they furnish both even and odd order modulation products, whereas the coils contribute only to the odd orders. Of these the third is generally the only one that is of any interest since higher odd orders are too small to produce any interference. Modulation need be considered only when measurements are made of small components in the presence of very large ones, as it is under these circumstances that conditions for modulation are the most favorable. This condition requires high sensitivity which is obtained by increasing the coupling resistance, and large resistance means greater interference voltage on the first amplifier and also less selectivity in the tuned circuits. The result is an increased load on all sections of the current analyzer, which causes modulation and may produce an error in the readings. It is,

therefore, important to know the limitations which this imposes upon it as a measuring device. Then with this information data can be taken within known limits of accuracy. Measurements have been

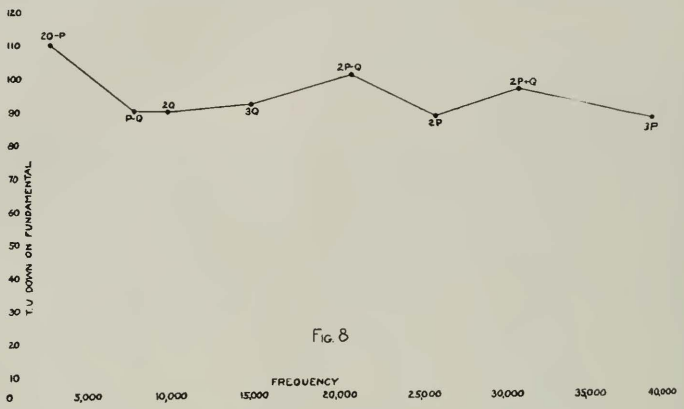


Fig. 8

Fig. 10—Second and third order modulation in heterodyne current analyzer together with 5,000 and 13,000 cycle band-pass filters with one or two input frequencies

made on the combined modulation occurring in the analyzer and also in the filters which were necessarily associated with the measurements. These were made with one input current and also with two input currents of different frequencies, which were applied simultaneously to the analyzer. One or two filters were therefore necessary to suppress all components except the fundamental currents desired, but any small amount of modulation which would occur in the filters would add to that produced in the analyzer so that the results shown by the curve in Fig. 10 represent the total modulation in both the filters and the analyzer. The modulation amplitude is expressed in terms of transmission units with respect to the current into the analyzer. This curve is quite irregular and depends somewhat upon the frequency. It is the lowest at 39,000 cycles where the modulation current is shown to be 86 *TU*[1] down or 0.00005 as large as the current into the analyzer. Measurements can, therefore, be made at this frequency of the modulation occurring in any device when its amplitude is not less than 66 *TU* below the amplitude of the fundamental, with a possible maximum error of 10 per cent. At other frequencies this amplitude may be less, as for instance at 21,000 cycles, measurements may be made up to 80

[1] For a discussion of this method of expressing current ratios see "The Transmission Unit and Telephone Reference Systems" by W. H. Martin, *Journal A. I. E. E.*, June 1924, Vol. XLIII, No. 6; also *Bell System Technical Journal*, July 1924, Vol. III, No. 3; also "The Transmission Unit," R. V. L. Hartley, *Electrical Communication*, July 1924, Vol. III, No. 1.

TU without exceeding the same percentage of error, or up to 60 *TU* with 1 per cent error due to undesirable modulation in the current analyzer.

Reliability. Use of the heterodyne current analyzer over a period of two years has proven it to be one of the most reliable means for making measurements. With proper maintenance, which consists only in maintaining constant "A" and "B" battery voltages and grid voltage, and with proper precautions as to shielding and balance, readings can be taken with a precision of 2 per cent.

Vacuum Tube Curves Obtained with Heterodyne Current Analyzer. A number of curves have been added to illustrate the application of the current analyzer, though of course these represent only a small part of the field of usefulness for which it is adapted.

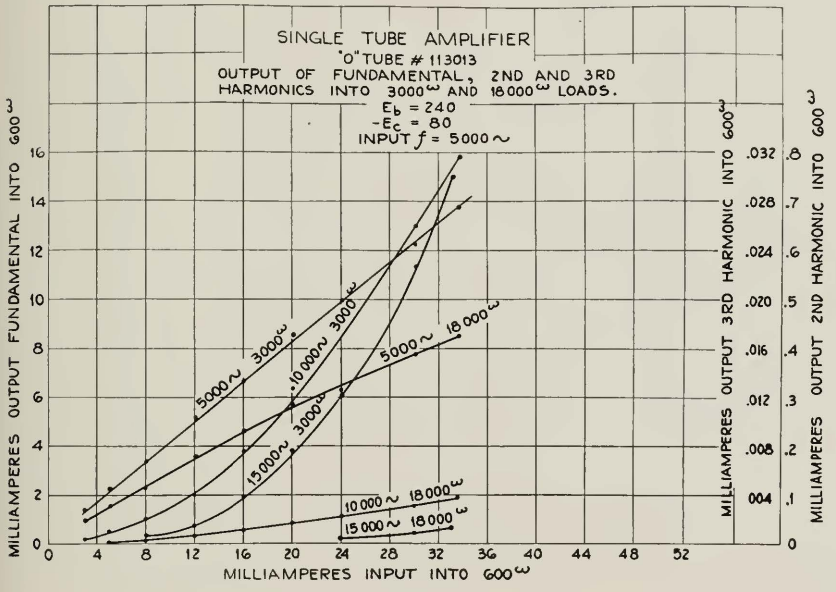


Fig. 11

The first set of curves, shown in Fig. 11, were taken on an "O" tube (104–D) to show how the fundamental current and the second and third harmonics produced in the tube changed with increase in the input amplitude of a single frequency. Two sets of curves are shown which were taken for two values of load impedance, one being equivalent to the normal tube impedance and the other being six times as large. The output currents have all been computed to show the equivalent output into 600 ohms which is a common reference standard of impedance used in telephone work.

The second set of curves, shown in Fig. 12, were taken with an "L" tube (101–D) but with two input frequencies of $Q = 5{,}000$ and $P = 13{,}000$ cycles applied simultaneously. The measurements in

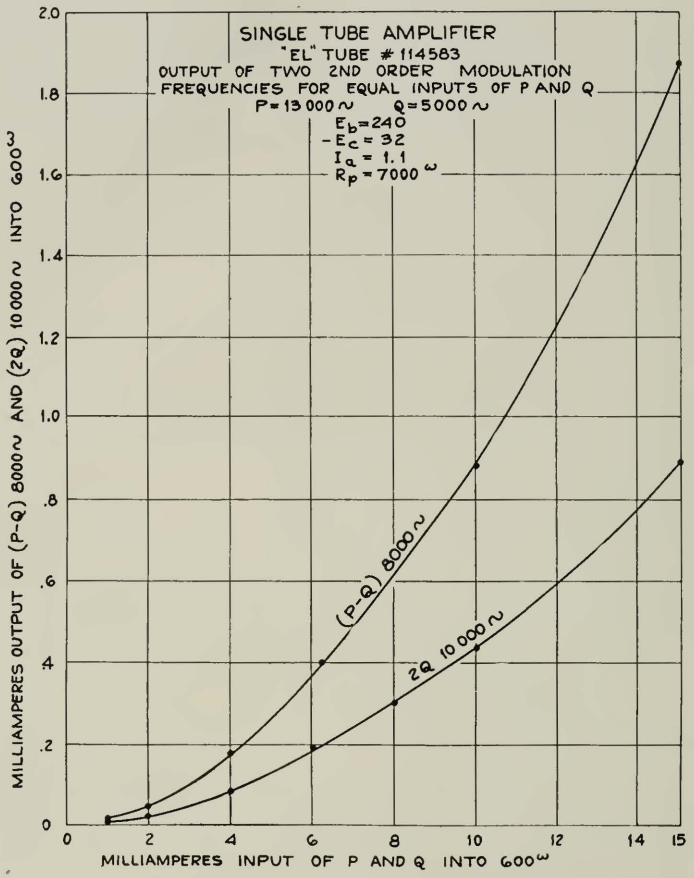


Fig. 12

this case were made on the modulation product consisting of the difference frequency $(P - Q) = 8{,}000$ cycles and of the second harmonic of Q, $2Q = 10{,}000$ cycles. Such curves furnish a quick and accurate means of studying the performance of the vacuum tube and also afford a convenient check on mathematical computations which, in this particular case, with a two frequency input, have indicated that the difference frequency amplitude should be twice that of the second harmonic.

In Fig. 13 are shown the third harmonic output of an "O" tube with a single input frequency of constant amplitude and with a variable load impedance. The harmonic current shown by this curve could

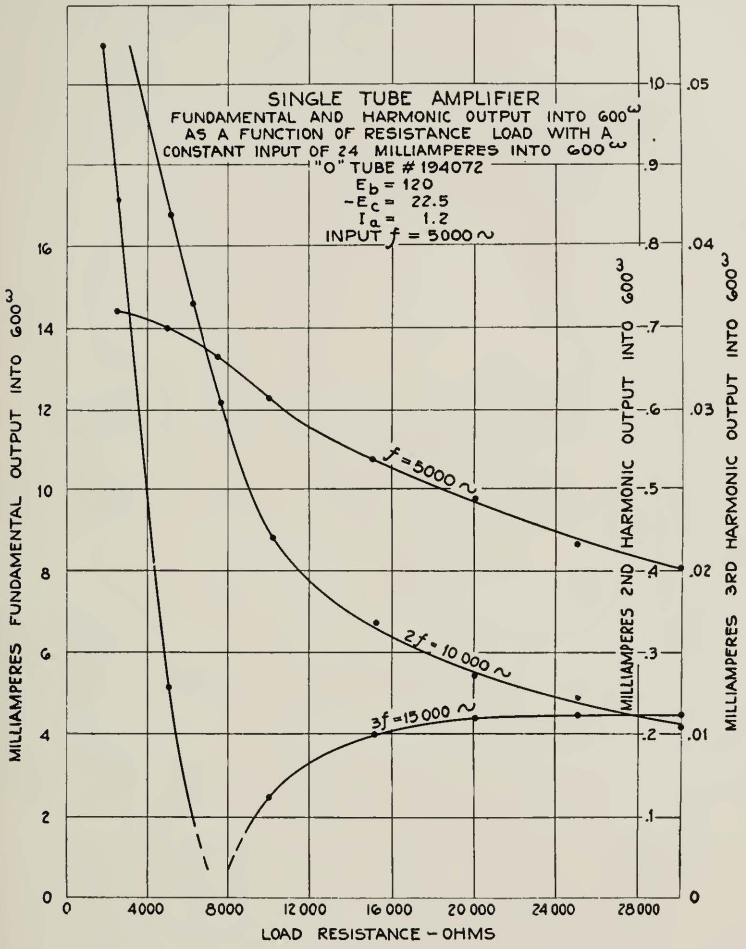


Fig. 13

not easily be predicted from computation due to the many factors involved, and can be measured only by an instrument of high sensitivity that can faithfully follow the varying amplitudes of a minute current.

Voltage Analyzer. In addition to its use as a current measuring device the current analyzer may be used, as previously mentioned, for measuring the voltage in circuits where the power dissipation is very

small, or where the voltage cannot be detected except by several stages of amplification such as are obtained in the current analyzer. To adapt it for this purpose it is only necessary to precede it by a simple circuit such as shown in Fig. 14. This consists of a single

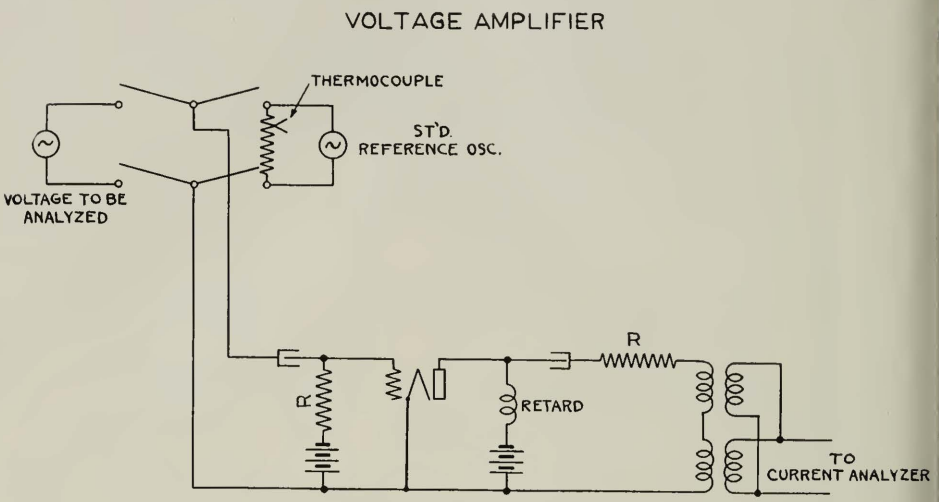


Fig. 14—Voltage amplifier

vacuum tube having a large resistance across the grid and filament. This resistance should be greater than the impedance across which the voltage is to be measured. The output side works into a step-down transformer through a resistance of several times the output impedance of the tube. This tends to straighten out the characteristic and to lower the tube modulation level. The object of the step-down transformer is of course to secure greater efficiency in working into the low input impedance of the heterodyne current analyzer.

The measuring procedure would be to apply the voltage to be evaluated across the high impedance input and adjust the analyzer in the usual manner; then substitute the output from the standard oscillator of the same frequency across the amplifier and adjust the amplitude to give the same meter deflection. In order to determine the voltage applied, the oscillator may be connected across a known resistance in parallel with the input, and the current into this resistance measured. The IR drop will then be a measure of the voltage applied.

The range of voltage which can be measured of course depends upon the biasing potential on the amplifier grid as it is not desirable that grid current flow through the high resistance and increase the tube modula-

tion. The most frequent use of the amplifier in the laboratory has been in measuring voltages around 10^{-2} volts but it can also be equally well used to measure much smaller values, of the order of 10^{-4} volts, when the frequency employed does not make the input impedance of the first tube too low.

In the choice and arrangement of the elements of this system and for many of the details of its adjustment recognition is particularly due to the extensive contributions of E. Peterson, W. A. Mueller and C. R. Keith.

Transatlantic Radio Telephony

By RALPH BOWN

MANY of the technical and scientific features of Transatlantic Radio Telephony have been discussed individually in considerable detail in engineering papers. Furthermore, through the agency of the newspapers much general information has been published regarding the development of commercial telephone service between the old world and the new. Most of this published material either is sketchy in nature or is concentrated upon some detail of the system and it is difficult to gain from it a connected picture of how the final result was built up through several years of continued effort. The following has been written in an attempt to provide such a connected story.

As soon as the successful experiments carried out by the Bell System engineers in 1915 had resulted in the reception of intelligible speech in Paris and Honolulu transmitted from near Washington, D. C., it became a foregone conclusion that sooner or later a serious attempt would be made to bridge the Atlantic Ocean by radio telephone service which would be available to the public at large.

While the 1915 experiments were successful, they also served to emphasize the tremendous difficulties which had to be overcome. The onset of war activities prevented continuing a direct attack on these difficulties but the developments incidental to the wartime use of radio had a profound effect on the instrumentalities necessary to their solution. In particular the development of vacuum tubes for transmitting purposes made considerable progress. Other radio developments carried out immediately subsequent to the war also aided the program.

When transatlantic telephony was taken up again for active consideration, it was obvious that the first requirement was for a transmitting station which would be sufficiently powerful to deliver satisfactory signals on the other side of the ocean. Since the amount of power which would be required to do this was unknown, it was decided to construct a transmitter which was sufficiently large to approach the economic limit of what it seemed it could be worth while to attempt to employ in a commercial undertaking. For this purpose there were available water-cooled vacuum tubes[1] each capable of handling about

[1] "A New Type of High-Power Vacuum Tube," W. Wilson: BELL SYSTEM TECHNICAL JOURNAL, Vol. 1, No. 1, July 1922, pp. 4–17.

10 kw. of power. It was decided that about 20 of these tubes were as many as could be reasonably expected to work satisfactorily in a parallel combination. In order to use these powerful tubes in the most advantageous and economical way, the transmitter was constructed to radiate what is called a single sideband carrier eliminated transmission.

In the ordinary radio telephone transmission such as is used in broadcasting, the radiation sent out consists of a carrier frequency together with two sidebands. The carrier transmits no intelligence but the complete message is transmitted in duplicate since each sideband contains the entire message. By eliminating one of the sidebands and the carrier it is possible to send out the intelligence using only one sideband. If the entire power capacity of the transmitting system is thus concentrated on a single sideband, the power is used several times more effectively.

Since it is more difficult to filter a single sideband away from its carrier and its brother sideband as the frequency becomes higher it was decided to produce the single sideband at a relatively lower frequency as is done in wire carrier telephony and then step it up by a modulation process to the desired position in the frequency range.[2] The voice was therefore modulated upon a 30-kilocycle carrier, and the single sideband produced by passing the modulated result through a band pass filter. This band is then combined in a second modulator with a frequency of 90 kilocycles and the resulting difference frequency, which is a sideband at 60 kilocycles, after passing through another band filter is ready to be amplified to high power for radiation from the antenna. Four preliminary stages of amplification are necessary before the final high power 20-tube amplifier is reached.[3]

When this transmitting apparatus first became available for experimental trial, it was set up at the large radio station at Rocky Point, Long Island, since the experiments were at that time being made in cooperation with the Radio Corporation of America, and the Radio Corporation arranged to lend one of its large and efficient antennas[4] at that station. Subsequently this antenna was leased for use in the final experiments and in giving a commercial service.

[2] "Production of Single Sideband for Transatlantic Radio Telephony," R. A. Heising: *Proceedings of the Institute of Radio Engineers*, Vol. 13, No. 3, June 1925, pp. 291–312.

[3] "Power Amplifiers in Transatlantic Radio Telephony," A. A. Oswald and J. C. Schelleng: *Proceedings of the Institute of Radio Engineers*, Vol. 13, No. 3, June 1925, pp. 313–361.

[4] For a description of this type of antenna known as a multiple tuned antenna see "Transatlantic Radio Communication," E. F. W. Alexanderson: *Proceedings of the American Institute of Electrical Engineers*, Vol. XXXVIII, Part II, 1919, p. 1089.

Simultaneously with the development of this transmitting apparatus, the art of measuring the strength of received radio signals and the amount of static, or radio noise, present at a receiving station had been developed.[5] Therefore in order to try out the effectiveness of the transmitting apparatus, engineers provided with suitable measuring equipment were dispatched to England and set up their apparatus near London. Satisfactory signals were received from the Rocky Point transmitter and in January 1923, it was possible to demonstrate one-way talking across the Atlantic Ocean on a much more satisfactory basis than had previously been possible.

Then there ensued a program of weekly tests wherein signals were sent from Rocky Point each hour for the 24 hours of one day each week and measurements of received signals, radio noise, and intelligibility tests of spoken words were made in England. This one-way telephone circuit was in other words used as a sample whereby the variations to which radio telephony is subject could be explored, catalogued and studied over an extended period of time so estimates could be made of the improvements which would be necessary before anything in the way of reliable communication could be established.

The British General Post Office became so interested in the subject as the result of the initial experiments that they decided to cooperate with the American Company to the fullest extent in determining what the possibilities of transatlantic radio telephony were. They therefore constructed an experimental receiving station and made arrangements to have a transmitter similar in general character to that being used at Rocky Point installed in the new high-powered telegraph station then under construction at Rugby.[6]

The study of transmission initiated in 1923 has been continued to the present time and a large volume of statistical information has been collected.[7] There are two main kinds of variation which have to be contended with. First, the strength of signal changes radically

[5] "Radio Transmission Measurements," Ralph Bown, C. R. Englund and H. T. Friis: *Proceedings of the Institute of Radio Engineers*, Vol. 11, No. 2, April 1923, pp. 115–152.

[6] "The Rugby Radio Station of the British Post Office," E. H. Shaughnessy: *Journal of the Institution of Electrical Engineers*, Vol. 64, June 1926, pp. 683–713. Also "Transatlantic Radio Telephony. Radio Station of the British Post Office at Rugby," E. M. Deloraine: *Electrical Communication*, Vol. 5, July 1926, pp. 3–21.

[7] "Transatlantic Radio Telephony," H. D. Arnold and Lloyd Espenschied: Bell System Technical Journal, Vol. II, No. 4, pp. 116–144, or *Journal of the American Institute of Electrical Engineers*, August 1923. Also "Transatlantic Radio Telephone Transmission," Lloyd Espenschied, C. N. Anderson and Austin Bailey: Bell System Technical Journal, Vol. IV, No. 3, pp. 459–507, or *Proceedings of the Institute of Radio Engineers*, Vol. 14, No. 1, February 1926, pp. 7–56.

with the time of day, being stronger at night. Second, the amount of radio noise present is usually less in the morning and increases towards the evening and well into the night. It is not the absolute strength of the signal which is controlling, but the extent to which it dominates the noise, therefore the ratio between the signal and the noise is the thing which indicates the satisfactoriness with which communication can be carried on. While signal transmission does not change widely between summer and winter, the amount of noise present in the summer time is usually very much greater than that in the winter time, so that the difficulties of communication in the summer are greatly increased.

It soon became apparent that the amount of increase in the signal-to-noise ratio which would be necessary to put conversation on anything like a practical basis would be so great that to hope to get it by increasing the power at the transmitting end was quite out of the question. Thus improvement had to be looked for at the receiving end and the problem became one not of increasing the signal strength, but of decreasing the amount of noise which was allowed to get into the receiving set along with the signal. There are three known ways of decreasing the effect of static in a case of this kind.

Since static is distributed over the entire frequency range, the first and most obvious one is to use such high selectivity that only the signal frequencies are permitted to enter the receiving set. This reduces the amount of static to that which is encompassed by the frequency band occupied by the signal.[8] If suitable band filters are employed it is possible to obtain a degree of selectivity such that practically all static noise which can be eliminated by this method is prevented.

A second method of reducing the amount of static is to employ receiving antenna systems which are directional, in other words, systems which are receptive only to signals coming from the direction of the transmitting station and are blind to interfering signals or interfering static coming from other directions. The most practical system which has so far been developed for doing this at long wave-lengths is the so-called wave antenna.[9] This consists of an open wire line three or four miles long which is grounded at both ends in the characteristic impedance of the wire-to-ground circuit. Thus it is substantially an aperiodic system, there being no reflections at the terminals. Radio waves which approach this line from the side produce relatively very

[8] "Selective Circuits and Static Interference," John R. Carson: BELL SYSTEM TECHNICAL JOURNAL, Vol. IV. No. 2, April 1925, pp. 265–279.

[9] "Wave Antennæ," H. H. Beverage, Chester W. Rice and E. W, Kellogg: *Journal of the American Institute of Electrical Engineers*, Vol. 42, March 1923, pp. 258–269; May 1923, pp. 510–519; and July 1923, pp. 728–738.

small currents in the grounding wires at the terminals. Waves which approach longitudinally build up currents in the line wire as they proceed along it and cause relatively large currents in the grounding wire at the distant end. By building such a line so that it points towards the transmitting station and attaching the receiving set at the end most distant from the transmitter it is possible to obtain a considerable degree of directivity. Further development of this simple basic antenna system is obtained by attaching to it various balancing arrangements. Or by combining several antennas together, it is possible still further to improve the directional characteristics.

In order to determine how much advantage could be obtained by the use of wave antenna systems, the British Post Office constructed one at their receiving station in England and the use of one at Riverhead, L. I. was borrowed by the telephone company from the Radio Corporation of America. Measurements taken over a considerable period of time showed that the signal-to-noise ratio on a good system of combined wave antennas was about ten times as great as that on a simple loop antenna. This was very gratifying since it meant that by the use of wave antennas the transmission would be improved just as much as if the transmitting station power had been multiplied 100 times.

At the receiving end quite aside from the special nature of the directive antenna systems, the amplifying and detecting apparatus must be of a character suited to the amplification and detection of single sideband-carrier eliminated signals. In order to detect signals of this kind it is necessary that a carrier be resupplied to the signal before detection. This is done by means of a local oscillator. Thus the signal actually supplied to the detector differs from the ordinary signals such, for instance, as are used in broadcasting, only by the fact that one of the sidebands is missing. This is of no importance since the complete signal may be detected from the carrier and one of the sidebands. The actual apparatus being used at the American receiving station is similar to the modulating apparatus employed at the transmitting end for producing a single sideband in that the process at the receiving end is substantially reversed. By means of a double detection type receiving set having a beating oscillator frequency of about 90 kc., the 60 kc. incoming sideband is reduced to approximately 30 kc. in the first detection. It is then passed through filters, amplified and has added to it the carrier frequency. The second detection brings it back to voice frequency and after further amplification it is ready to go on to the wire line to the terminal.

A third way in which it is possible to avoid the effects of static is

to place the receiving station in a more northerly latitude, bringing the signals down to the business centers by wire. In order to determine the extent to which this was useful, measurements were made at Green Harbor, Mass., and at Belfast, Me., over a considerable period of time. These measurements were made on telegraph signals specially transmitted by the British General Post Office from its telegraph stations. It was found that in Maine the signal-to-noise ratio was, at least during the important hours of the day, something like six or eight times that which was obtained on Long Island. In other words, the improvement was as great as would have been obtained by multiplying the power of the English transmitting station some fifty times. It was therefore decided to build a receiving station equipped with wave antennas at Houlton, Me.

These two improvements, the one in the antenna and the other in its location, taken together comprise an astounding advance in the battle against static. To get the same results by increasing the power of the transmitting station while using older receiving methods it would be necessary to employ transmitting apparatus rated at one million kilowatts, a power which is obviously far beyond either the technical or economic possibilities.

The British Post Office, having in mind that all Great Britain was already more northerly in latitude than Maine, decided to build a temporary receiving station near Wroughton, England, leaving the question of a more northern location for later experiments.

On both sides of the Atlantic, suitable wire circuits had been arranged to tie the transmitting and receiving stations, to terminal points in New York and London. There were then available early in 1926 the means whereby a complete channel could be set up from New York to London and one from London to New York. These two channels were operated on different radio frequencies, the American transmitter sending on 57 kc. while the British transmitter sent at about 52 kc.

At this point, with the major radio problems if not solved, at least well in hand, the undertaking became more a telephone toll circuit problem for the time being. The simplest way to connect up a system of this kind is to follow the practice which is employed for long 4-wire telephone circuits. Where the circuit needs to become a 2-wire circuit for termination in an office where it may be switched to subscribers, the outgoing and incoming wires are brought together through a hybrid coil or 3-winding transformer. This well-known device, by means of a balancing arrangement, has the property of directing currents incoming on the receiving leg of the 4-wire circuit into the

2-wire line without permitting them to go into the outgoing leg of the 4-wire circuit. The currents coming from the 2-wire line go into both sides of the 4-wire circuit but travel on the receiving leg only until they meet with a repeater which, being directed against them, prevents further travel. The amount of amplification which can be maintained in such a circuit is dependent upon the effectiveness of the balance maintained between the real 2-wire line and the artificial line or network at the hybrid coil. The transatlantic circuit was set up for initial two-way experiments in accordance with this procedure. Since the east-bound and west-bound channels were on different frequencies, the selectivity of the receiving sets prevented any cross-fire from the local transmitter into the local receiving circuit.

Since it is necessary to deliver signals to the distant receiving station of the maximum possible amplitude in order to maintain a favorable signal-to-noise ratio, it was essential that the transmitters be kept loaded up to full output even though the voice currents coming from the speakers might vary widely due to differences in voices and differences in attenuation of connected 2-wire circuits. This was done by changing the gain in the repeaters, the operation being carried out by a control operator in a manner similar to that employed in broadcasting stations. In order to maintain the overall gain around the circuit constant to avoid singing difficulties, it was necessary to change the amplification at the receiving end in such a manner as to compensate for changes at the transmitting end.

Experimental operation of the system on this basis was hindered by the fact that the two frequency bands being employed for the two oppositely directed channels were also being used by a number of radio telegraph stations, some of these being so powerful as to produce interfering signals which very seriously hampered telephone conversation. It was evident that some arrangement must be made to enable the telephone communications to be carried on in frequency bands which were used by them exclusively. The fact that radio telephony inherently requires a wider band for its accomplishment than does radio telegraphy made it desirable to use every device available to narrow the band occupied in order to reduce to a minimum the necessary displacement of existing telegraph services. The employment of single sideband carrier eliminated transmission had already cut in half the frequency space required over that which would be needed if the ordinary form of modulated transmission were used. In order to cut down still further the width of frequency band occupied it was decided to attempt to operate both the east-bound and the west-bound channels on exactly the same frequency band. If this

could be done the entire system would utilize only about 3000 cycles.

In this sort of arrangement, it is evident at once that selectivity at the receiving station is of no further avail in preventing interference from the local transmitter and that unless means are provided greatly to reduce this crossfire or to set up the circuit in some fashion so that it is harmless, the local circuit from transmitter to receiver with return by wire will be in a singing condition, since it is not practicable to obtain at the hybrid coil anything like a sufficient balance to prevent this.

It was found that with the American receiver in Maine some 500 miles from the transmitting station, the local signals were so reduced by distance that the further reduction which could be obtained by virtue of the antenna directional characteristics was sufficient to permit operation. At the English end, however, due to the proximity of the receiving station to the transmitting station, this so-called "radio balance" method of operating could not be employed. This difficulty had been foreseen and there had been developed a switching device based upon certain similar switching devices called echo suppressors which are employed in long toll circuits.[10] The function of this apparatus was in part to supplement the hybrid coil in its office of preventing received signals from getting into the transmitting line. The arrangement is one in which switching means are employed alternately to disable the transmitting or receiving side of the radio circuit automatically in response to the voice currents produced by the speakers at the two ends. Each end of the system was provided with a device of this character operating on substantially the same principles. Briefly, the functioning of the device is as follows:

When no one is speaking on the circuit the transmitting voice paths are blocked at both the New York and London ends of the system but the receiving paths are open so that incoming radio signals pass freely through to the ears of the subscribers. When a speaker, for instance, in America, speaks, his voice currents actuate the device to block off his receiving path and to open his transmitting path so that his voice goes out. Since the other end of the circuit is in a receiving condition, the voice currents travel through the entire system to the listener's ear. When the American speaker has finished, his apparatus is automatically restored to the receiving condition and the British speaker is, by the functioning of the apparatus in London, able to

[10] "Echo Suppressors for Long Telephone Circuits". A. B. Clark and R. C. Mathes: *Journal of the American Institute of Electrical Engineers*, Vol. XLIV, June 1925, pp. 618–626. Also "Telephone Repeaters," C. Robinson and R. M. Chamney: *The Electrician*, December 12, 1924, pp. 665–667.

speak through the circuit. Certain interlocking arrangements are provided so that the voice of only one speaker can go through the entire system at one moment. In this way, the two speakers are prevented from talking simultaneously without either one hearing the other. In addition to facilitating two-way operation of the radio channels on the same frequency band the voice operated devices have other valuable features.

The difficulties of reducing the transmission to the narrowest possible band having been overcome, it was necessary to find a free band of this width. Negotiations by the British Post Office people with European stations and by the American Telephone and Telegraph Company with United States stations finally resulted in the moving of a sufficient number of stations to open up a free band having its central frequency at 60 kc. and this frequency is being employed in service.

The above description covers substantially the system which is being used at the present time for giving the commercial transatlantic radio telephone service. This service is not as yet free from difficulties due to unsatisfactory performance of the radio portions of the system. Further development work is being pursued in an attempt to improve these matters. On the English side the British Post Office, after having made comparative measurements of signals and noise in various parts of Scotland, has undertaken and now has under construction a new receiving station at Cupar near Dundee, Scotland. This will provide the greater freedom from radio noise which can be obtained by increasing the latitude of the receiving station. At both the receiving ends of the system, improvements are being made in the directive characteristics of the receiving antennas.

So far very little has been said about the operation of the system. At the New York and London terminals where the transmitting and receiving circuits join, there is a considerable amount of apparatus which includes the automatic switching devices, the repeaters with their gain controls, and a variety of measuring apparatus for determining and maintaining the characteristics of the entire system. This apparatus is under the charge of men called technical operators. Two of them, one in New York and the other in London, have the duty of maintaining the best possible transmission conditions on the system by making the most favorable adjustments. The local transmitting and receiving stations are under their charge in so far as apparatus adjustments which affect the circuit performance are concerned. Communication between the stations and the terminal is provided by means of order wires.

As the circuit passes out of the realm of the technical operator going

towards the wire system of the country, it consists of an ordinary two-wire trunk which goes to an operating position in the long distance office. At each end of the circuit two telephone operators are employed. One of these operators makes contact with the telephone network in her country to make ready connections for attachment to the transatlantic link. The other operator directs her attention to the transatlantic link and to dealings with her correspondent at the other end in the way of passing call information, making the final connections, pulling down the connections when subscribers have finished, and so on. From the subscribers' standpoint a call is made in the same way as any other long distance call. He asks for "long distance," gives the information regarding the person he wishes to reach in England and then awaits the return call from the long distance operator. When the person called has been located and the transatlantic link is available, the subscriber receives a ring and is connected with his correspondent. They talk back and forth in exactly the same manner as they would over any wire toll circuit and except for the possibility of occasional noises on the circuit which are obviously of radio origin it is difficult for them to realize that their voices are crossing the Atlantic by radio.

A Study of the Regular Combination of Acoustic Elements, with Applications to Recurrent Acoustic Filters, Tapered Acoustic Filters, and Horns

By W. P. MASON

Synopsis: The use of combinations of tubes to produce interference between sound waves and a suppression of certain frequencies originates with Herschel (1833), and was applied by Quincke to stop tones of definite pitch from reaching the ear. Following the development of electrical filters, G. W. Stewart showed that combinations of tubes and resonators could be devised which would give transmission characteristics at low frequencies similar to electrical filters. The assumptions made by Stewart in the development of his theory are that no wave motion need be considered in the elements, and that the lengths of the elements employed are small compared to the wave-length of sound.

The present paper considers primarily regular combinations of acoustic elements, such as straight tubes, and shows that the equations for recurrent filters, tapered filters and horns can be obtained in this manner. The assumption of no wave motion in the elements, made by Stewart, is removed and also account is taken of the viscosity and heat conduction dissipation. The principal difference between acoustic and electric filters is that the former have an infinite number of bands. The effect of using filters between varying terminal impedances is also determined.

Studying next the combination of filters having the same propagation characteristics but in which the conducting tube areas increase in some regular manner, it is shown that a tapered filter results which has a transforming action in addition to its filtering properties. It is shown that if straight tubes are employed and the distance between successive changes in areas is made small we obtain the horn equations first developed by Webster. The general combination of acoustic elements is then considered, and a proof of several theorems has been given.

STEWART, in a series of papers,[1] has studied the recurrent acoustic filter as an analogue of the electric filter with lumped constants. If due account is taken of the wave motion occurring in the individual elements themselves, it appears that the nearest electrical analogue of the acoustic filter is a combination of electric lines.

In the present paper we study primarily regular combinations of acoustic elements, such as straight tubes, and show that the equations for recurrent filters, tapered filters, and horns can be obtained in this manner. The effect of viscosity and heat conduction dissipation has been taken into account, and a consideration of the effect of varying terminal impedances has been included.

I. Equations of Propagation of a Plane Wave in a Uniform Tube

The propagation of plane waves of sound in uniform tubes has been discussed in a number of places,[2] but generally the results obtained are

[1] *Phys. Rev.*, 20, 528 (1922); 23, 520 (1924); 25, 90 (1925).

[2] Rayleigh's "Theory of Sound," Vol. II, p. 318. Lamb's "The Dynamical Theory of Sound," p. 193.

only a determination of the propagation constant, that is, a determination of the attenuation and phase change per unit length, or as more often stated, the attenuation and velocity characteristics. If we solve the differential equations in the manner first employed by Heaviside in the solution of the equation of the electric line, we obtain one more parameter, namely, the characteristic impedance of the tube.

The differential equation, given by Rayleigh,[2] for the propagation of plane waves of sound in a tube of uniform cross-section is

$$\left(1 + \frac{R}{S}\sqrt{\frac{\mu}{2\omega\rho}}\right)\frac{\partial^2\xi}{\partial t^2} + \frac{R}{S}\sqrt{\frac{\mu\omega}{2\rho}}\frac{\partial\xi}{\partial t} = c^2\frac{\partial^2\xi}{\partial x^2}, \tag{1}$$

where ξ denotes the displacement of the fluid at a distance x from one end of the tube,

μ = the coefficient of viscosity of the medium,
ρ = the density of the medium,
R = perimeter and S = cross-sectional area of pipe,
ω = frequency of vibration times 2π,
$C = \sqrt{\frac{P_0\gamma}{\rho}}$ = velocity of sound in medium,
γ = ratio of specific heats of medium.

This equation is valid for tube diameters and frequencies such that

$$\sqrt{\frac{\rho\omega}{2\mu}}\frac{S}{R} > 1$$

and hence can be used for all frequencies of interest in connection with acoustic filters.

Kirchoff [3] extended the theory to take account of the losses due to heat conduction in the medium. His results indicate that in order to take account of this effect, the square root of the coefficient of viscosity should be replaced by a quantity γ', given by

$$\gamma' = \sqrt{\mu} + \left(\sqrt{\gamma} - \frac{1}{\sqrt{\gamma}}\right)\sqrt{v},$$

where v is the coefficient of heat conductivity of the medium. By the kinetic theory of gases v has the value $5/2\ \mu$.

The most useful solution for our present purpose is obtained by writing

$$\xi = e^{i\omega t}(A\cosh\alpha x + B\sinh\alpha x), \tag{2}$$

[3] Rayleigh, "Theory of Sound," Vol. II, p. 325.

where A and B are constants and α by analogy with an electric line is the propagation constant of the tube. Substituting (2) in (1), we see that (2) is a solution provided

$$\alpha^2 = -\frac{\omega^2}{C^2}\left[\left(1 + \frac{R}{S}\sqrt{\frac{\gamma'^2}{2\omega\rho}}\right) - i\frac{R}{S}\sqrt{\frac{\gamma'^2}{2\omega\rho}}\right]. \tag{3}$$

Now α can be written $\alpha = a + ib$, where a is the attenuation constant and b the phase constant. If we solve for a and b, assuming

$$\frac{R}{S}\sqrt{\frac{\gamma'^2}{2\omega\rho}}$$

is a small quantity, we obtain

$$\alpha = a + ib = \frac{1}{2}\frac{R}{CS}\sqrt{\frac{\gamma'^2\omega}{2\rho}} + i\frac{\omega}{C}\left[1 + \frac{1}{2}\frac{R}{S}\sqrt{\frac{\gamma'^2}{2\omega\rho}}\right]. \tag{4}$$

We are generally interested in the volume velocity $S\xi = V$, so we can rewrite equation (2) as

$$V = i\omega S e^{i\omega t}[A \cosh \alpha x + B \sinh \alpha x]. \tag{5}$$

To determine one constant of equation (5), let x equal zero. Then

$$V_{x=0} = V_1 = i\omega e^{i\omega t} SA$$

or

$$A = \frac{V_1}{i\omega S e^{i\omega t}}. \tag{6}$$

We have the additional relation

$$P - P_0 = -P_0\gamma\frac{\partial\xi}{\partial x} = p, \tag{7}$$

where p denotes the excess pressure. Substituting (2) in (7), and differentiating, we have

$$p = -P_0\gamma e^{i\omega t}(A\alpha \sinh \alpha x + B\alpha \cosh \alpha x).$$

Putting $x = 0$, we have

$$p_{x=0} = p_1 = -P_0\gamma e^{i\omega t}(B\alpha)$$

or

$$B = -\frac{p_1}{\alpha P_0\gamma e^{i\omega t}}. \tag{8}$$

Substituting the value of A and B in (5) and (7), we have

$$\left.\begin{aligned} V &= V_1 \cosh \alpha x - \frac{p_1 i\omega S \sinh \alpha x}{P_0\gamma\alpha}, \\ p &= p_1 \cosh \alpha x - V_1 \frac{(P_0\gamma\alpha)}{i\omega S} \sinh \alpha x. \end{aligned}\right\} \tag{9}$$

$(P_0\gamma\alpha)/(i\omega)$ is, by analogy with the electric line, the characteristic impedance [4] per square centimeter of the tube. It is the ratio of p_1/ξ_1 for an infinitely long tube. For since $\cosh \alpha x = \frac{1}{2}(e^{ax} + e^{-ax})$ while $\sinh \alpha = \frac{1}{2}(e^{ax} - e^{-ax})$, then when x approaches infinity, and dissipation exists in the tube, $\cosh \alpha x$ approaches $\sinh \alpha x$, and both approach infinity. Hence the ratio of P_1/V_1 equals $P_0\gamma\alpha/i\omega S$. The propagation constant α has the physical significance that e^{-ax} equals the ratio of V to V_1 or p to p_1, when we are dealing with an infinitely long tube, as can be seen by substituting $p_1/V_1 = P_0\gamma\alpha/i\omega S$ in (9) and solving for the above ratios. The real part of α, i.e. a, determines the rate at which the linear or volume velocity, or pressure, decreases with distance, while the imaginary part b determines the phase of pressure or velocity with respect to the initial values, and hence is known as the phase constant and gives the phase rotation per unit length of pipe. Now since the velocity of propagation C' is

$$C' = \frac{\omega}{b},$$

we have by equation (4)

$$C' = C\left[1 - \frac{1}{2}\frac{R}{S}\sqrt{\frac{\gamma'^2}{2\omega\rho}}\right].$$

The attenuation constant and the velocity reduce to the familiar Helmholtz formulæ, for circular sections.[5]

We write (9) as

$$\left.\begin{aligned} V &= V_1 \cosh \alpha x - \frac{p_1 S}{Z_L} \sinh \alpha x, \\ p &= p_1 \cosh \alpha x - \frac{V_1 Z_L}{S} \sinh \alpha x, \end{aligned}\right\} \tag{10}$$

where Z_L represents the specific characteristic impedance $P_0\gamma\alpha/i\omega$.

[4] The analogy between pressure and electromotive force, volume velocity and current, and impedance to ratio of pressure and volume velocity was first pointed out by Webster[9]. Another system in which force and e.m.f., and linear velocity and current are related, is very convenient when we are dealing with combinations of mechanical elements such as masses and elasticities and no account has to be taken of the area. In the first system, the total impedance is Z_L (per sq. cm.) divided by S whereas in the second system it is $Z_L S$. We follow the first system expressing, however, the impedance in terms of the impedance per square centimeter, which is the same on either systems of units.

[5] See Lamb, "Dynamical Theory of Sound," p. 193, or Rayleigh, "Theory of Sound," Vol. II, p. 319.

The value of the specific characteristic impedance $P_0\gamma\alpha/i\omega$ becomes on substituting in the value of α

$$Z_L = \sqrt{P_0\gamma\rho}\left[\left(1+\frac{1}{2}\frac{R}{S}\sqrt{\frac{\gamma'^2}{2\omega\rho}}\right) - i\frac{1}{2}\frac{R}{S}\sqrt{\frac{\gamma'^2}{2\omega\rho}}\right]. \qquad (11)$$

If we assume no dissipation, $\gamma' = 0$ and $Z_L = \sqrt{P_0\gamma\rho}$. In any case at fairly high frequencies Z_L approaches $\sqrt{P_0\gamma\rho}$. For example, for air in a circular tube 1 centimeter in diameter, Z_L departs from its final value $\sqrt{P_0\gamma\rho}$ by less than 5 per cent at 100 cycles. The attenuation constant a increases as the square root of the frequency, while the phase constant b is little affected by the dissipation and at high frequencies approaches the value ω/C.

II. Effect of a Junction or of a Change in Area of the Conducting Tube

Suppose that we have a straight conducting tube, with a sidebranch as shown in Fig. 1. Let the excess pressure of the incoming plane wave

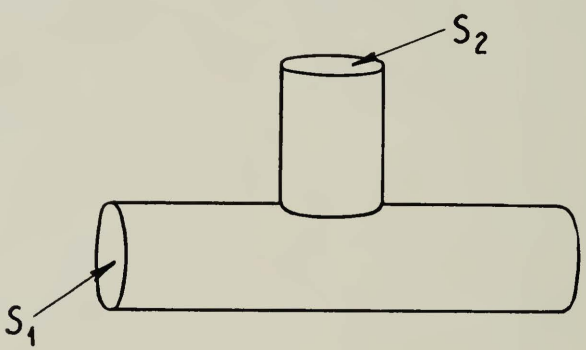


Fig. 1—An acoustic junction

be p_1. The ordinary assumption is that the width of the junction is small compared with a wave-length and hence the pressure is practically constant in the sidebranch, and main branch over the portion in immediate contact with the sidebranch. It states also that the algebraic sum of the volume displacements at a junction of tubes is zero. If S_1 is the area of the main conducting tube, S_2 the area of the branch tube, $\dot{\xi}_1$ the linear velocity of the incoming wave in the conducting tube, $\dot{\xi}_2$ the linear velocity of the outgoing wave from the junction and $\dot{\eta}$ the linear velocity in the branch tube at the junction, we can write the equation

$$\dot{\xi}_1 S_1 = \dot{\xi}_2 S_1 + \dot{\eta} S_2 \qquad \text{or} \qquad V_1 = V_2 + V'.$$

We have now that $\dot{\eta} = p_1/Z_S$ where Z_S is the impedance per unit area of the sidebranch, or the ratio of the excess pressure to the linear velocity. Substituting this value in the above equation, we have

$$\left.\begin{aligned} V_2 &= V_1 - \frac{p_1 S_2}{Z_S}. \\ \text{We have also} \qquad p_2 &= p_1, \end{aligned}\right\} \tag{12}$$

where p_2 is the excess pressure in the conducting tube on the outgoing side. The equations are exactly equivalent to Kirchoff's laws, and hence any equation for a combination of acoustic elements will also apply to the combinations of equivalent electric elements.

A slightly better approximation than the above has been obtained by solving completely the case of three pistons placed in the sides of a rectangular box. This corresponds closely to the condition considered here, if we have rectangular tubes, since the waves can be considered plane up to the junction point with little possibility of error. The solution obtained indicates that the main effect of the junction point is to add an end correction to all the tubes entering the junction. For example, we will measure the length of the main conducting tube, between sidebranches, from the center of the sidebranches rather than the edge, as the approximation given first would imply. Also the length of the sidebranch should be measured from the center of the conducting tube, rather than the edge. For other types of junctions, different end corrections will apply to the sidebranch tubes. For example if the width of the junction is large compared to the width of the sidebranch, we should expect Rayleigh's theoretical value of .82 R to apply where R is the radius of the sidebranch tube. Hence the equations for a junction are equivalent to Kirchoff's laws with the additional proviso that end corrections shall be added to tubes entering a junction.

The effect of a change of area of the conducting tube can be obtained with the same assumptions as above. If we have one conducting tube of area S_1, joined to a second of area S_2, we can write

$$\dot{\xi}_1 S_1 = \dot{\xi}_2 S_2 \qquad \text{or} \qquad V_1 = V_2, \tag{13}$$

where $\dot{\xi}_1$ is the linear velocity in the first tube and $\dot{\xi}_2$ in the second tube. We have also that the pressures in the adjoining tubes are equal. Hence

$$p_2 = p_1 \qquad \text{and} \qquad V_2 = V_1. \tag{14}$$

This equation is of the same order of approximation as the second approximation given above for a junction, since we measure the length from one change of area to the next change.

Equation 14 has been found to hold well as long as the change in area is small while equation 12 holds well as long as the length of a junction is less than half of a wave-length.

III. Recurrent Filters

With the aid of equations (10), (12), and (14), we can obtain the propagation characteristics of any structure employing straight tubes, sidebranches, and changes in area of conducting tubes.

Among the simplest of these are recurrent filters. Fig. 2 shows an

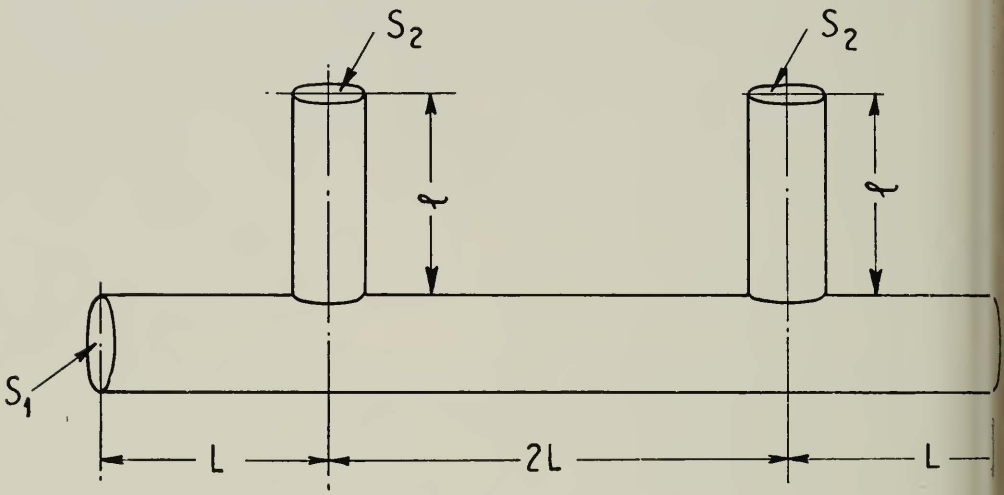


Fig. 2—A typical acoustic filter

example of this type of structure, a main conducting tube, with equally spaced sidebranches. In order to make the structure symmetrical, we let the distance L between one end and the first sidebranch equal one half the distance between two sidebranches. We can then write with regard to the first tube

$$\left.\begin{aligned} V_2 &= V_1 \cosh \alpha_1 L - \frac{p_1}{Z_{L_1}} S_1 \sinh \alpha_1 L, \\ p_2 &= p_1 \cosh \alpha_1 L - V_1 \frac{Z_{L_1}}{S_1} \sinh \alpha_1 L, \end{aligned}\right\} \tag{15}$$

where α_1 and Z_{L_1} refer to the conducting tube. For the junction, we have by (12)

$$\left.\begin{aligned} V_3 &= V_2 - \frac{p_2}{Z_S} S_2, \\ p_3 &= p_2. \end{aligned}\right\} \tag{16}$$

Combining with (15), we have

$$\left.\begin{aligned} V_3 &= V_1\left(\cosh \alpha_1 L + \frac{Z_{L_1} S_2}{Z_S S_1}\sinh \alpha_1 L\right) \\ &\quad - p_1 S_1\left[\frac{\sinh \alpha_1 L}{Z_{L_1}} + \frac{S_2 \cosh \alpha_1 L}{Z_S S_1}\right], \\ p_3 &= p_1 \cosh \alpha_1 L - \frac{V_1 Z_{L_1}}{S_1}\sinh \alpha_1 L. \end{aligned}\right\} \tag{17}$$

The pressures and volume velocities p_4 and V_4 at one half the distance between the first and second sidebranches are again

$$\left.\begin{aligned} V_4 &= V_3 \cosh \alpha_1 L - \frac{p_3 S_1}{Z_{L_1}}\sinh \alpha_1 L, \\ p_4 &= p_3 \cosh \alpha_1 L - V_3 \frac{Z_{L_1}}{S_1}\sinh \alpha_1 L. \end{aligned}\right\} \tag{18}$$

Combining with (17), we obtain

$$\left.\begin{aligned} V_4 &= V_1\left(\cosh 2\alpha_1 L + \frac{Z_{L_1} S_2}{2Z_S S_1}\sinh 2\alpha_1 L\right) \\ &\quad - \frac{p_1 S_1}{Z_{L_1}}\left(\sinh 2\alpha_1 L + \frac{Z_{L_1} S_2}{Z_S S_1}\cosh^2 \alpha_1 L\right)\cdot \\ p_4 &= p_1\left(\cosh 2\alpha_1 L + \frac{Z_{L_1} S_2}{2Z_S S_1}\sinh 2\alpha_1 L\right) \\ &\quad - \frac{V_1 Z_{L_1}}{S_1}\left(\sinh 2\alpha_1 L + \frac{Z_{L_1} S_2}{Z_S S_1}\sinh^2 \alpha_1 L\right)\cdot \end{aligned}\right\} \tag{19}$$

These equations apply to the first section of the filter. By comparison with equation (10) we see that we can write equation (19) as

$$\left.\begin{aligned} V_4 &= V_1 \cosh \Gamma - \frac{p_1 S_1}{Z_0}\sinh \Gamma, \\ p_4 &= p_1 \cosh \Gamma - V_1 \frac{Z_0}{S_1}\sinh \Gamma, \end{aligned}\right\} \tag{20}$$

where

$$\cosh \Gamma = \left(\cosh 2\alpha_1 L + \frac{Z_{L_1} S_2}{2Z_S S_1}\sinh 2\alpha_1 L\right),$$

$$Z_0 = Z_{L_1}\sqrt{\frac{1 + \frac{Z_{L_1} S_2}{2Z_S S_1}\tanh \alpha_1 L}{1 + \frac{Z_{L_1} S_2}{2Z_S S_1}\coth \alpha_1 L}} \tag{21}$$

and

$$\sinh \Gamma = \sinh 2\alpha_1 L \sqrt{\left(1 + \frac{Z_{L_1} S_2}{2 Z_S S_1} \tanh \alpha_1 L\right)\left(1 + \frac{Z_{L_1} S_2}{2 Z_S S_1} \coth \alpha_1 L\right)}.$$

Z_0 and Γ are sometimes called the equivalent line parameters. If we have n sections of the type discussed above, we can write n equations of the kind given by (20). If we eliminate all the terms except for the first and last sections, it can be shown that

$$\left.\begin{aligned} V_n &= V_1 \cosh n\Gamma - \frac{p_1 S_1}{Z_0} \sinh n\Gamma, \\ p_n &= p_1 \cosh n\Gamma - \frac{V_1 Z_0}{S_1} \sinh n\Gamma. \end{aligned}\right\} \tag{22}$$

We see then that Γ represents the propagation constant of one section and Z_0 its specific characteristic impedance. They have the physical interpretation, that Z_0 represents the specific impedance looking into an infinite sequence of these sections, while Γ represents the ratio of excess pressure or volume velocity between one section and the next, when we are dealing with an infinite number of sections, or with a finite number, terminated in the characteristic impedance of the filter.

It is customary in electric filter design to determine the characteristics of a dissipationless filter, and to regard dissipation as causing a slight change in the filter characteristic, which usually occurs most prominently in the pass bands. If we neglect dissipation, equation (21) becomes

$$\cosh \Gamma = \left[\cos\left(\frac{2\omega L}{C}\right) + \frac{i\sqrt{P_0 \gamma \rho} S_2}{2 Z_S S_1} \sin\left(\frac{2\omega L}{C}\right)\right],$$

$$Z_0 = \sqrt{P_0 \gamma \rho} \sqrt{\frac{1 + \frac{i\sqrt{P_0 \gamma \rho} S_2}{2 Z_S S_1} \tan\left(\frac{\omega L}{C}\right)}{1 - \frac{i\sqrt{P_0 \gamma \rho} S_2}{2 Z_S S_1} \cot\left(\frac{\omega L}{C}\right)}}. \tag{23}$$

The propagation constant Γ is in general a complex number $A + iB$. The real part represents a diminution of the volume velocity or the pressure, while the imaginary part represents a phase change, as can be seen from the fact that the ratio of pressure or volume velocity is

$$\frac{p_2}{p_1} = e^{-\Gamma} = e^{-(A+iB)} = e^{-A}(\cos B - i \sin B).$$

Now $\cosh \Gamma = \cosh (A + iB) = \cosh A \cos B + i \sinh A \sin B$. Hence we see from equation (23), if Z_S is an imaginary quantity, the expression for $\cosh \Gamma$ is always real, and hence either $\sinh A$ or $\sin B$

is always zero. Hence either the attenuation constant A is zero, or the phase shift is zero, π radians or some multiple of π radians. Now since cosh A can never be less than 1 while cos B must lie between $+1$ and -1, then when the expression for cosh Γ is between -1 and $+1$, the attenuation constant A is zero and cos B equals the expression in (23). When the value of cosh Γ is outside the limits ± 1, the phase shift is 0, π, or some multiple and the attenuation constant A is given by the expression in (23).

The specific characteristic impedance Z_0, given in (23), can be shown to be a real quantity within the transmitted band and an imaginary quantity outside the transmitted band.

The type of filter obtained with the structure shown in Fig. 2 depends on the sidebranch impedance Z_S. As long as Z_S is of such a value as to make the expression for cosh Γ greater in magnitude than 1, an attenuation band occurs, while if cosh Γ is less than 1, a pass band occurs. The cut-off frequencies of the band occur when cosh $\Gamma = \pm 1$. From equation (23) the cut-off frequencies occur when

$$Z_S = \frac{i\sqrt{P_0\gamma\rho}S_2}{2S_1}\cot\left(\frac{\omega L}{C}\right) \quad \text{or} \quad Z_S = -\frac{i\sqrt{P_0\gamma\rho}S_2}{2S_1}\tan\left(\frac{\omega L}{C}\right). \tag{24}$$

A. Low Pass Filter

The model shown in Fig. 2 can be used to obtain the different types of recurrent filters possible by acoustic means. One of the simplest types of filters in the electrical case is the low pass filter. No exact analogue of this filter exists in the acoustic case, as every acoustic filter has more than one band, but a filter which passes low frequencies and attenuates high frequencies can be designed.

Suppose that the sidebranch used is a straight tube closed at one end. Then by equation (10), the impedance Z_S, when the tube is terminated in an infinite impedance, is

$$Z_S = Z_{L_2}\coth \alpha_2 l,$$

where Z_{L_2} and α_2 are respectively the specific characteristic impedance and propagation constant of the sidebranch, and l its length measured to the center of the conducting tube. Substituting this in the expression for cosh Γ and Z_0, we have

$$\left.\begin{aligned} \cosh\Gamma &= \left(\cosh 2\alpha_1 L + \frac{Z_{L_1}S_2 \sinh 2\alpha_1 L}{2Z_{L_2}S_1 \coth \alpha_2 l}\right), \\ Z_0 &= Z_{L_1}\sqrt{\frac{1 + \dfrac{Z_{L_1}S_2 \tanh \alpha_1 L}{2Z_{L_2}S_1 \coth \alpha_2 l}}{1 + \dfrac{Z_{L_1}S_2 \coth \alpha_1 L}{2Z_{L_2}S_1 \coth \alpha_2 l}}}. \end{aligned}\right\} \tag{25}$$

If we assume no dissipation, and substitute the values of α_1 and Z_L given in section (I), we have

$$\cosh \Gamma = \left[\cos\left(\frac{2\omega}{C} L\right) - \frac{S_2 \sin\left(\frac{2\omega}{C} L\right)}{2S_1 \cot\left(\frac{\omega}{C} l\right)} \right], \tag{26}$$

$$Z_0 = \sqrt{P_0 \gamma \rho} \sqrt{\frac{1 - \frac{S_2}{2S_1}\left(\frac{\tan \frac{\omega}{C} L}{\cot \frac{\omega}{C} l}\right)}{1 + \frac{S_2}{2S_1}\left(\frac{\cot \frac{\omega}{C} L}{\cot \frac{\omega}{C} l}\right)}}. \tag{27}$$

An example of the type of filter obtained by acoustic means, is given when we let $l = 3L$. Fig. 3 gives a plot of the value of Γ for several ratios of S_2/S_1. Fig. 4 shows the corresponding values of the specific characteristic impedance Z_0.

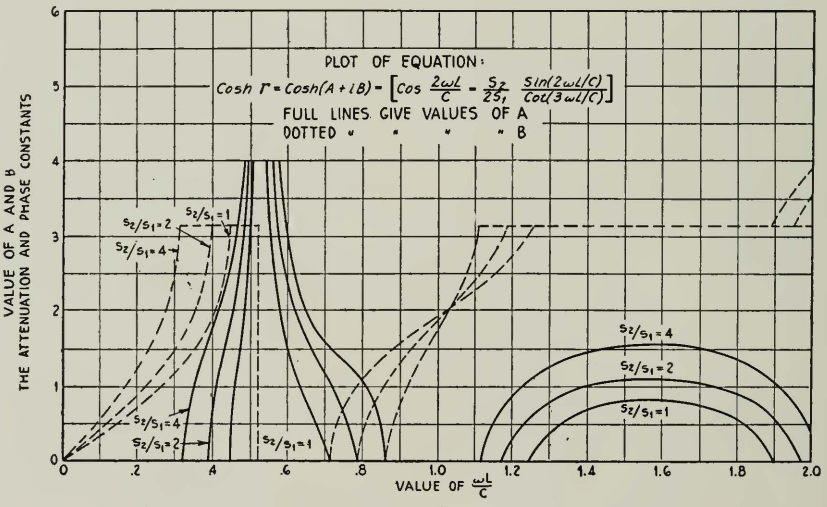


Fig. 3—Propagation constants for a low pass type of filter

A knowledge of Γ will determine the ratio of pressures or volume velocities, if we have an infinite sequence of sections, or if we terminate a finite sequence in the impedance Z_0. If however the terminating impedance is not the characteristic impedance, $e^{-\Gamma}$ no longer represents the ratios of pressures between adjacent sections.

What is generally desired is a knowledge of the effect produced by inserting the filter in a given acoustic system. With the aid of Thévenin's theorem, which is proved for an acoustic system in Appendix I, and equations (20) and (21), this effect can be obtained. Thévenin's

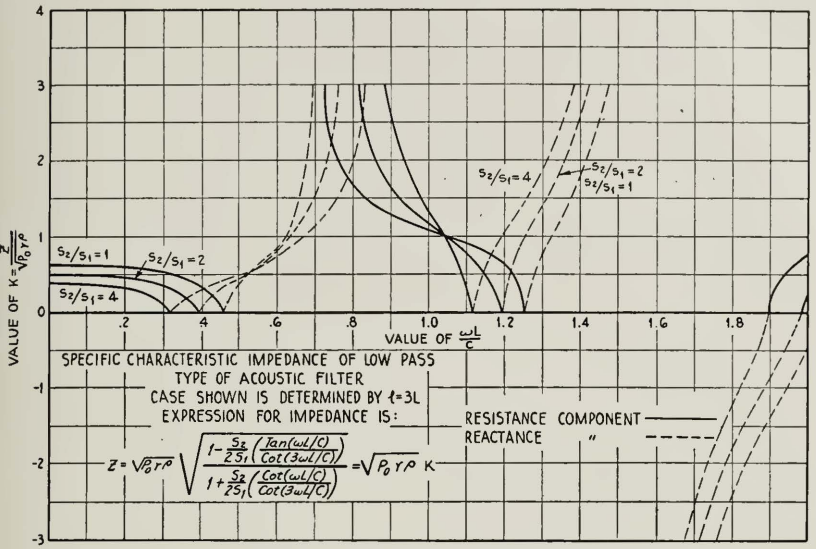


Fig. 4—Specific characteristic impedances for a low pass type of filter

theorem states: If a source of simple harmonic pressure p_0 and of internal impedance Z_T, per square centimeter, is connected to an acoustic system, and if the specific impedance Z_R terminates the system, the volume velocity at the termination of the system will be $p_0'/[(Z_T'/S_1) + (Z_R/S_n)]$, where p_0' is the pressure at the terminating end when this is closed through an infinite impedance, and Z_T' is the impedance per sq. cm. looking back into the acoustic system when this terminated in the impedance Z_T. S_1, and S_n are the areas at the input and output junctions, respectively.

Making use of Thévenin's theorem, the effect of inserting a filter in a given system is the same as the effect obtained by inserting this filter between a source of pressure p_0, with an internal impedance of Z_a/S_1 and a terminating impedance Z_b/S_n, where Z_a/S_1 and Z_b/S_n are respectively the total impedances looking toward the source, and away from the source at the insertion junction of the acoustic system. We have from equation (20)

$$V_2 = V_1 \cosh \Gamma - \frac{p_1 S_1}{Z_0} \sinh \Gamma.$$

$$p_2 = p_1 \cosh \Gamma - \frac{V_1 Z_0}{S_1} \sinh \Gamma,$$

Making use of the above, we can write

$$p_0 = p_1 + \frac{V_1 Z_a}{S_1}.$$

Substituting this, the above equation takes the form

$$\left.\begin{aligned} p_2 &= p_0 \cosh \Gamma - \frac{V_1}{S_1}[Z_0 \sinh \Gamma + Z_a \cosh \Gamma], \\ V_2 &= V_1 \left(\cosh \Gamma + \frac{Z_a}{Z_0} \sinh \Gamma \right) - \frac{P_0 S_1}{Z_0} \sinh \Gamma. \end{aligned}\right\} \tag{28}$$

Eliminating V_1 and substituting $V_2 Z_b / S_1$ for p_2, since here the area remains constant at the two junctions, we have

$$V_2 = \frac{p_0 S_1}{\left[Z_b \cosh \Gamma + \frac{Z_a Z_b}{Z_0} \sinh \Gamma + Z_a \cosh \Gamma + Z_0 \sinh \Gamma \right]}.$$

The most useful way of writing this equation is

$$V_2 = \left(\frac{p_0 S_1}{2 Z_b} \right) \left(\frac{2 Z_0}{Z_0 + Z_a} \right) \left(\frac{2 Z_b}{Z_0 + Z_b} \right) (e^{-\Gamma}) \times \left[\frac{1}{1 - e^{-2\Gamma} \left(\frac{Z_0 - Z_a}{Z_0 + Z_a} \right) \left(\frac{Z_0 - Z_b}{Z_0 + Z_b} \right)} \right]. \tag{29}$$

The volume velocity in the termination of the acoustic system, if the filter were not inserted, is obviously $p_0/[(Z_a/S_1) + (Z_b/S_1)]$. Hence the effect of inserting the filter at any junction is to change the volume velocity of the system by the factor

$$\left(\frac{Z_a + Z_b}{2 Z_b} \right) \left(\frac{2 Z_0}{Z_0 + Z_a} \right) \left(\frac{2 Z_b}{Z_0 + Z_b} \right) (e^{-\Gamma}) \times \left[\frac{1}{1 - e^{-2\Gamma} \left(\frac{Z_0 - Z_a}{Z_0 + Z_a} \right) \left(\frac{Z_0 - Z_b}{Z_0 + Z_b} \right)} \right]. \tag{30}$$

A physical interpretation of equation (30) can be obtained in terms of the transmission and reflection factors first introduced by Heaviside.[6] Heaviside showed that at a junction, a reflection of a wave takes place if the impedances looking towards the source and away from the source are not equal. He showed that the current reflected on striking a junction, will be the unmodified current in the line multiplied by the

[6] Heaviside "Electromagnetic Theory" Vol. II, page 79.

factor, $(Z_I - Z_T)/(Z_I + Z_T)$, while the current transmitted to the terminating side of the junction will be the unmodified current in the line multiplied by the factor $2Z_T/(Z_I + Z_T)$ where Z_I and Z_T are respectively the impedances looking towards and away from the source at the junction. We see then that the second and third factors are transmission factors, determining respectively the transmission from the input impedance Z_a to the inserted structure, and from the inserted structure to the output impedance Z_b. The first factor is the inverse of the transmission factor determining the transmission from the impedance Z_a to the impedance Z_b. The fourth factor is the transfer factor and gives the reduction in volume velocity due to attenuation. The fifth factor has been called the interaction factor, and it gives the change in volume velocity in the termination due to repeated reflections of the volume velocity within the structure. All of these factors reduce to 1 except the transfer factor when $Z_a = Z_b = Z_0$. It will be noted that all factors except the transfer factor cancel out if $Z_a = Z_0$, or $Z_b = Z_0$.

The effect on the pressure due to inserting a filter can be shown to be given also by equation (30).

If the terminating impedances are resistances about equal to an average of the resistance value of Z_0, the effect of these is generally to introduce some loss in the pass band, when the characteristic impedance differs materially from the terminating impedances due to a reflection of the sound wave at the junction points. Since the characteristic impedance of a non-dissipative filter goes either to zero or infinity at the cut-off frequency, the effect of the reflection loss is generally to narrow the pass bands of the filter.

The effect of dissipation, when we take account of the viscosity effects by equations (20) or (21), is two-fold. It changes slightly the position of the band in the frequency range, due to a small change in the velocity of propagation. This is generally negligible. The other effect is to introduce attenuation in the pass band, due to absorption and dissipation of the sound wave.

B. High Pass Filter

An analogous type of high pass filter, which will attenuate the low frequencies and pass the high frequencies, can be made from the structure shown in Fig. 2 by using side tubes which are open on the outer end. The termination at the end of an open tube has been shown by Rayleigh [7] to be a mass with some resistance due to radiation. We could substitute this relation in equation (10) to determine

[7] Rayleigh, "Theory of Sound," Vol. II, p. 106.

the impedance Z_S looking into the sidebranch. Another approximation used with organ pipes is to consider the tube extended by a length .57 times the radius of the tube, and to consider this extended tube terminated in a zero impedance.

The impedance Z_S for this case is from (10)

$$\frac{p_1 S_2}{V_1} = Z_S = Z_{L_2} \tanh \alpha_2 l',$$

where l' is the corrected length of the pipe. Substituting this value in equation (21), we have

$$\cosh \Gamma = \left[\cosh 2\alpha_1 L + \frac{Z_{L_1} S_2 \sinh 2\alpha_1 L}{2 Z_{L_2} S_1 \tanh \alpha_2 l'} \right],$$

$$Z_0 = Z_{L_1} \sqrt{\frac{1 + \dfrac{Z_{L_1} S_2 \tanh \alpha_1 L}{2 Z_{L_2} S_1 \tanh \alpha_2 l'}}{1 + \dfrac{Z_{L_1} S_2 \coth \alpha_1 L}{2 Z_{L_2} S_1 \tanh \alpha_2 l'}}}. \tag{31}$$

For no dissipation these equations become

$$\cosh \Gamma = \left[\cos \left(\frac{2\omega L}{C} \right) + \frac{S_2}{2S_1} \left[\frac{\sin \left(\dfrac{2\omega L}{C} \right)}{\tan \left(\dfrac{\omega l'}{C} \right)} \right] \right],$$

$$Z_0 = \sqrt{P_0 \gamma \rho} \sqrt{\frac{1 + \dfrac{S_2 \tan \left(\dfrac{\omega L}{C} \right)}{2S_1 \tan \left(\dfrac{\omega l'}{C} \right)}}{1 - \dfrac{S_2 \cot \left(\dfrac{\omega L}{C} \right)}{2S_1 \tan \left(\dfrac{\omega l'}{C} \right)}}}.$$

Fig. 5 shows a plot of Γ for several ratios of S_2/S_1, when $l' = 3L$.

C. *Band Pass Type of Filter*

The high pass type of filter discussed above can also be considered as a band pass type of filter, in that an attenuation occurs at zero frequency, then a pass band, and a second attenuation band. A different arrangement of the pass bands can be obtained from the structure shown in Fig. 2, by inserting two sidebranches at one junction point, one of which is open at the outside end and the other closed.

An example of the type of characteristic obtained, is given by the special case where the lengths of both tubes are the same and equal to

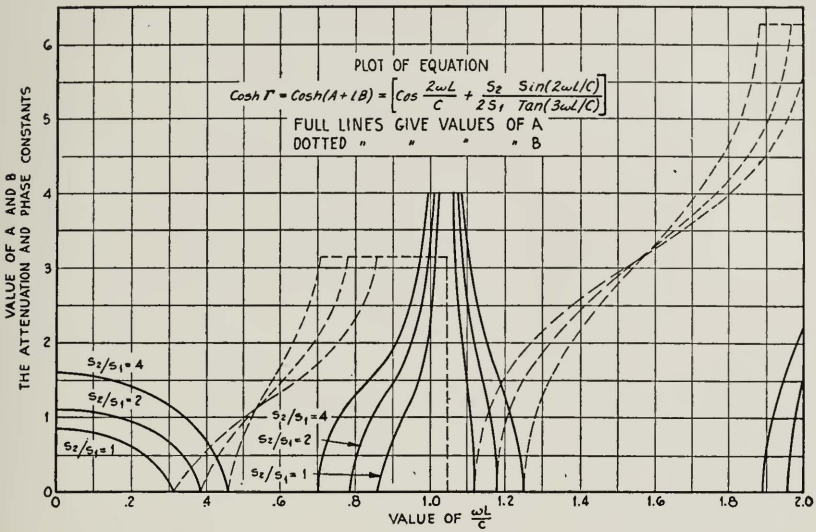


Fig. 5—Propagation constants for a high pass type of filter

$3L$. If S_2 is the area of the open tube and S_3 that of the closed tube, then neglecting dissipation, we find

$$\cosh \Gamma = \cos\frac{2\omega L}{C} + \frac{1}{2S_1}\left[\frac{S_2}{\tan\dfrac{3\omega L}{C}} - \frac{S_3}{\cot\dfrac{3\omega L}{C}}\right]\sin\frac{2\omega L}{C}.$$

A plot of A, the attenuation constant, for several values of S_2/S_1 and S_3/S_1 is given in Fig. 6.

D. Other Types of Sidebranches

We have so far considered only the characteristics obtained where we employ straight tubes. A number of cases can be solved in which the elements employed are not straight tubes although we cannot take account of the viscosity dissipation in these cases. As an example, the characteristics of a filter will be worked out, which employs a straight tube for the conducting tube and conical tubes closed on the end for the sidebranches. We can make use of equation (21) to determine Γ and Z_0, if we insert the proper value of Z_S for the conical tube.

It is evident that for a conical tube, the proper type of wave is a

spherical wave, in place of the plane wave employed for a straight tube. For this case we can write [8] for a simple harmonic wave

$$\frac{\partial^2(r\varphi)}{\partial t^2} = C^2\frac{\partial^2(r\varphi)}{\partial r^2}\ ;\quad \dot{\eta} = -\frac{\partial\varphi}{\partial r} \text{ and } \frac{p}{\rho_0} = \dot{\varphi},$$

where φ is the velocity potential, $\dot{\eta}$ the linear velocity for the spherical wave, p the pressure, ρ the average density of the medium, and r the

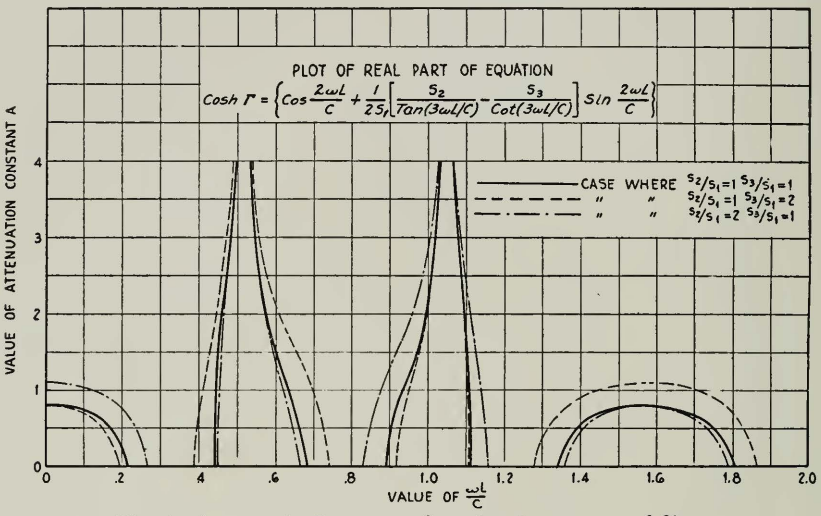


Fig. 6—Attenuation constants for a band pass type of filter

distance from the apex of the cone. The solution for this case is

$$r\varphi = A\sin\frac{\omega}{C}r + B\cos\frac{\omega}{C}r.$$

Hence we can determine $\dot{\eta}$ and p as

$$\dot{\eta} = A\left[\frac{\sin\frac{\omega}{C}r}{r^2} - \frac{\frac{\omega}{C}\cos\frac{\omega}{C}r}{r}\right] + B\left[\frac{\cos\frac{\omega}{C}r}{r^2} + \frac{\frac{\omega}{C}\sin\frac{\omega}{C}r}{r}\right]$$

and

$$p = i\omega\rho\left[\frac{A\sin\frac{\omega}{C}r + B\cos\frac{\omega}{C}r}{r}\right].$$

If now we set $\dot{\eta} = 0$ when $r = x_2$ and determine the ratio of $p/\dot{\eta}$ at

[8] Lamb, "Dynamical Theory of Sound," p. 206. Rayleigh, "Theory of Sound," Vol. II, p. 114.

$r = x_1$, we find

$$Z_S = \frac{p}{\dot{\eta}}$$

$$= -i\sqrt{P_0\gamma\rho}\left[\frac{\cos\frac{\omega}{C}(x_2 - x_1) - \frac{\sin\frac{\omega}{C}(x_2 - x_1)}{\frac{\omega}{C}x_2}}{\cos\frac{\omega}{C}(x_2 - x_1)\left[\frac{1}{\frac{\omega}{C}x_2} - \frac{1}{\frac{\omega}{C}x_1}\right] + \sin\frac{\omega}{C}(x_2 - x_1)\left[1 + \frac{1}{\frac{\omega^2}{C^2}x_1x_2}\right]}\right]. \quad (32)$$

If we substitute this value of Z_S in equation (21), we can readily determine the value of Γ and Z_0. Fig. 7 shows a plot of A and B for this case assuming $(x_2 - x_1) = L$.

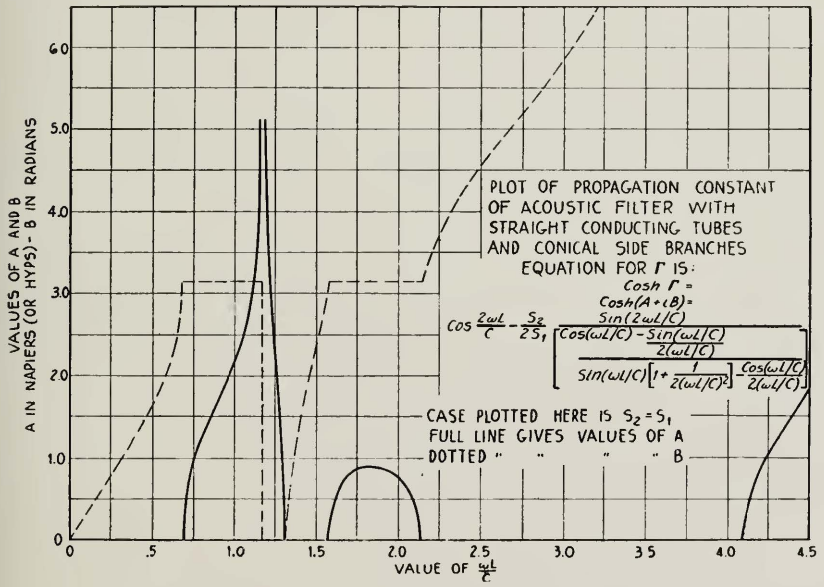


Fig. 7—Propagation constant of a low pass type of filter.

IV. Tapered Filter Structures and Horns

In addition to recurrent filters, other types of filters exist. If, for example, we connect sections with the same propagation constants and

characteristic impedances, but whose conducting tube areas increase in some regular manner, a tapered filter is obtained whose characteristics differ from those of a recurrent filter. The distinguishing property introduced by a tapered filter, in addition to its filtering property, is a transformer action which increases the pressure by a given ratio and decreases the volume velocity in the same ratio, or vice versa, thus giving a transforming action and a complete transmission of power over the pass band. This is a useful property, if acoustic systems of different impedances are to be connected together. Horns are the limiting cases of tapered acoustic filters and hence their study has considerable practical importance.

The typical section of a tapered filter considered here is one built up from two symmetrical structures with the same propagation constants and characteristic impedances per square centimeter, but with different cross-sectional areas. If we use any of the recurrent filters discussed in Section III, then, for example, since

$$\cosh \Gamma = \left[\cos\left(\frac{2\omega L}{C}\right) - \frac{S_2 \sin\left(\frac{2\omega L}{C}\right)}{2S_1 \cot \frac{\omega l}{C}} \right]$$

for the low pass filter, to keep the same value of Γ when we vary the conducting tube area it will be necessary to keep the ratio of the areas constant and to leave all values of L and l the same. Similarly for the other types of filters.

If $\Gamma/2$ is the propagation constant of each of the symmetrical structures, Z_0 the characteristic impedance per square centimeter for each structure, S_1 the cross-sectional area of the first structure and S_2 that of the second, we can write three sets of equations for the two structures and the junction point. These are

$$\left.\begin{aligned} p_1' &= p_1 \cosh\frac{\Gamma}{2} - V_1 \frac{Z_0}{S_1} \sinh\frac{\Gamma}{2}, \\ V_1' &= V_1 \cosh\frac{\Gamma}{2} - \frac{p_1 S_1}{Z_0} \sinh\frac{\Gamma}{2}, \end{aligned}\right\}$$

$$p_1'' = p_1'; \quad V_1'' = V_1',$$

$$\left.\begin{aligned} p_2 &= p_1'' \cosh\frac{\Gamma}{2} - V_1'' \frac{Z_0}{S_2} \sinh\frac{\Gamma}{2}, \\ V_2 &= V_1'' \cosh\frac{\Gamma}{2} - \frac{p_1'' S_2}{Z_0} \sinh\frac{\Gamma}{2}. \end{aligned}\right\}$$

Combining these equations, we obtain

$$\left.\begin{aligned} p_2 &= p_1\left[\left(\frac{S_1+S_2}{2S_2}\right)\cosh\Gamma+\left(\frac{S_2-S_1}{2S_2}\right)\right] \\ &\qquad -\frac{V_1Z_0}{S_1}\left(\frac{S_1+S_2}{2S_2}\right)\sinh\Gamma, \\ V_2 &= V_1\left[\left(\frac{S_1+S_2}{2S_1}\right)\cosh\Gamma-\left(\frac{S_2-S_1}{2S_1}\right)\right] \\ &\qquad -\frac{p_1S_1}{Z_0}\left(\frac{S_1+S_2}{2S_1}\right)\sinh\Gamma, \end{aligned}\right\} \tag{33}$$

or for simplicity we write

$$\left.\begin{aligned} p_2 &= p_1A-\frac{V_1Z_0}{S_1}B, \\ V_2 &= V_1C-\frac{p_1S_1}{Z_0}D. \end{aligned}\right\} \tag{34}$$

In order to express the propagation in terms of some known functions we will first obtain some relations between the impedances of the sections and the ratios of p_2/p_1 and V_2/V_1. We can write the above equations as

$$\frac{p_2}{p_1}=A-\frac{Z_0}{Z_1}B, \quad \frac{V_2}{V_1}=C-\frac{Z_1}{Z_0}D,$$

where $Z_1/S_1 = p_1/V_1$. Eliminating Z_1, we have

$$\frac{p_2}{p_1}\frac{V_2}{V_1}-A\frac{V_2}{V_1}-C\frac{p_2}{p_1}=BD-AC=-1 \tag{35}$$

as can be seen by multiplying together the above expressions. Solving for the ratio of V_2/V_1 in terms of p_2/p_1, A and C, we have

$$\frac{V_2}{V_1}=\frac{C\dfrac{p_2}{p_1}-1}{\dfrac{p_2}{p_1}-A}.$$

Multiplying both sides by p_1/p_2, we have

$$\frac{p_1}{V_1}\frac{V_2}{p_2}=\frac{C-\dfrac{p_1}{p_2}}{\dfrac{p_2}{p_1}-A}.$$

Now since $\frac{p_1}{V_1} = \frac{Z_1}{S_1}$ and $\frac{p_2}{V_2} = \frac{Z_2}{S_2}$, we have

$$\frac{Z_2}{S_2} = \frac{Z_1}{S_1}\left[\frac{\frac{p_2}{p_1} - A}{C - \frac{p_1}{p_2}}\right]. \tag{36}$$

Z_2/S_2 and Z_1/S_1 then are respectively the terminating impedance and input impedance necessary to give a structure specified by the factors A, B, C, D the pressure ratio p_2/p_1. To solve for the input impedance we take the first of equations (34) and obtain

$$Z_1 = \frac{Z_0 B}{A - \frac{p_2}{p_1}}. \tag{37}$$

Hence by virtue of (36), the terminating impedance Z_2/S_2 becomes

$$\frac{Z_2}{S_2} = \frac{\frac{Z_0}{S_1}B}{\frac{p_1}{p_2} - C}. \tag{38}$$

Equations (37) and (38) state that there is a relation between the input impedance and the pressure ratio, and the output impedance and the pressure ratio. When one is specified and the constants of the section Z_0, A, B, C, D are given, the others are known.

Suppose now that we wish to join a second structure of this type to the first, assuming that the cross-sectional area at the junction is the same for both. We must have now Z_2, the specific output impedance of the first section, equal to Z_1', the specific input impedance of the second section. Hence we can write.

$$\frac{\frac{Z_0}{S_1}B}{\frac{p_1}{p_2} - C} = \frac{\frac{Z_0}{S_2}B'}{\left(A' - \frac{p_3}{p_2}\right)} \text{ or } \frac{S_2}{S_1}B\left[A' - \frac{p_3}{p_2}\right] = B'\left[\frac{p_1}{p_2} - C\right],$$

where the primes refer to the constants of the second section and where p_3/p_2 is the pressure ratio of the second section. Substituting in the values of A', B, B', C, we have

$$\left(\frac{(S_1 + S_2)(S_2 + S_3)}{2S_1S_3}\right)\cosh\Gamma + \frac{S_1S_3 - S_2^2}{2S_1S_3} = \left(\frac{S_2 + S_3}{2S_3}\right)\frac{p_1}{p_2} + \left(\frac{S_1 + S_2}{2S_1}\right)\frac{p_3}{p_2}. \tag{39}$$

Equation (39) gives the relationship between p_1/p_2 and p_3/p_2 which must be satisfied if the output impedance of one section equals the input impedance of the next section. If we specify a value of p_2/p_1, then the value of p_3/p_2 is determined. The impedance Z_2' terminating the second section is also determined and hence the pressure ratio of the third section, etc. Hence if we specify a value of p_2/p_1, we also determine the propagation characteristic of any other section in a series of sections. The pressure ratios will not in general be constant from section to section.

We can write $p_2/p_1 = Ke^{-\delta}$ since this will represent any phase or amplitude change. Similarly we can write p_3/p_2 as $K'e^{-\delta'}$. Substituting these values in (39), we have

$$\left(\frac{(S_1+S_2)(S_2+S_3)}{2S_1S_3}\right)\cosh\Gamma+\frac{S_1S_3-S_2^2}{2S_1S_3}=\left(\frac{S_2+S_3}{2S_3}\right)\frac{e^{\delta}}{K}+\left(\frac{S_1+S_2}{2S_1}\right)K'e^{-\delta'}. \tag{40}$$

Now if the value of δ remains unchanged from section to section a great simplification results, for in order to determine the overall pressure ratio we have only to multiply the number of sections by δ. Hence it is desirable to determine for what rate of taper this condition is met and also how good an approximation it is for all rates of taper.

If we set $\delta = \delta'$ and multiply through by $e^{-\delta}$, we obtain

$$e^{-2\delta}-\left[\frac{\left(\frac{(S_1+S_2)(S_2+S_3)}{2S_1S_3}\right)\cosh\Gamma+\frac{S_1S_3-S_2^2}{2S_1S_3}}{\frac{S_1+S_2}{2S_1}K'}\right]e^{-\delta}+\frac{\left(\frac{S_2+S_3}{2S_3}\right)\frac{1}{K}}{\left(\frac{S_1+S_2}{2S_1}\right)K'}=0.$$

Similarly the equation for the next two sections is

$$e^{-2\delta''}-\frac{\left[\left(\frac{(S_2+S_3)(S_3+S_4)}{2S_2S_4}\right)\cosh\Gamma+\left(\frac{S_2S_4-S_3^2}{2S_2S_4}\right)\right]}{\left(\frac{S_2+S_3}{2S_2}\right)K''}e^{-\delta''}+\frac{\left(\frac{S_3+S_4}{2S_4}\right)\frac{1}{K'}}{\left(\frac{S_2+S_3}{2S_2}\right)K''}=0.$$

If we are to have $\delta = \delta''$, we must have

$$e^{-\delta}\left[\left[\frac{S_2+S_3}{S_3K'}-\frac{S_3+S_4}{S_4K''}\right]\cosh\Gamma + \left[\frac{S_1S_3-S_2^2}{(S_1+S_2)S_3K'}-\frac{S_2S_4-S_3^2}{(S_2+S_3)S_4K''}\right]\right] = \left[\frac{(S_2+S_3)S_1}{S_3(S_1+S_2)KK'}-\frac{(S_3+S_4)S_2}{(S_2+S_3)S_4K'K''}\right]. \quad (41)$$

Since the term on the left is complex, while that on the right is a numeric, each must separately vanish if we are to have this equality. Similarly the terms within the bracket of the left hand side would have to vanish.

We see that the two terms on the left do not vanish simultaneously unless we satisfy the progression equation

$$2S_1S_3^2 - 2S_2^2S_4 + (S_3 - S_2)(S_1S_4 + S_2S_3) = 0. \quad (42)$$

This equation is satisfied by a system whose area increases exponentially with the distance. The terms involving $S_1S_3 - S_2^2$ and $S_2S_4 - S_3^2$ are always very small no matter what the rate of progression. Hence is is desirable to see if neglecting these terms we can still satisfy the above conditions. The most useful value of the two terms on the right hand side of equation (41) is 1. Hence setting each term equal to 1 and solving for K' and K'', we find that

$$K' = \sqrt{\frac{S_2}{S_3}}; \quad K'' = \sqrt{\frac{S_3}{S_4}}.$$

We see then that if we neglect second order quantities, we can represent with good approximation the pressure ratio of any tapered filter by the expression

$$\frac{p_2}{p_1} = \sqrt{\frac{S_n}{S_{n+1}}}\, e^{-\delta},$$

where δ is the propagation constant of a tapered structure. For a complete solution, δ is not constant except for a progression which satisfies equation (42).

A. Exponentially Tapered Filters and Horns

If we assume that the area of a given section is e^{2t} times as large as that of the section preceding, equation (40) reduces to

$$2e^{-t}[\cosh\Gamma\cosh t] = K'e^{-\delta} + \left(e^{-2t}\times\frac{1}{K}\right)e^{\delta}. \quad (43)$$

We choose now $K' = \sqrt{\frac{S_2}{S_3}} = e^{-t}$ and $K = \sqrt{\frac{S_1}{S_2}} = e^{-t}$.

Then

$$\cosh \delta = \cosh \Gamma \cosh t.$$

To show that δ is the propagation constant for an infinite sequence of such sections, it is necessary to show that δ is the same for any two sections. But equation (43) holds good for any two sections, and hence δ is the same, and represents a solution for an infinite sequence of sections. Now

$$e^{-\delta} = \cosh \delta - \sinh \delta = \cosh \Gamma \cosh t - \sqrt{\sinh^2 \Gamma \cosh^2 t + \sinh^2 t}.$$

Hence

$$\frac{p_2}{p_1} = Ke^{-\delta} = e^{-t}[\cosh \Gamma \cosh t - \sqrt{\sinh^2 \Gamma \cosh^2 t + \sinh^2 t}]$$

and

$$\frac{V_2}{V_1} = e^{-t}e^{-\delta}\left[\frac{C - e^{t}e^{\delta}}{e^{-t}e^{-\delta} - A}\right]$$
$$= e^{t}[\cosh \Gamma \cosh t - \sqrt{\sinh^2 \Gamma \cosh^2 t + \sinh^2 t}],$$

and hence the pressure and volume velocity have the same propagation constant δ but an inverse multiplying factor.

The specific impedance Z_1, looking into a given section, is by equation (37)

$$Z_1 = \frac{Z_0 B}{A - Ke^{-\delta}} = Z_0\left[\frac{\sqrt{\tanh^2 t + \sinh^2 \Gamma} - \tanh t}{\sinh \Gamma}\right] \qquad (44)$$

and similarly Z_2, the specific terminating impedance, can be shown equal to Z_1. Hence the impedance per square centimeter at the junction points is the same for each section.

To observe the action of a tapered filter, let us obtain the product of the pressure by the volume velocity and see how these are propagated. Since the specific impedance is the same from section to section, this will represent also the power propagation. Now since

$$\cosh \delta = \cosh \Gamma \cosh t,$$

a pass band occurs when $1 \geqq \cosh \delta \geqq -1$ and hence the band occurs only when Γ is imaginary, since $\cosh \Gamma < 1$ and > -1, or when the filter repeated recurrently is in its pass band. Furthermore the pass band for the tapered structure will not be as wide as that for a similar recurrent structure, since for the tapered structure the band

occurs when $\cosh \Gamma = \pm 1/\cosh t$ while in the recurrent structure, the band occurs when $\cosh \Gamma = \pm 1$. One result of this is that no low pass filter exists in exponentially tapered structures.

Considering now the pressure and volume velocity ratios when $1 \geqq \cosh \delta \geqq -1$, the absolute value of $e^{-\delta}$ is 1. Hence over the band the ratios of pressure and of volume velocity from section to section are respectively e^{-t} and e^{t} or $\sqrt{S_1/S_2}$ and $\sqrt{S_2/S_1}$. Hence one section multiplies the pressure by a ratio $\sqrt{S_1/S_2}$, and the volume velocity by the factor $\sqrt{S_2/S_1}$. Therefore a tapered structure of this kind is equivalent to a transformer of turns ratio $\sqrt{S_1/S_2}$, and a filter of somewhat narrower bands than for the filter repeated recurrently.

To specify completely a filter of this type requires three parameters. Two such parameters have been developed above and are δ, the propagation constant of a tapered filter, and Z_{R_1}, the specific recurrent impedance in one direction. These are given by

$$\left.\begin{aligned} \cosh \delta &= \cosh \Gamma \cosh t, \\ Z_{R_1} &= Z_0 \left[\frac{\sqrt{\tanh^2 t + \sinh^2 \Gamma} - \tanh t}{\sinh \Gamma} \right]. \end{aligned}\right\} \tag{45}$$

We take as the third parameter Z_{R_2}, the specific recurrent impedance in the opposite direction. We can readily determine that Z_{R_2}, the impedance looking in the opposite direction from that used to specify Z_{R_1}, but obtained at the same junction point, is

$$Z_{R_2} = \frac{Z_0 B}{\frac{p_1'}{p_2'} - A}.$$

It is desirable to have the same propagation constant serve for the two directions, hence we let $p_2'/p_1' = K_1 e^{-\delta}$. Since K represents a transformer change of the pressure in one direction, we find, when going in the opposite direction, that the pressure should change by the inverse of K, so $K_1 = 1/\mathrm{K}$. Substituting these values for p_1'/p_2',

$$Z_{R_2} = \frac{Z_0 B}{K e^{\delta} - A}. \tag{46}$$

Hence for an exponentially tapered filter

$$Z_{R_2} = Z_0 \left[\frac{\sqrt{\tanh^2 t + \sinh^2 \Gamma} + \tanh t}{\sinh \Gamma} \right]. \tag{47}$$

In terms of the parameters, δ, Z_{R_1} and Z_{R_2} we can express p_2, p_1, V_2,

and V_1 as

$$\begin{aligned} p_2 &= e^{-t}\left[p_1\left[\cosh\delta + \left[\frac{Z_{R_2} - Z_{R_1}}{Z_{R_2} + Z_{R_1}}\right]\sinh\delta \right] \right. \\ &\qquad \left. - \frac{V_1}{S_1}\left[\frac{2Z_{R_1}Z_{R_2}}{Z_{R_1} + Z_{R_2}}\right]\sinh\delta \right], \\ V_2 &= e^{t}\left[V_1\left[\cosh\delta + \left[\frac{Z_{R_1} - Z_{R_2}}{Z_{R_1} + Z_{R_2}}\right]\sinh\delta \right] \right. \\ &\qquad \left. - p_1 S_1\left[\frac{2\sinh\delta}{Z_{R_1} + Z_{R_2}}\right]\right]. \end{aligned} \tag{48}$$

If now the elements of our structure are non-dissipative straight tubes, instead of a general filter structure, and the length of these tubes between changes of area is made very small, it is evident that the structure reduces to an exponential horn. We now let the ratio $S_1/S_2 = e^{-2t}$ be expressed as

$$\frac{S_1}{S_2} = e^{-2Tl} = e^{-2t},$$

where l is the distance between changes in area and T a new taper constant. Then Γ, for a straight tube, neglecting dissipation, becomes $\Gamma = i\omega l/c$ and hence

$$\begin{aligned} \cosh\delta &= \cosh\frac{i\omega l}{C}\cosh Tl \\ &= \left(1 - \frac{\omega^2 l^2}{2!C^2} + \frac{\omega^2 l^2}{4!C^4} + \cdots\right)\left(1 + \frac{(Tl)^2}{2!} + \frac{(Tl)^4}{4!} + \cdots\right) \\ &= 1 + \frac{l^2\left(T^2 - \dfrac{\omega^2}{C^2}\right)}{2!} + l^4\frac{\left(T^4 - 6\dfrac{\omega^2}{C^2}T^2 + \dfrac{\omega^4}{C^4}\right)}{4!} \cdots. \end{aligned}$$

This reduces for small values of l to

$$\cosh\delta = \cosh\left(l\sqrt{T^2 - \frac{\omega^2}{C^2}}\right).$$

Hence

$$\frac{p_n}{p_1} = e^{-nTl}e^{-n\delta} = e^{-nl\left(T + \sqrt{T^2 - \frac{\omega^2}{C^2}}\right)} = e^{-L\left(T + \sqrt{T^2 - \frac{\omega^2}{C^2}}\right)}$$

since $nl = L$, the total length of the horn.

As long as $T^2 > (\omega^2/C^2)$, an attenuation band exists, while if $\omega^2/C^2 > T^2$, the expression becomes

$$\frac{p_n}{p_1} = e^{-TL}\left[\cos\left(L\sqrt{\frac{\omega^2}{C^2} - T^2}\right) - i\sin\left(L\sqrt{\frac{\omega^2}{C^2} - T^2}\right)\right]$$

and a pass band occurs.

The complete equation for the horn, equivalent to equation (48), becomes

$$p_2 = e^{-LT}\left\{\left[\cosh\left(L\sqrt{T^2-\frac{\omega^2}{C^2}}\right) + \frac{T}{\sqrt{T^2-\frac{\omega^2}{C^2}}}\sinh\left(L\sqrt{T^2-\frac{\omega^2}{C^2}}\right)\right]p_1 - \frac{iV_1\sqrt{P_0\gamma\rho}\,\frac{\omega}{C}\sinh\left(L\sqrt{T^2-\frac{\omega^2}{C^2}}\right)}{S_1\sqrt{T^2-\frac{\omega^2}{C^2}}}\right\},$$

$$V_2 = e^{+LT}\left\{\left[\cosh\left(L\sqrt{T^2-\frac{\omega^2}{C^2}}\right) - \frac{T}{\sqrt{T^2-\frac{\omega^2}{C^2}}}\sinh\left[\left(L\sqrt{T^2-\frac{\omega^2}{C^2}}\right)\right]\right]V_1 - \frac{iS_1p_1\frac{\omega}{C}\sinh\left(L\sqrt{T^2-\frac{\omega^2}{C^2}}\right)}{\sqrt{P_0\gamma\rho}\sqrt{T^2-\frac{\omega^2}{C^2}}}\right\}. \tag{49}$$

These expressions can be derived from Webster's [9] differential equations for an exponential horn. Exponential horns have also been discussed by a number of writers.[10]

B. Tapered Filters Whose Area Increases as the Square of the Distance

One other example of a tapered filter, for which an approximate solution can be obtained, will be considered because of its bearing on the straight or conical horn. Let us assume that the area S_1 of a typical section of a tapered filter chain is n^2E, while that of the section next to it is equal to $(n+1)^2E$, where E is a small constant. Sub-

[9] A. G. Webster, "Acoustic Impedance, and The Theory of Horns and of the Phonograph," *Nat. Acad. of Science*, Vol. 5, 1919, p. 275. The solution given by Webster for the exponential horn appears to have some typographical errors.

[10] Hanna and Slepian (*Trans. A. I. E. E.*, 43, 1924, p. 393); H. C. Harrison (British Patent No. 213,525, 1925); I. B. Crandall, "Theory of Vibrating Systems and Sound," D. Van Nostrand, 1926, p. 158.

stituting these values in equation (40), we obtain

$$\left[\left(\frac{(n^2+(n+1)^2)((n+1)^2+(n+2)^2)}{2(n)^2(n+2)^2}\right)\cosh\Gamma + \frac{n^2(n+2)^2-(n+1)^4}{2(n)^2(n+2)^2}\right] = K'\left(\frac{n^2+(n+1)^2}{2(n)^2}\right)e^{-\delta} + \left(\frac{(n+1)^2+(n+2)^2}{2(n+2)^2}\right)\frac{e^{\delta}}{K}.$$

If we substitute $K' = \sqrt{\frac{S_2}{S_3}} = \frac{n+1}{n+2}$ and $K = \sqrt{\frac{S_1}{S_2}} = \frac{n}{n+1}$ and neglect 1 as compared with n^3, we have

$$\cosh\delta = \left[\left(\frac{2n^2+1}{2n^2}\right)\cosh\Gamma - \frac{1}{2n^2}\right]. \tag{50}$$

If again our changes in areas are very small and hence n very large, we can neglect 1 compared with $2n^2$ and obtain

$$\cosh\delta = \cosh\Gamma, \quad \text{or} \quad \delta = \Gamma.$$

Either of these solutions will hold for any other pair of sections if we neglect 1 as compared with n^3 for the first of 1 compared with n^2 for the second. Hence for either solution, the propagation constant is little affected for this type of taper. The specific characteristic impedances Z_{R_1} and Z_{R_2} become

$$Z_{R_1} = \frac{Z_0\sinh\Gamma}{\frac{1}{n}+\left(\frac{2n^2}{2n^2+1}\right)\sqrt{\left(\frac{2n^2+1}{2n^2}\cosh\Gamma-\frac{1}{2n^2}\right)^2-1}},$$
$$Z_{R_2} = \frac{Z_0\sinh\Gamma}{-\frac{1}{n}+\frac{2n^2}{2n^2+1}\sqrt{\left(\frac{2n^2+1}{2n^2}\cosh\Gamma-\frac{1}{2n^2}\right)^2-1}}. \tag{51}$$

If we neglect 1 as compared with n^2, these expressions reduce to

$$Z_{R_1} = \frac{Z_0 n\sinh\Gamma}{1+n\sinh\Gamma}; \quad Z_{R_2} = \frac{Z_0 n\sinh\Gamma}{-1+n\sinh\Gamma}. \tag{52}$$

These impedances represent the impedances per square cm. looking in both directions at the input junction of the filter, whose area is n^2E. As we move in either direction these impedances change since n itself changes. If n becomes sufficiently large and Γ is not zero, the two characteristic impedances approach the value Z_0.

To express p_2 and V_2 in terms of p_1 and V_1 and these three parameters, we can obtain the equations.

$$p_2 = \frac{n}{n+1}\left[p_1 \left[\cosh\delta + \left(\frac{Z_{R_2}{}^I - Z_{R_1}{}^I}{Z_{R_2}{}^I + Z_{R_1}{}^I} \right) \sinh\delta \right] - \frac{V_1}{S_1}\left[\frac{2Z_{R_1}{}^I Z_{R_2}{}^I}{Z_{R_1}{}^I + Z_{R_2}{}^I} \right] \sinh\delta \right],$$

$$V_2 = \frac{n+1}{n}\left[V_1 \left[\cosh\delta + \left(\frac{Z_{R_1}{}^O - Z_{R_2}{}^O}{Z_{R_1}{}^O + Z_{R_2}{}^O} \right) \sinh\delta \right] - p_1 S_1 \left[\frac{Z_{R_1}{}^I + Z_{R_2}{}^I}{2Z_{R_1}{}^I Z_{R_2}{}^I} \right] \times \left[\sinh\delta \left[1 - \left(\frac{Z_{R_2}{}^I - Z_{R_1}{}^I}{Z_{R_2}{}^I + Z_{R_1}{}^I} \right)\left(\frac{Z_{R_2}{}^O - Z_{R_1}{}^O}{Z_{R_2}{}^O + Z_{R_1}{}^O} \right) \right] + \cosh\delta \times \left[\left(\frac{Z_{R_2}{}^I - Z_{R_1}{}^I}{Z_{R_2}{}^I + Z_{R_1}{}^I} \right) - \left(\frac{Z_{R_2}{}^O - Z_{R_1}{}^O}{Z_{R_2}{}^O + Z_{R_1}{}^O} \right) \right] \right] \right], \tag{53}$$

as can readily be seen by comparing these expressions with the equations, $p_2 = p_1 A - (V_1/S_1)Z_0 B$; $V_2 = V_1 C - (p_1/Z_0)S_1 D$. In the above expression the letter I indicates that the impedances Z_{R_1} and Z_{R_2} are to be taken at the input junction, while the letter O indicates that they are to be taken at the output junction.

The effect of this type of tapering is to change the propagation constant scarcely at all, but to lower the characteristic impedances in the neighborhood of the cut-off frequencies. This tends to produce large reflection losses and hence effectively the band is narrowed. A transforming action equivalent to a transformer of turns ratio $\sqrt{S_1/S_2}$ occurs as before.

To obtain the equation for a straight horn, we let S_1, a typical area of the horn, equal

$$S_1 = n^2 K = K'(nl)^2 = K'(x_1)^2,$$

where $nl = x_1$, the distance from the apex of the horn, and l the length of an individual section. Γ becomes $i\omega l/C$, and Z_{R_1} and Z_{R_2} are

$$Z_{R_1} = \frac{\sqrt{p_0\gamma\rho}\, in\, \frac{\omega}{C} l}{1 + in\frac{\omega}{C} l} = \frac{\sqrt{p_0\gamma\rho}\, i\, \frac{\omega}{C} x_1}{1 + i\frac{\omega}{C} x_1},$$

and (54)

$$Z_{R_2} = \frac{\sqrt{p_0\gamma\rho}\, i\, \frac{\omega}{C} x_1}{-1 + i\frac{\omega}{C} x_1}.$$

Substituting these values in equation (53), we obtain the equation

$$p_2 = \frac{x_1}{x_2}\left\{ p_1 \left[\cos\frac{\omega}{C}(x_2 - x_1) + \frac{\sin\left(\frac{\omega}{C}(x_2 - x_1)\right)}{\frac{\omega}{C}x_1} \right] - i\frac{V_1}{S_1}\sqrt{P_0\gamma\rho}\sin\frac{\omega}{C}(x_2 - x_1) \right\},$$

$$V_2 = \frac{x_2}{x_1}\left\{ V_1\left(\cos\frac{\omega}{C}(x_2 - x_1) - \frac{\sin\left(\frac{\omega}{C}(x_2 - x_1)\right)}{\frac{\omega}{C}x_2} \right) - i\frac{p_1 S_1}{\sqrt{P_0\gamma\rho}} \times \left[\left(1 + \frac{1}{\left(\frac{\omega}{C}\right)^2 x_1 x_2} \right)\sin\frac{\omega}{C}(x_2 - x_1) + \left(\frac{1}{\frac{\omega}{C}x_2} - \frac{1}{\frac{\omega}{C}x_1} \right)\cos\frac{\omega}{C}(x_2 - x_1) \right] \right\}. \quad (55)$$

If we introduce two lengths ϵ_1 and ϵ_2 defined by $\tan(\omega/C)\epsilon_1 = (\omega/C)x_1$ and $\tan(\omega/C)\epsilon_2 = (\omega/C)x_2$ and take account of the fact that the impedance as defined here must be multiplied by $i\omega$ to correspond to the impedance defined by Webster, then it is evident that the above equation corresponds to the relation given by Webster.[9, 10]

It is interesting to compare the relations obtained above involving the assumptions introduced in Section II with the solution involving no assumptions. This can be done for the conical horn, since its solution can be obtained using spherical waves. In Section III–*D*, the impedance looking into a conical horn was obtained when an infinite impedance terminated the horn. If we set $V_2 = 0$ in the last of equations (55) and solve for the ratio of p_1/V_1, it is evident that the impedance agrees with that given in Section III–*D*. Hence it is evident that both methods give the same solution.

Many other types of tapered filters can be solved in a similar manner, but no more will be considered here.

V. General Network Equations and Network Parameters

We can combine a number of symmetrical structures to form a general network. For any symmetrical structure we can write the

[10] The solution for the conical horn has been discussed in more detail by I. B. Crandall, "Theory of Vibrating Systems and Sound," D. Van Nostrand, 1926, p. 152.

equations

$$p_2 = p_1 \cosh \Gamma_1 - V_1 \frac{Z_{0_1}}{S_1} \sinh \Gamma_1,$$

$$V_2 = V_1 \cosh \Gamma_1 - \frac{p_1 S_1}{Z_{0_1}} \sinh \Gamma_1.$$

Suppose then that we wish to join this structure to other structures, with different characteristics and with different area conducting tubes. At the junction of the structures, we have by equation (14)

$$p_3 = p_2, \qquad V_3 = V_2.$$

Combining these with the above, we have

$$p_3 = p_1 \cosh \Gamma_1 - V_1 \frac{Z_{0_1}}{S_1} \sinh \Gamma_1,$$

$$V_3 = V_1 \cosh \Gamma_1 - \frac{p_1 S_1}{Z_{0_1}} \sinh \Gamma_1.$$

Writing a set of equations similar to the above for the second structure and combining, we have

$$\left.\begin{aligned} p_4 = {} & p_1 \left(\cosh \Gamma_1 \cosh \Gamma_2 + \frac{S_1}{S_2} \frac{Z_{0_2}}{Z_{0_1}} \sinh \Gamma_1 \sinh \Gamma_2 \right) \\ & - V_1 \frac{Z_{0_1}}{S_1} \left(\sinh \Gamma_1 \cosh \Gamma_2 + \frac{Z_{0_2}}{Z_{0_1}} \frac{S_1}{S_2} \cosh \Gamma_1 \cosh \Gamma_2 \right), \\ V_4 = {} & V_1 \left(\cosh \Gamma_1 \cosh \Gamma_2 + \frac{Z_{0_1}}{Z_{0_2}} \frac{S_2}{S_1} \sinh \Gamma_1 \sinh \Gamma_2 \right) \\ & - \frac{p_1 S_1}{Z_{0_1}} \left(\sinh \Gamma_1 \cosh \Gamma_2 + \frac{S_2}{S_1} \frac{Z_{0_1}}{Z_{0_2}} \cosh \Gamma_1 \sinh \Gamma_2 \right). \end{aligned}\right\}$$

We can also write this in the form

$$p_4 = p_1 \begin{vmatrix} \cosh \Gamma_1 & \frac{S_1}{S_2} \frac{Z_{0_2}}{Z_{0_1}} \sinh \Gamma_2 \\ -\sinh \Gamma_1 & \cosh \Gamma_2 \end{vmatrix} - V_1 \frac{Z_{0_1}}{S_1} \begin{vmatrix} \sinh \Gamma_1 & \frac{S_1}{S_2} \frac{Z_{0_2}}{Z_{0_1}} \sinh \Gamma_2 \\ -\cosh \Gamma_1 & \cosh \Gamma_2 \end{vmatrix},$$

$$V_4 = V_1 \begin{vmatrix} \cosh \Gamma_1 & \frac{Z_{0_1}}{Z_{0_2}} \frac{S_2}{S_1} \cosh \Gamma_2 \\ -\sinh \Gamma_1 & \cosh \Gamma_2 \end{vmatrix} - \frac{p_1 S_1}{Z_{0_1}} \begin{vmatrix} \sinh \Gamma_1 & \frac{Z_{0_1}}{Z_{0_2}} \frac{S_2}{S_1} \sinh \Gamma_2 \\ -\cosh \Gamma_1 & \cosh \Gamma_2 \end{vmatrix}.$$

In fact if we combine η structures of this kind, we can write the equations

$$\left.\begin{aligned} p_\eta &= p_1 A - V_1 \frac{Z_{0_1}}{S_1} B, \\ V_\eta &= V_1 C - \frac{p_1 S_1}{Z_{0_1}} D, \end{aligned}\right\} \tag{56}$$

where

$$A = \begin{vmatrix} \cosh \Gamma_1 & \frac{S_1}{S_2}\frac{Z_{0_2}}{Z_{0_1}} \sinh \Gamma_2 & -\frac{S_1}{S_3}\frac{Z_{0_3}}{Z_{0_1}} \sinh \Gamma_3 \cdots \\ -\sinh \Gamma_1 & \cosh \Gamma_2 & \frac{S_2}{S_3}\frac{Z_{0_3}}{Z_{0_2}} \sinh \Gamma_3 \cdots \\ \sinh \Gamma_1 & -\sinh \Gamma_2 & \cosh \Gamma_3 \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cosh \Gamma_\eta \end{vmatrix} \tag{57}$$

$$B = \begin{vmatrix} \sinh \Gamma_1 & \frac{S_1}{S_2}\frac{Z_{0_2}}{Z_{0_1}} \sinh \Gamma_2 & -\frac{S_1}{S_3}\frac{Z_{0_3}}{Z_{0_1}} \sinh \Gamma_3 \cdots \\ -\cosh \Gamma_1 & \cosh \Gamma_2 & \frac{S_2}{S_3}\frac{Z_{0_3}}{Z_{0_2}} \sinh \Gamma_3 \cdots \\ \cosh \Gamma_1 & -\sinh \Gamma_2 & \cosh \Gamma_3 \quad \cdot \quad \cdot \\ -\cosh \Gamma_1 & \sinh \Gamma_2 & \sinh \Gamma_3 \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cosh \Gamma_\eta \end{vmatrix} \tag{58}$$

$$C = \begin{vmatrix} \cosh \Gamma_1 & \frac{S_2}{S_1}\frac{Z_{0_1}}{Z_{0_2}} \sinh \Gamma_2 & -\frac{S_3}{S_1}\frac{Z_{0_1}}{Z_{0_3}} \sinh \Gamma_3 \cdots \\ -\sinh \Gamma_1 & \cosh \Gamma_2 & \frac{S_3}{S_2}\frac{Z_{0_2}}{Z_{0_3}} \sinh \Gamma_3 \cdots \\ +\sinh \Gamma_1 & -\sinh \Gamma_2 & \cosh \Gamma_3 \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cosh \Gamma_\eta \end{vmatrix} \tag{59}$$

$$D = \begin{vmatrix} \sinh \Gamma_1 & \frac{S_2}{S_1}\frac{Z_{0_1}}{Z_{0_2}} \sinh \Gamma_2 & -\frac{S_3}{S_1}\frac{Z_{0_1}}{Z_{0_3}} \sinh \Gamma_3 \cdots \\ -\cosh \Gamma_1 & \cosh \Gamma_2 & \frac{S_3}{S_2}\frac{Z_{0_2}}{Z_{0_3}} \sinh \Gamma_3 \cdots \\ \cosh \Gamma_1 & -\sinh \Gamma_2 & \cosh \Gamma_3 \quad \cdot \quad \cdot \\ -\cosh \Gamma_1 & \sinh \Gamma_2 & -\sinh \Gamma_3 \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cosh \Gamma_\eta \end{vmatrix} \tag{60}$$

Among these four determinants there is one relation

$$AC - BD = 1, \tag{61}$$

as can be seen by multiplying them together.

Hence to completely specify the characteristics of the structure three parameters are required. A number of possible sets of parameters exist whose usefulness depends on the type of structure to which they are applied. The set of parameters having the greatest use in

connection with electrical networks are the image parameters which include two image impedances and an image transfer constant. We define these constants as follows for the acoustic case.

If we have a network terminated in impedances Z_{I_1} and Z_{I_2} (per square centimeter of area) at the beginning and at the end of the network, then these impedances are the image impedances of the structure if they terminate the structure in such a way that at either termination junction, the impedance looking in either direction is the same.

The image transfer constant θ may be defined as one half the natural logarithm of the vector ratio of the product of the pressure by the volume velocity, at the input junction point, and this product for the output junction point, when the network is terminated in its image impedances.

Hence

$$\theta = \frac{1}{2} \log_e \frac{p_1 V_1}{p_\eta V_\eta}.$$

To determine the image impedances, we have one set of equations

$$\left.\begin{aligned} p_\eta &= p_1 A - V_1 \frac{Z_{0_1}}{S_1} B, \\ V_\eta &= V_1 C - \frac{p_1 S_1}{Z_{0_1}} D. \end{aligned}\right\} \tag{62}$$

This gives the pressure and volume velocity propagated in one direction. We need also the equation of propagation in the opposite direction. This can evidently be written

$$\left.\begin{aligned} p_\eta' &= p_1' A' - V_1' \frac{Z_{0_\eta}}{S_\eta} B', \\ V_\eta' &= V_1' C' - \frac{p_1' S_\eta}{Z_{0_\eta}} D', \end{aligned}\right\} \tag{63}$$

where p_η' and V_η' represent the pressure and volume velocity at the beginning and V_1' and p_1' at the end of the structure. A' can be obtained from A by cyclically permuting the subscripts. By writing the expansions for these quantities we can show that

$$A' = C; \quad C' = A; \quad B' = \frac{Z_{0_1}}{Z_{0_\eta}} \frac{S_\eta}{S_1} B; \quad D' = \frac{Z_{0_\eta}}{Z_{0_1}} \frac{S_1}{S_\eta} D. \tag{64}$$

Eliminating the ratio V_η / V_1 from (62) and writing $p_1 / V_1 = Z_{I_1} / S_1$ and $p_\eta / V_\eta = Z_{I_2} / S_\eta$, we obtain

$$Z_{I_1} Z_{I_2} D + Z_{0_1} \left[Z_{I_1} \left(\frac{S_\eta}{S_1} A \right) - Z_{I_2} C \right] - Z_{0_1}^2 \left(\frac{S_\eta}{S_1} B \right) = 0. \tag{65}$$

From (63) eliminating the ratio V_η'/V_1' and writing $p_\eta'/V_\eta' = Z_{I_1}/S_1$ and $p_1'/V_1' = Z_{I_2}/S_\eta$ and substituting the values in (64), we have

$$Z_{I_1}Z_{I_2}D + Z_{0_1}\left(Z_{I_2}C - Z_{I_1}\frac{S_\eta}{S_1}A\right) - Z_{0_1}^2\frac{S_\eta}{S_1}B = 0. \tag{66}$$

Solving (65) and (66) simultaneously, we find

$$Z_{I_1} = Z_{0_1}\sqrt{\frac{BC}{AD}}; \quad Z_{I_2} = Z_{0_1}\frac{S_\eta}{S_1}\sqrt{\frac{AB}{CD}}. \tag{67}$$

From the definition of θ and equations (62), we can show that

$$\cosh\theta = \sqrt{AC}. \tag{68}$$

In terms of these parameters, the effect upon the pressure or volume velocity in the termination of an acoustic system, due to inserting the structure into the system, will be given by multiplying the terminal pressure or volume velocity by the factor

$$\frac{\frac{Z_A}{S_1}+\frac{Z_B}{S_\eta}}{\frac{2Z_B}{S_\eta}} \times \frac{2\sqrt{\frac{S_1}{S_n}}\sqrt{Z_{I_1}Z_{I_2}}}{Z_{I_1}+Z_A} \times \frac{2Z_B}{Z_{I_2}+Z_B} \times e^{-\theta} \times \frac{1}{1-\frac{Z_{I_2}-Z_B}{Z_{I_2}+Z_B}\times\frac{Z_{I_1}-Z_A}{Z_{I_1}+Z_A}\times e^{-2\theta}}, \tag{69}$$

where Z_A and Z_B are respectively the impedances, per square centimeter, of the acoustic system at the insertion junction looking towards and away from the source.

Appendix I. Proof of Thévenin's Theorem for an Acoustic System

The proof of Thévenin's theorem as stated in Section III can be obtained directly from the general network equations given in Section V. These equations are

$$p_2 = p_1A - V_1\frac{Z_0}{S_1}B,$$

$$V_2 = V_1C - \frac{p_1S_1}{Z_0}D,$$

where $AC - BD = 1$. If we connect at the input end a source of

pressure p_0, whose specific internal impedance is Z_T, we can write

$$p_0 = p_1 + V_1 \frac{Z_T}{S_1} .$$

Inserting this result in the above equation, we obtain

$$\left.\begin{aligned} p_2 &= p_0 A - \frac{V_1}{S_1}(Z_T A + Z_0 B), \\ V_2 &= V_1\left(C + \frac{Z_T}{Z_0} D\right) - \frac{p_0 S_1}{Z_0} D. \end{aligned}\right\} \tag{70}$$

To obtain the pressure when an infinite impedance is used at the termination, we let $V_2 = 0$, and solving for V_1 we have

$$V_1 = \frac{p_0 D S_1}{Z_0 C + Z_T D} . \tag{71}$$

Substituting this in the first of equations (70), we have

$$p_2 = \frac{p_0 Z_0}{(Z_0 C + Z_T D)} = p_0', \tag{72}$$

which is the terminal pressure for an infinite terminating impedance.

Eliminating V_1 from (70) and substituting $V_2 Z_R / S_\eta = p_2$, we have

$$V_2 = \frac{p_0 Z_0}{(Z_0 C + Z_T D)} \times \frac{1}{\frac{Z_R}{S_\eta} + \frac{Z_0}{S_1}\left(\frac{Z_T A + Z_0 B}{Z_0 C + Z_T D}\right)} . \tag{73}$$

We can show now that

$$\frac{Z_0}{S_1}\left(\frac{Z_T A + Z_0 B}{Z_0 C + Z_T D}\right) = \frac{Z_T'}{S_1},$$

which is the impedance at the terminating junction looking toward the source, when the specific impedance Z_T terminates the input end. From equations (63) and (64), we can write

$$p_2' = p_1' C - V_1' \frac{Z_{0_1}}{S_1} B,$$

$$V_2' = V_1' A - \frac{p_1' S_1}{Z_0} D.$$

Substituting $V_2'(Z_T/S_1) = p_2'$ and solving for the ratio p_1'/V_1', we have

$$\frac{p_1'}{V_1'} = \frac{Z_T'}{S_1} = \frac{Z_0}{S_1}\left(\frac{Z_T A + Z_0 B}{Z_0 C + Z_T D}\right). \tag{74}$$

Hence we can express V_2 in equation (73) as

$$V_2 = \frac{p_0'}{\frac{Z_R}{S_\eta} + \frac{Z_T}{S_1}}$$

which is Thévenin's theorem.

Appendix II. Determination of Loss for a Constant Volume Velocity Source

Another type of insertion effect desired in some cases is the effect caused by inserting filter structures in an acoustic system in which the source supplies a constant volume velocity. One such acoustic system is the phonograph.

In order to obtain this effect we first prove the theorem: If a source of constant volume velocity V_1 is connected to the input of an acoustic system, and if the impedance Z_R (per square centimeter) is used to terminate the system, the volume velocity V_2 will be $p_0''/[(Z_R/S_1) + (Z_c/S_\eta)]$ where p_0'' is the pressure at the termination of the system when the system is closed through an infinite impedance, and Z_c is the specific impedance of the acoustic system at the output junction looking toward the source when the system is terminated in in an infinite impedance at the input junction. S_1 and S_η are the areas at the input and output junctions, respectively.

To prove this we substitute the value of p_0 given by (71) in the first of equations (70) and obtain for the pressure, with an infinite impedance termination

$$p_2'' = \frac{V_1 Z_0}{S_1 D}. \tag{75}$$

Then eliminating p_0 from equations (70), and inserting the value $p_2 = V_2 Z_R / S_\eta$, we obtain

$$V_2 = \frac{V_1 Z_0}{S_1 D}\left[\frac{1}{\frac{Z_R}{S_\eta} + \frac{A Z_0}{S_1 D}}\right]. \tag{76}$$

From equation (74) we see that the impedance looking toward the source is $(Z_0 A/S_1 D)$ if we make Z_T approach infinity. Hence

$$V_2 = p_0''\left[\frac{1}{\frac{Z_R}{S_\eta} + \frac{Z_c}{S_1}}\right].$$

To obtain the insertion loss for a constant current source, then, it is only necessary to substitute Z_c for Z_a in equation (30). One special case of interest is the case where the acoustic filter is connected directly to the source. In this case $Z_c = \infty$ and the insertion effect is determined by the factor

$$\left(\frac{2Z_0}{Z_0 + Z_b}\right) \times e^{-\Gamma} \times \left(\frac{1}{1 + \frac{Z_0 - Z_b}{Z_0 + Z_b} e^{-2\Gamma}}\right). \tag{77}$$

Contemporary Advances in Physics—XIII. Ferromagnetism

By KARL K. DARROW

Magnetism was revealed to Europeans by pieces of a mineral later to be called lodestone, which lay scattered in the fields of Magnesia in Asia Minor, and were endowed with the curious power of attracting iron. They who first noticed it were apparently Greeks of the period before the practice of writing; for legends of the discovery were transmitted by the Greeks of later centuries, legends entangled with tales of Cretan shepherds and the myth of Medea. Electricity was disclosed, evidently in the same dim period and region, by fragments of amber on which friction conferred the remarkable power of attracting shreds and flakes of light materials.

By these quaint phenomena electricity and magnetism were disclosed to the European world before the beginnings of written history; and the intimations were recorded in writings of classical antiquity, and handed down from generation to generation. Yet two millennia and more were destined to flow past, before sufficiently many further data should be gathered to make possible the forming of a valid conception of either. The nineteenth century arrived, before anyone detected the signs that the two are but different aspects of one fundamental entity. Obviously the early hints were not sufficient; but it would not be well to conclude that therefore the Greeks were unwise. If they are indicted for stupidity because they did not understand the lodestone and the electrified amber, the indictment lies also against ourselves. For these are instances of ferromagnetism and of frictional electricity; which is to say, they belong to provinces which to this day are not fully incorporated into the empire of the theory of electricity and magnetism.

How then does it happen that the phenomena earliest discovered must still be listed among the least well understood? There is nothing unusual in this. There is no general reason for expecting that the phenomena which occur spontaneously and frequently and conspicuously in Nature should be the easiest to understand. On the contrary, it frequently happens that they are much less instructive and interpretable than others which can be brought to pass only by careful choice of conditions and skilful experimentation. The history of physics abounds in instances of such contrasts, and there is none more striking than the one with which I am to deal. Many phenomena of

magnetism are well explained by the contemporary theory, many seem admirably clear; but none of these was or could have been witnessed by the Greeks. We know much about the magnetic properties of gases, dilute solutions, free atoms, elements and compounds which are so feebly magnetizable that before 1830 they were not supposed to be "magnetic" at all; we are still perplexed by the behaviour of iron and lodestone. This is the reason why there are textbooks of magnetism, in which hundreds of pages are devoted to the data and the theories of a number of effects most difficult to perceive and known to none but physicists, while the magnets of daily experience are dismissed with a chapter or two of mere description. As for electrified amber and its kindred, they are fortunate to have a few paragraphs of any modern treatise on electricity bestowed upon them.

Frictional electricity is not a very striking phenomenon, nor is it valuable in engineering; consequently it has been allowed to slip into obscurity, shunned by cautious students on the hunt for problems promising immediate returns. Ferromagnetism is not so unobtrusive. Much of the electric machinery which has transformed the world since Napoleon derives all its efficacy from certain blocks of iron or magnetizable alloy, enmeshed among the wires. So useful a property of matter does not consent to lie neglected; physicists are forced to hearken to its insistent demands for attention. Ambition to achieve some technical advance supplies a strong incentive; and there is a feeling of humiliation that a quality of matter so conspicuous and so remarkable, and so remarkably limited to a particular class of substances not in other ways exceptional, should not be properly connected with the structure of contemporary physics. For these and other motives, there are always physicists engaged in the struggle with the problem of ferromagnetism—no mean struggle, for the difficulties are truly serious. It was a tough problem which was offered to the Greeks and which they rejected, when they saw the lodestone, took note of it, and left it for the modern world to study.

Some of the difficulties of ferromagnetism may be peculiar to it. Others, it is to be feared, are examples of the troubles which are reserved for scientists by the internal properties of solid bodies generally, and which physicists will some day be forced to confront when the obvious problems of gases and free atoms are exhausted, if they are not sooner incited by curiosity or by the requirements of engineering. Most of the great conquests of recent physics have been achieved through the study of gases, or of those properties of matter which are the same for the solid as for the gaseous state. It is but natural to

wish to postpone as long as possible the attack upon the intrinsic properties of solids; but there is no evading it in the study of ferromagnetism, for this is a property of solids only, and not even of transparent solids at that. One would wish at least to be permitted to restrict the study to pure elements or simple compounds; but many of the most interesting of the ferromagnetics belong among those bewildering substances the alloys, which form what the mathematicians describe as a *continuum* beside and among the great yet finite number of chemical compounds. If one were to work only with perfectly pure iron (supposing that one could get such a substance, or could recognize it when he had it!) the problem would not yet be simple; for every species of mechanical and thermal treatment, and magnetization itself, would transform the iron into a new sort of solid.

These difficulties I will stress in the pages of this article. There is another. The information about a ferromagnetic substance—the prime material required for theorizing or for practical applications—is usually furnished in the form of so-called *I*-vs-*II* curves; that is to say, relations between the "intensity of magnetization" and the "magnetizing field." These curves play the part of the ultimate data of experience. Yet they are not ultimate data; the "magnetizing field" is seldom actually measured, the "intensity of magnetization" almost never. These entities *I* and *II* are deduced from experience by means of a theory. The theory is indispensable. If an uninstructed person were presented with a number of variously-shaped pieces of iron, and a battery and a coil of wire with which to produce any desired magnetic field, and any number of measuring-instruments, he would find it extremely hard to select something to measure that might yield a coherent and intelligible set of data. He would be able to show in a vague way that the greater the magnetic field acting upon any piece of iron, the more powerful a magnet it becomes; but if he were to search for some precise measurable quantity that could serve as a measure of the power of the magnet, and that would be *characteristic of iron as a substance and not merely characteristic of individual pieces of iron as individuals*, his search would be a long one. From what I have just called "the theory" he would find out what to measure, and how to calculate from it the value of something characteristic of iron and not affected by the shape of the pieces; he would find out how to trace an "*I*-vs-*II*" curve. This curve would serve in turn as a basis for theories of ferromagnetism; but theory would have entered already into the preparation of the curve. I shall therefore devote the first section of this article to the principles according to which such curves are determined from the immediate data. Any

reader who feels that these principles are familiar, or self-evident, or unimportant, may leap to the second section and the third, in which the *I*-vs-*II* curves are accepted as the data of experience.

I shall not venture a definition of ferromagnetism until nearly the end of the article. Such a definition is not easy to make, unless one takes refuge in the statement that "ferromagnetism is the kind of magnetism displayed by iron." I can only regret the frequency with which such ponderous words as *ferromagnetism* and *permeability* and *susceptibility* and *magnetization* and *magnetostriction* must needs appear. The subject is encumbered by its heavy vocabulary; it ought to have a new one made up entirely of short and vivid words.

A. Analysis of the Magnetization of Magnetized Bodies.

Let us imagine a collection of magnets such as one frequently sees, horseshoe magnets for example, with their ends painted red and blue. We know that (if the painting was done properly) the red end of each attracts the blue ends and repels the red ends of the others; the blue end attracts those which the red end repels and repels those which the red end attracts. It seems as if the ends of the magnets were covered with invisible substances—one kind on all the red ends, the other on all the blue ends—so constituted, that a sample of either substance attracts all samples of the other sort, repels all samples like itself. Coulomb found that if the magnets were long and slender, so that the power of attracting and repelling was concentrated very closely about the extremities of each, these extremities attracted or repelled one another according to an inverse-square law. That suggested gravitation and electric force; which suggested in turn that, even as matter is the source of gravitation and electric charge is the source of electric force, so also there is an invisible thing called magnetism which inhabits iron—or rather, two invisible things, positive magnetism and negative magnetism, which may be pulled and pushed around inside and over the surface of a piece of iron. This notion of a pair of invisible and mobile fluids is very helpful, and I shall use it in several passages; yet the reader must not take it as corresponding to the actual reality. We cannot imagine two or even one perfectly mobile magnetic fluid, for a well-known reason.

The reason is, that even though a magnet may appear to carry nothing but positive magnetism on one of its ends and nothing but negative magnetism on the other, yet it is not possible to cut off anywhere a piece containing only one of these kinds. In fact it is not possible anywhere to cut off a piece not containing equal quantities of the two kinds of magnetism. Any piece of matter always contains as much

positive magnetism as negative; so also does any smaller fragment broken off from the piece, and any still smaller bit broken out of the fragment, and so forth until the original piece is crumbled into dust, each particle of which still contains as much magnetism of either sign as of the other.*

Now this requires that when we subdivide a magnetized piece of iron into tiny parcels or volume-elements, not by the hammer nor the file but by the exercise of the imagination, these volume-elements must themselves be imagined as magnets each invested with a positive pole and a negative pole and a magnetic axis pointing in some particular direction. I am not implying atoms by these "parcels"—we shall as yet have nothing to do with atoms. The process of dividing a substance into imaginary small volume-elements has nothing in common with the construction of atoms or atom-models; quite the contrary! It is a process which every physicist undertakes, whenever he desires to analyze the flow of water or the vibrations of air or the strain of a twisted rod or any of a multitude of problems concerning pieces of matter, which, whatever his views about atoms, he intends to regard as continuous media for the nonce. Well! in dealing with magnetism, it is not sufficient to conceive these volume-elements as cubical or otherwise-shaped bits of matter entirely uniform and isotropic in their qualities; they must be conceived as being little magnets themselves.

This is the reason why we are taught to imagine a piece of magnetized iron as a collection of tiny cubes, each bearing positive magnetism spread like a coat of paint over one side, and negative magnetism over the side opposite; or as a bundle of filaments which, where they come out to the surface of the piece, divide it into a pattern of area-elements each of which is overspread with magnetism positive or negative; or as a pile of laminæ, somewhat like a nest of saucers, each of which is covered with magnetism of the two signs on its two sides. This is the reason why, developing the first of these conceptions (which contains implicitly the other two), we are taught to picture a function called the *intensity of magnetization*, which has a definite value at each point within the magnet, and may be visualized with the aid of the imaginary cubes. Select a point in the interior of the magnet, and imagine it surrounded by a cubical volume-element of thickness d and face-area d^2 and volume d^3; and imagine two opposite sides of the cube to be covered with magnetism of opposite signs painted on with a surface-density I, so that each side bears a quantity Q which

* The best evidence for this statement is the fact that magnets in a uniform magnetic field such as that of the earth experience no force tending to displace them bodily though they experience a torque tending to orient them.

is Id^2. This cube would be a minute magnet having the moment* Qd which is Id^3, directed normally to the two sides coated with magnetism; for I is a quantity possessing both magnitude and direction, a vector quantity and not a scalar—this is a way of expressing the complexity to which an allusion was made in the last paragraph. The piece of magnetized material is to be visualized as the assembly of all these little cubes, each having a magnetic moment equal to its volume multiplied into the value of I prevailing in it. The force exerted by the piece anywhere outside of its volume is to be considered as the sum of the forces there exerted by all the little magnets. The entity I plays the rôle of a magnetic-moment-per-unit-volume. It is this entity which is defined as the *intensity of magnetization* of the material.

This, it may be objected, is something quite unverifiable; for one cannot penetrate into the interior of a piece of iron, and find out whether it contains such an entity as this vector I. Quite so! and this is another of the great difficulties in ferromagnetism, though not peculiar to ferromagnetism alone, for it besets in greater or less degree every problem of the properties of solid bodies. The state of affairs within a piece of magnetized iron is the leading problem of ferromagnetism, indeed it is the one problem which contains all the rest. But there is no way of ascertaining that state of affairs, for there is no way of putting a measuring-instrument into a piece of iron. One might scoop a hole in the iron to make a place for the magnetometer, but then the magnetometer would be in the hole and not in the iron. The field of magnetic force outside the magnet can be plotted, the lines of force in the field can be followed up to the very edge of the magnetized material, but there they dive and they disappear. When one sees a sketch of a magnet and its environment, in which the lines of force coming up from all sides to the surface of the magnet are connected in pairs by "lines of induction" passing through the body of the magnet, he should realize that while the lines of force outside are a map of a field which can be explored, the lines of induction within are hypothetical altogether.

Why then take the trouble of conceiving entities such as these, intensity of magnetization I and induction B, since they are solely imagined to exist in a locality where there is no possible means of penetrating to seek them? The reason is this, and this only: Confined though they are within the bodies of the magnets, they facilitate the

* "Magnetic moment" is usually defined by inviting the reader to imagine a magnet so long and slender that its "magnetism" is concentrated almost completely at its ends or "poles"; the moment of such a magnet is the product of its length into the amount of magnetism, or "polestrength," at either end. Actual magnets have no true poles. The moment of an actual magnet is the torque which a unit field exerts upon it when it is normal to the direction in which the field tends to set it.

understanding of the effects which the magnets produce outside. Induction and intensity of magnetization are things which are supposed to exist *inside* a solid magnetic body, to make it possible to predict what effects that body produces in the world outside of itself—the only region which can be entered with or without measuring-instruments.

Now if a magnet were delivered over by Nature in fixed and permanent state, so that nothing which could be done to it would alter its behavior towards surrounding objects, the problem of determining I would be relatively simple. It would amount to this: to build up a structure of little cubical magnets occupying the same volume as the actual magnet, and producing everywhere outside that volume the same field as the actual magnet is observed to produce. In other words, it would consist in seeking a function I of the coordinates x, y, z of the points within the volume of the magnet, fulfilling the following condition: when this volume is subdivided into small cells of volume dv, and each is treated as a magnet of moment Idv, and the forces exerted by all these little magnets at any point outside of the volume are summed together, their sum shall turn out to be the same as the force which the actual magnet is observed to exert at that point.

This however is not the whole of the actual problem. The force which a magnet exerts at any particular point in its vicinity depends upon the magnetic fields which are impressed upon it by external objects—other magnets, or electric currents, or the earth itself. It becomes a different magnet when it is subjected to a different field. The process of finding a function I fulfilling the condition made above must therefore be carried through anew whenever the exterior fields acting upon the magnet are changed.

This variability makes the problem much more difficult. Yet in some cases it can be dealt with, in the same manner as the more restricted problem of analyzing an unchanging magnet into volume-elements; and in dealing with it, the first foundations of a theory of magnetism are laid down.

A piece of iron is observed to become a different magnet, whenever the impressed magnetic field is changed. Very well! we will try to describe the difference, by assuming that each of the volume-elements into which we have mentally divided the piece becomes itself a different magnet. The change in the magnetism of the piece is all too likely to be complicated and obscure; but we will simplify by supposing that the magnetization of each of the volume-elements depends upon the magnetic field prevailing in it, according to some law which is

the same for all the volume-elements; that there is a fixed relation between the intensity of magnetization at a point and the field existing at that point, which is the same everywhere within the supposedly uniform piece of iron, which is a quality of that particular kind of iron. If there is no such relation, the whole procedure is likely to be futile. If there is such a relation, it is the fundamental fact of magnetism; and the first business of the student of magnetism is to determine it for as many substances, under as many conditions, as he can. We shall presently see that most research in ferromagnetism is devoted to determining this relation, by methods which would not yield self-consistent results, did it not exist.

But we shall attain nothing by merely assuming that there is such a relation, unless we make another assumption concerning the field prevailing within the magnet; for it is quite inaccessible, we cannot enter in to measure it. Let us therefore suppose that the field produced at any point inside the magnet, by the objects outside—be they laboratory magnets, or electric currents, or the earth itself—is the same as they would produce at that point were the magnet taken away, leaving them the same. The outer parts of the magnet are supposed *not* to shield the inner parts from the magnetic influences of the outer world. This is a natural corollary of the supposition we have tacitly made already, that the outer volume-elements of the magnet do not shield the outer world from the magnetic forces due to the inner volume-elements. We assume it; and we assume that the intensity of magnetization and the magnetic field, the vectors I and H, are parallel to one another,* and that there is a relation between their magnitudes which is the same for every point within the magnet.

On proceeding to test this set of assumptions by the appeal to experiment, we encounter results which at first sight seem to destroy them. For instance, let us immerse a short rod of iron (quite demagnetized to begin with) in the uniform magnetic field produced within a long cylindrical tube by an electric current flowing through a coil of wire, a solenoid, evenly wrapped around the tube. The field H_e which the current would produce within the tube were the iron not there is uniform in magnitude and direction, everywhere parallel to the axis of the solenoid. By the last assumption, this is the field which the current produces everywhere inside the iron. We map the magnetic field produced by the rod in its vicinity, and determine

* There are cases, neither few nor trivial, in which I and H cannot always be supposed parallel; for instance, when the magnet is a large crystal, or when it is a plate of metal which has been cold-rolled, or when the direction of the magnetizing field is changed after the substance is already perceptibly magnetized. But if I were to expound the most general actual case, this article would never come to an end.

the function I which describes the magnetization which would produce such a field. The vector I is not uniform throughout the iron, either in direction or in magnitude. Though H_e is the same everywhere within the metal, I varies from point to point. This result by itself seems to demolish the assumptions.

The contradiction however is only apparent; it vanishes if we make due allowance for the field produced at every part of the magnet by the other parts, for *the effect of the magnet upon itself.* Continuing to use the illustration of the short rod in the uniform impressed field: the distribution of elementary magnets which the function I expresses, and which produces at every point outside the iron a calculable field agreeing with the field there observed, should also produce a calculable field at every point within the iron. Considering that we have assumed that the force due to even the innermost volume-element of the magnet is exerted unimpeded everywhere in the outside world, we cannot consistently avoid assuming that its force is exerted unimpeded upon the other volume-elements as well. Thus it is reasonable to suppose that if the value of I at any point in the iron is controlled by the magnetic field there prevailing, then the truly controlling field comprises not only the one (H_e) due to the external agencies, but also the other (H_i) due to the multitude of little magnets presumed to constitute the piece of iron. The value of I should depend on the resultant H of H_e and H_i. In the present case of the short rod inside the solenoid, the vector H_e is uniform, but the vector H_i varies from point to point, and consequently so does the resultant H of H_e and H_i, and consequently so does I. More properly, I should not use such a word as "consequently" at all; both I and H vary from point to point, either accounting for the other, either being cause and either being effect.

This, by the way, is one of the reasons why as a rule it is not possible to analyze the magnetization of a magnet by cutting it into little pieces and measuring the moment of each separately. When such a piece is isolated from the rest of the magnet, the field acting upon it is changed even though all the external field-producing agencies remain the same. The other reason for not cutting up a magnet is, that the stresses exerted on the material in the process of cutting are likely to change it into some very different ferromagnetic material—but of this, more later.

The problem of determining I now assumes its full scope. For every magnet, or let us say for every piece of magnetized iron, there should be a function I describing its magnetization, defined at every point within it and satisfying these conditions:

First, it should account for the field due to the magnet at every point outside;

Second, its value at every point inside the magnet should be a definite function of the thing which we have just tentatively defined as "the magnetic field" at that point; *viz.* the resultant of that field which the external agencies would produce were the magnet away, and that which the magnetization I should itself produce.

Or, in other words: it should be possible to build up a reproduction of the magnetized piece of iron out of little magnets, the magnetic moment of each depending in a perfectly definite way on the force exerted on it by the other little magnets and by the external world, and all together producing the same effects in the external world as the piece of iron does.

In saying "it should be possible" I do not mean to imply that there is an obligation resting upon Nature to construct magnetizable objects in such a way that it is possible. One could not prove *a priori* that she does. One must take variously shaped pieces of magnetizable metal and observe their behavior in various impressed fields, and ascertain for each whether or not there is a function I. In so doing, one is liable to encounter very great mathematical difficulties. In fact, the difficulties are likely to prove insuperable unless the piece of metal is shaped in one or other of a few definite ways, and the impressed field is uniform and properly oriented.

Let us attack the problem from the other side, and enquire first whether it is possible so to shape a piece of iron and so to orient the impressed field, that the extra field due to the magnetization should vanish everywhere within the iron, and the actual field should everywhere be identical with the impressed field—so that although there is a function I differing from zero, yet $H_i = 0$ and $H = H_e$ everywhere inside the iron. This condition would be realized, if one could make an infinitely long straight rod and expose it to an infinitely extended uniform field parallel to its axis. It is very nearly realized along the middle of a wire several hundred times as long as it is thick, set parallel to the earth's field or along the axis of a solenoid somewhat longer than the wire itself. It is very nearly realized within the substance of a ring-shaped piece of metal pervaded everywhere by an impressed field following the curvature of the ring; a field of this character can be produced by wrapping a current-carrying wire around the ring.

In these cases, or rather in the ideal cases to which these are close approximations, the vectors H_e and I are uniform throughout the metal; the relation between their magnitudes is the relation between

"magnetizing field" and "intensity of magnetization," which is characteristic of the metal and is the cardinal fact of ferromagnetism.

Next we enquire whether it is possible so to shape the metal and so to orient the impressed field, that the actual field within the metal shall be uniform all through it even though not the same as the impressed field—so that I and H_e and H_i shall all three differ from zero, and the resultant H of H_e and H_i shall be uniform throughout the magnet. This condition is realized, if the piece of metal is an ellipsoid and the impressed field is uniform and directed parallel to one of its axes. In this case the ellipsoid is magnetized uniformly, and the extra field H_i which it produces within itself is uniform and oppositely directed,* "antiparallel," to the impressed field. The actual field H is uniform and points everywhere in the same direction as H_e, and its magnitude is equal to the difference between the magnitudes of H_e and H_i. The magnitude of H_i is proportional to that of I, as might be expected, so that

$$H = H_e - NI.$$

The factor N ("demagnetizing factor") depends upon the ratios between the axes of the ellipsoid, and Maxwell developed formulæ for it.

In these cases of ellipsoids, the relation between I and H_e, which is what the data usually supply, is not the true relation between the intensity of magnetization and the magnetizing field. However, the more significant relation between I and $H_e - NI$ can be deduced from the other by a simple graphical artifice. Ellipsoids of different shapes yield very different I-vs.-H_e curves; but the I-vs.-H curves into which these are translated in the aforesaid manner agree with one another, and with the curves obtained from closed rings or exceedingly long wires, very well indeed. Did they not agree, the whole theory would be upset; this procedure therefore is a manner of testing the theory.†

Incidentally, the field H_i produced by the magnet within itself may be far from insignificant. To take an example from Ewing:

* Unless the metal was not properly demagnetized before the application of the field.

† The artifice mentioned above consists in drawing upon the graph, on which orthogonal axes for I and for H_e have already been laid off, an additional axis passing through the origin and inclined to the I-axis at an angle of which the tangent is N. If now the I-vs.-H_e curve is plotted in the usual way, the value of H corresponding to any point P upon the curve is given by the length of the line drawn parallel to the H_e-axis and connecting P with the new axis.

In dealing with rods or other magnets shaped differently from ellipsoids, N may be determined empirically by plotting the I-vs.-H_e curve and drawing an axis so inclined to the I-axis that when the curve is referred to the new axis it coincides with the curve obtained with an ellipsoidal or ring-shaped magnet of the same material; the value of N is then the tangent of the angle between the new axis and the I-axis.

inside a sphere of soft iron exposed to the earth's magnetic field, H_i amounts to 84/85 of H_e, so that only 1/85 of the external field is active within the iron. Since the discovery of permalloy, this instance can be bettered. Within a sphere of suitably prepared permalloy exposed to a field of 10,000 gauss, 0.9996 of that field is counteracted by the magnetized volume-elements themselves.

This counterbalancing of part of the impressed field is sometimes called the *demagnetizing effect of the poles*—a rather unfortunate term, which affords me a pretext for discussing these alleged "poles." The pole of a magnet is like the end of the rainbow; if one were to tunnel into a magnet to get the pole, one would not find it. Or, to draw a better simile from geometrical optics, the poles of a magnet are like virtual images behind a mirror. The virtual image is a point which we reach by retracing the light-rays backward to the surface of the mirror and then prolonging them straight ahead until they all intersect, even though the light-rays themselves actually came up to the mirror from some other direction; the magnet-pole is a point which we reach by prolonging the lines of force down into the substance of the magnet and carrying them on until they meet, although the lines of force actually supposed to prevail within the magnet may not converge at all. The poles, in fact, are like all the other entities supposed to exist inside a magnet—they are imagined, in order to describe and predict the field which the magnet produces outside of itself. For instance, the external field due to an ellipsoid magnetized parallel to an axis is precisely that which two "poles," properly placed upon the axis and endowed with the proper equal amounts of positive and negative magnetism, would produce. If one chooses to visualize these "poles" rather than the ellipsoid, there is nothing to impede him.*

Again it is permissible, in the case of the ellipsoid and in some others, to visualize only the "magnetization of the surface"—to imagine the surface painted over with magnetism, laid on with a density governed by a certain law. At any point P on the surface of the ellipsoid, let $\mathbf{I}$ represent the magnetization of the material, which as we have seen is a vector; let I stand for the magnitude of this vector; let ds stand for the area of a small element of the surface containing P; let θ stand for the angle between the outward-pointing

* The inexactitude of this concept of "poles" leads to some curious lapses of logic in most expositions of the theory of magnetism (including, I am afraid, this one). Even in Maxwell we read: "The ends of a long thin magnet are commonly called its poles. . . . In all actual magnets the magnetization deviates from uniformity, so that no single points can be taken as the poles. Coulomb, however, by using long thin rods magnetized with care, succeeded in establishing the law of force between two like magnetic poles."(!)

Some use the terms "poles" or "polestrength" in the sense assigned to the word 'magnetism" on p. 298.

normal to ds and the vector **I**. Magnetism in the amount $I \cdot ds \cdot \cos \theta$ is to be spread upon ds; magnetism is to be spread over the surface of the ellipsoid with surface-density $I \cdot \cos \theta$. This film of "magnetism" would produce, everywhere outside of the ellipsoid, the same field as the poles or the continuous magnetization which we have imagined as existing inside the ellipsoid. Furthermore, it has a firmer basis in experience than do the poles. For, if a beam of polarized light is directed against the surface of a magnetized ellipsoid, the reflected beam is curiously altered; this effect, known by the name of its discoverer Kerr, is sometimes extremely complicated, but in magnitude it is always proportional to the value of the imagined quantity $I \cdot \cos \theta$ at the point where the reflection occurs; and by promenading a spot of light over a magnetized piece of iron and analyzing at every point the reflected beam, one can actually find how $I \cdot \cos \theta$ varies all over the surface. This property endows the vector I with a physical reality.

There is still one of the effects which a magnet produces outside of itself, which requires our attention; did it not exist, magnets would not play nearly so great a rôle as they do in the life of the world.

Hitherto I have implied that one maps out the external field of a magnet by exploring it with some one of the known field-measuring devices, of which there are several: the magnetometer needle, the bismuth wire which changes its resistance according to the field impressed upon it, the plate of glass which rotates a traversing beam of plane-polarized light to an extent proportional to the field. There is another method essentially different from these, and capable of measuring something which they cannot. One may set up a loop of wire in the neighborhood of the to-be-magnetized piece of metal; suddenly impress the magnetizing field; and measure the sudden rush of charge around the loop. This rush of charge is proportional to the mean value of the magnetic field thus suddenly created in the region enclosed by the loop.* One might map a field in this manner; but that is not the unique feature of the method.

We consider a special and actual case. Take an unmagnetized ring of iron; cut out a thin segment, leaving two nearly parallel end-surfaces facing one another across a narrow gap; take a loop slightly larger than the cross-section of the ring, suspend it in the gap,

* The E.M.F. around the loop at any instant is equal to the time-derivative of the surface-integral, over any surface bounded by the loop, of the component of the magnetic field normal to the surface; in technical language it is equal to the rate of change of the flux of magnetic field through the loop. The rush of charge is equal to the quotient of the time-integral of this E.M.F., which is the difference between the initial and final values of the surface-integral, by the resistance of the loop.

parallel to the end-surfaces; apply an impressed field H_e by sending a current through a coil wrapped around the ring. The rush of charge in the loop testifies that the field established in the gap is vastly greater than H_e, a fact which can be confirmed by the magnetometer or any other field-measuring device. The field in the gap is, in fact, the resultant of H_e and a field due to the magnetized iron. We call it B. Replace the segment, closing the ring; encircle the restored segment with the loop as with a collar; repeat the experiment (after carefully demagnetizing the ring, so as to start afresh from the same condition). The rush of charge is the same. The apparent inference is, that the field B continues to subsist inside the iron forming the closed ring; and the method of the loop seems to be competent to measure, if not the actual force within the metal, at least the average of its values—which would contradict in part my former statement that the field within the iron is unreachable by measurement.

The contradiction involves one of the most confusing assumptions in the theory of ferromagnetism. The field B is greater than the impressed field H_e, whereas the actual field H, which we have been postulating within the iron in order to explain its magnetization, is smaller than H_e. To prove this for the ring might be difficult, since it is a property of the complete ring that the field due to its own magnetization is zero everywhere outside of it as well as inside (so that, incidentally, the method of the loop is the only one giving even an intimation that the ring is a magnet). With an ellipsoid the demonstration is easy. Wrap the loop like a girdle around the middle of an ellipsoid of iron, and suddenly magnetize the iron by impressing a uniform field H_e parallel to one of its axes and normal to the plane of the loop. Measure the rush of charge; it attests that the field established through the loop is much greater than H_e. But the field within the iron, as we have seen already, has been set equal to $H_e - NI$, hence to a value smaller than H_e, in order to account for the field outside.

It is necessary, therefore, to add a third vector B to the pair of vectors I and H which we have already conceived as existing in the depths of the magnet. It is this vector, the alteration of which governs the rush of charge which occurs through a loop encircling the magnet when the magnetization is changed. The rush of charge is proportional to the change in the mean value of B throughout the magnet in the plane of the loop—not to the mean value of H. Making this the definition of B, and considering all the data assembled from experiments on rings and ellipsoids and rods of various proportions, it is found that the observations made upon their external fields by

field-measuring devices and the observations made by the method of the loop are all reconcilable with one another, provided that the vector B is made parallel to I and H and equal to

$$B = H + 4\pi I.$$

The vector B is known as the *induction*. The relation between B and H is often plotted instead of the relation between I and H; naturally if either relation is known the other can readily be found. The ratio of the magnitudes of B and H is called *permeability* and denoted by μ; the ratio of the magnitudes of I and H is called *susceptibility* and denoted by κ or σ or χ.

One might think that this quantity B should be identified with the magnetic field which is supposed to exist within the metal and to magnetize it. Though all the textbooks beseech the student not to confuse the induction with the field (he is usually asked to imagine himself digging variously shaped infinitely small holes within a magnet, and putting an instrument into each to measure the magnetic force inside it), the distinction has an obstinate way of not becoming clear. We should get just as self-consistent sets of curves if we were to plot I against $(H_e + H_i + 4\pi I)$ as we do when plotting I against $(H_e + H_i)$; it would merely be tantamount to adding 4π to the "demagnetizing factor." As a matter of fact nearly everyone, as soon as he begins to theorize about the state of affairs inside magnetized bodies (or polarized dielectrics), promptly assumes that the acting field is something different from the resultant of H_e and H_i. Some make it equal to $(H + \frac{4}{3}\pi I)$, attributing the term $\frac{4}{3}\pi I$ to an action of the molecules which are neither very close to nor very far from the point where the field is being evaluated. Some (Weiss and his many followers) make it equal in ferromagnetic metals to the sum of H and a term nI, the factor n being so enormous that the postulated field is millions of times as great as H and thousands of times as great as B. The extra field, they say, is "not magnetic"; but this distinction is more obscure than the other. Nobody really knows what the field inside a magnetized solid is. The best policy is to continue plotting I and B as functions of H, regarding H as the independent variable sanctioned by tradition.

B. The Relation between Intensity of Magnetization and Magnetic Field

Since all of the actions of magnets are interpreted by supposing that in every magnetizable substance the intensity of magnetization is controlled by the magnetic field in a definite and peculiar way—

that for every magnetizable substance there is a distinctive *I*-vs.-*II* relation—it is evident that this relation, if it exists, must be the fundamental fact of magnetism. The first object of research in ferromagnetism is to discover it for all of the ferromagnetic materials; the second, to devise for each of these materials a model, accounting for the particular form of *I*-vs.-*II* relation which it displays.

On setting about to collate the recorded samples of *I*-vs.-*II* curves, one promptly encounters the last and greatest of the troubles of ferromagnetism. There are infinitely many such curves to be collected, for there is a limitless variety of ferromagnetic substances!

This is not always realized, because of the unfortunate practice of referring to "the three ferromagnetic metals, iron, nickel and cobalt," as though there were but three *I*-vs.-*II* relations to be determined. But in addition, there are ferromagnetic alloys: binary alloys of iron with nickel, of nickel with cobalt, of cobalt with iron; ternary and yet more complex alloys comprising these and other elements, or consisting entirely of elements none of which by itself is ferromagnetic. Anyone acquainted with the diversities of alloys would be prepared to find a truly vast variety of qualities exhibited by these; and he would not be disappointed. Indeed, an alloy may contain one of its elements in so small a proportion as to appear quite negligible—so small, as to be considered a mere casual impurity—so slight, as to be difficult to detect and difficult to expel—and yet so great, as to influence the magnetization in the most drastic fashion. Iron containing a fraction of a per cent of carbon differs as much from pure iron, in regard to its magnetic properties, as either differs from nickel. (Perhaps even what is now called "pure" iron contains a minimal amount of some undetected yet potent impurity, the ultimate removal of which will reveal a whole new set of phenomena!) So there is not a triad, but a multitude of ferromagnetic substances, each of which may be expected to have a distinctive *I*-vs.-*II* relation of its own.

But for each of these substances there is, as it turns out, not one but a legion of *I*-vs.-*II* relations. The curve depends very much on the temperature of the sample—to such an extent, indeed, that as the temperature is raised, the ferromagnetism varies rapidly, diminishes, and finally vanishes. The curve depends also upon the mechanical stresses prevailing in the material, compression and tension and torsion and the complicated combinations of these. It is also liable to be altered by an electric current flowing in the material.

Degree of crystallization likewise matters a great deal. Most of the samples of metal used in the past have consisted of very great numbers of very small crystals, millions of them to a cubic centimeter.

Recently it became possible to make individual crystals so large that one of them, or an aggregate of a few, is by itself large enough to serve as a sample for magnetic testing. The *I*-vs.-*II* relation for an individual crystal is very different from the relation for a mass of tiny crystals of the same material. In fact, the vector **I** is usually not parallel to the vector **H**. A magnetic field, applied to an ellipsoid cut from a single crystal, magnetizes it askew unless the field, and an axis of the ellipsoid, and an axis of the crystal happen to be all parallel to one another. In the polycrystalline mass, these deviations between direction of field and direction of magnetization must be averaged, and cancel one another out; for otherwise, the universally made assumption that *I* is parallel to *II* would not have been effectual. No doubt it is fortunate that Nature, with a rare benevolence, simplified the data first presented to the students of magnetism by this averaging and this cancellation. We cannot however conclude with safety that an assemblage of small crystals will behave just like an assemblage of equally many large ones. Evidently the size of the crystal must influence its *I*-vs.-*II* relations, or else the boundaries between adjacent crystals affect the magnetization, or there is something inherent which changes concurrently with the degree of crystallization. At any rate, whenever the crystallization of a sample is varied, the *I*-vs.-*II* curve is liable to feel it.

Composition and strain, temperature and current, state of crystallization—one must be prepared to find a new way of dependence of *I* upon *II* for each combination and every gradation of these; and yet the half has not been told. Whenever stress or heat are applied to a magnetizable substance, they alter its *I*-vs.-*II* relation, *not merely while they are being applied, but after they are withdrawn.* After such an experiment one may restore the original temperature and the original stress or freedom from stress, but the material is no longer quite the same. Vibrations and concussions, compressions and tensions and twistings, bending and tapping and cold-rolling and hammering, heating and cooling, annealing and quenching, *the very act of magnetization itself*—each of these is liable not merely to affect the *I*-vs.-*II* curve while it prevails, but to transform the substance permanently into another and a distinct ferromagnetic substance, with a system of magnetization-curves distinct from what the sample showed beforehand.

If we could see into the penetralia of a piece of iron, and discern the conditions and the arrangements of its atoms, it is probable that we should see that every such agency leaves behind it some definite and enduring change; and then we should not wonder at (for example)

the fact that an iron wire, which undergoes the experience of being violently pulled and then relaxed, displays very different *I*-vs.-*H* curves before and after this adventure. In certain cases we do observe some sort of an attendant change, as for instance when the iron wire has been stretched so forcefully that it is permanently lengthened, or cold-rolled so vigorously that the X-ray diffraction-pattern due to its little crystals is affected. In other cases we observe no concurrent change whatever, and are forced to assume that there has been an internal alteration of the metal, for which there is no evidence beyond the testimony of the changed magnetization-curve. In the same way, we are prone to assume that when "the burnt child dreads the fire," something is altered within his brain-cells, for which there is no evidence except his change of conduct. As a rule, one would not speak of the brain-cells; one would say that the child has a memory of the painful burn. The ferromagnetic substance also changes its conduct after each experience, as though it remembered. No one, I presume, supposes that it actually has a consciousness which remembers; but the actual responsible alteration, whatever it may be, is often as far beyond detection as the alteration in the brain-cells. The resulting change in conduct, the result of this "memory" of the metal, is what is known as *hysteresis*.

All this makes the designing of a model for a ferromagnetic substance a very difficult and perplexing problem indeed, as we shall discover in due time. For the moment we are concerned only with knowing how much of the biography of a piece of metal must be recorded, in order to give background and value to a determination of its *I*-vs.-*H* curve. A curve inscribed "*This is the I-vs.-H curve for iron*" would not be worth much, no matter how carefully it had been determined nor how nearly pure the iron had been. At this point the physicist must betake himself to the foundry and the rolling-mill, and confer with the metallurgist, and learn the usage of a number of uncouth words such as *swaging* and *sintering* and *cold-working* and *quenching*, and grasp the distinction between cast-iron and wrought-iron and pig-iron and soft steel and hard steel, and observe a number of processes which were discovered so long ago that originally they were practiced without the least assistance from the guiding hand of pure science. The curve for his sample of metal must be labelled with the processes which the sample underwent before and after it came into his hands. Even yet it is not completely settled how many of the details of these processes should be recorded, nor how far back the history of the sample should be traced. One piece of knowledge, however, dispenses us from the risks of this uncertainty; it is known that a long-continued

annealing,* followed by a gradual cooling, obliterates the traces of earlier experiences; and consequently a sample of unknown (or known) antecedents can be restored, by putting it through this process, to a standardized initial state.

Imagine, then, a sample of nickel which since its latest rejuvenation by annealing has undergone a recorded set of experiences; for instance, that it has been "stretched almost to the point of rupture, bent into a circle, and allowed to restraighten itself" (I quote an actual case investigated by R. Forrer). It might now be thought that, so long as the greatest care is exercised to avoid subjecting the metal to new stresses, concussions or heatings, the I-vs.-H curve would be fixed for good. Not so! for in order to determine the I-vs.-H curve, the metal must be magnetized; and magnetization, like stress and heating, is one of the events that leave an imprint, one of the experiences which the metal remembers. If two I-vs.-H curves are measured in succession, the second is generally not the same as the first; during the process of ascertaining the first, the material was changed into a new one. To predict or classify an I-vs.-H curve, one must not only know the composition of the substance, not only have records of its entire mechanical and thermal history since it was last rendered forgetful of its past by annealing, but also have the protocol of all its magnetizings since the last occasion when it was "completely demagnetized"—whether by the annealing which effaced all the memories, or by the gentler process † prescribed by Ewing which cancels the imprints of past magnetizations without destroying those of past stresses and heatings.

To make some choice among this staggering mass of data, it is suitable to concentrate one's attention on two, or rather on one and a group, of the infinite multitude of curves. The first of the chosen curves is obtained by applying to a sample which is freshly demagnetized a magnetizing field H_e which is increased by consecutive small steps, and measuring the field of the magnet after, or (by the method of the loop) the increase of the induction in the magnet during, each of these steps. From either of these sets of data, after making the allowances and the reductions indicated in the first section of this article, one may determine the I-vs.-H curve for steadily-increasing magnetizing fields applied to a piece of metal initially demagnetized.‡

* I use the word "anneal" to denote a long-continued maintenance at a high temperature, irrespective of the rate of cooling thereafter.

† By applying an alternating magnetic field of which the amplitude is at first greater than any field which has been applied to the magnet, and thence diminishes gradually to zero. However, the effect of this process is not quite thoroughgoing.

‡ There is a risk that the increase in magnetization at a certain step may be so great that, when due allowance is made for the demagnetizing effect of the magnet upon itself, it will be found that H has actually decreased in spite of the increase of H_e.

This is what is sometimes called the normal magnetization curve, sometimes the initial curve; I will adopt the latter term.

At this point it is well to recall that most of the curves actually found in the literature are B-vs.-H curves, not I-vs.-H. Since in the right-hand member of the equation $B = H + 4\pi I$ the second term is usually enormously greater than the first, a B-vs.-H curve usually looks exactly like an I-vs.-H curve plotted on a smaller scale. At very high fieldstrengths, however, a B-vs.-H curve continues climbing upward with a constant slope while the corresponding I-vs. H curve runs parallel to the H-axis.

The Initial Curve

The form of the initial curve is peculiar and distinctive. Departing from the origin of the (I, H) coordinate-plane, it ascends, bends upward, passes through a point of inflection, bends over but never quite turns downward; it goes off towards a horizontal asymptote, toward a maximum or *saturation* value of magnetization. Nearly all initial curves display these features, the point of inflection and the saturation; but in all other details, in the lengths and curvatures of the arcs before and after the point of inflection, in the scale of the curve and of its parts, they differ very much from one substance to another, and are altered very much by mechanical and thermal treatments.

Well-annealed substances, iron and nickel and permalloy for instance, display curves which tempt the onlooker to divide them into three segments: a slowly-rising and eventually upward-bending arc starting from the origin, a relatively steep-climbing portion including the point of inflection, a final arc drawing itself close up to the asymptote. A good example is shown in Figure 1. The distinction is accentuated by the hysteresis-loops which originate from the various points of the curve. In the prevalent theories of magnetization, as we shall eventually find, these segments are supposed to result from different processes occurring inside the metal. I will therefore adopt this separation of the curve into three parts, warning the reader to remember that at best there is always something arbitrary in subdividing a continuous curve, and at worst there are substances in which the division into three segments becomes quite impossible to make.

The first segment, extending from the origin to what some call the instep of the curve, may be regarded as a parabolic arc so long as the field is rather low—for iron and nickel, inferior to about one gauss; and for these metals it is sensibly a straight line so long as the field is below say a tenth, or to be safe a hundredth of a gauss.

This rather nebulous statement might be made precise by expressing I as a power-series in H, after this fashion:

$$I = aH + bH^2 + cH^3 + \ldots,$$

and citing experimental values of the coefficients a, b, c, ...; from this a student equipped with a measuring-instrument could determine

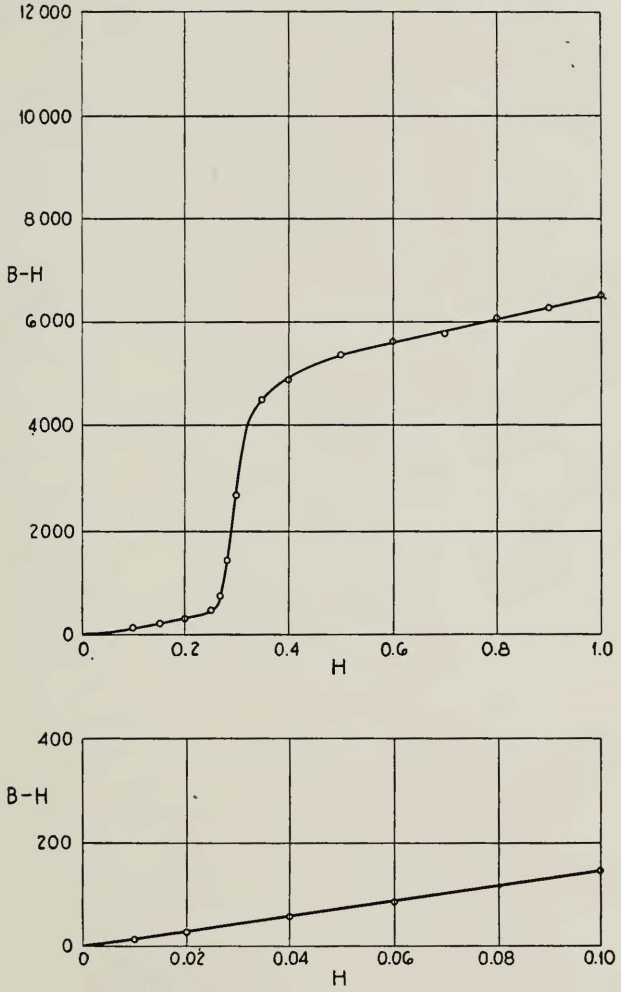


Fig. 1—Initial curve for a permalloy, displaying the three segments. The lower curve is part of the first segment of the upper, on a larger scale. (After L. W. McKeehan.)

the value of H at which the terms after the first become so small that with his apparatus he could not detect them. I mention this to

show how valueless is the mere statement that "the early portion of the magnetization-curve is very nearly straight."

Values of the coefficient a, which is frequently called *initial susceptibility*—the corresponding value of $(1 + 4\pi a)$ is called *initial permeability* and denoted by μ_0—are rather often determined; it is an important constant of each material. Some pairs of values of a and b are quoted by Ewing, others by Bidwell, and some others were determined by the pupils of Weiss. According to one of these latter (Renger) the values for very pure freshly-annealed iron at room-temperature are: $a = 49.9$, $b = 108$. Tempered steel however yielded values of 2.23 for a, and 0.032 for b; from which, and from a mass of other observations on metals hardened by stretching, one sees that the effect of hardening is to lower a a great deal and b a great deal more, so that the curve slopes less sharply upward and does not begin to bend appreciably for a much longer way. I cannot quote all of the relevant data; but it is worth remembering that Rayleigh made measurements so delicate that he was able to follow the curve (for unannealed iron) all the way from .04 to .00004 gauss. Over this range his magnetometer reported no variation in the ratio of I to H. For nickel the detectable upward curvature commences at a much higher fieldstrength—five gauss, according to Ewing.

The alloys of iron and nickel, containing more than 30 per cent of the latter element, develop extraordinary magnetic properties when they are submitted to certain heat-treatments,* as G. W. Elmen discovered towards 1915. For these, the first segment of the magnetization-curve shrinks to a small fraction of the length it has for iron; the two-term formula

$$I = aH + bH^2$$

becomes visibly inadequate at 0.02 gauss, as the curve sweeps upward into its rapidly-rising stage. The value of a for some of these "perm-alloys" is as great as 8000, the value of b as great as 4000.

As the value of H is increased the later terms in the power-series for I bulk larger, and eventually the first segment of the curve passes over into what I have called the second. In this second section the ratio of I to H rapidly rises, and attains the enormous values which form one of the distinguishing marks of ferromagnetic substances, and are responsible for much of their utility in the world of engineering. Plotted against H, the ratio of I to H appears as a curve with a high

* For samples of a certain specified shape and size, this is the heat treatment which was recommended: "They are first heated at about 900° C. for an hour and allowed to cool slowly, being protected from oxidation throughout these processes. They are then reheated to 600° C., quickly removed from the furnace, and laid upon a copper plate which is at room-temperature."

sharp maximum, and so also does the more-commonly-plotted ratio of B to H, the permeability μ.

$$\mu = B/H = (H + 4\pi I)/H = 1 + 4\pi(I/H).$$

Pure iron attains much higher values of permeability than does either of the other metals which can be ferromagnetic when pure. By careful purifying and long annealing, T. D. Yensen elevated $\mu_{max.}$

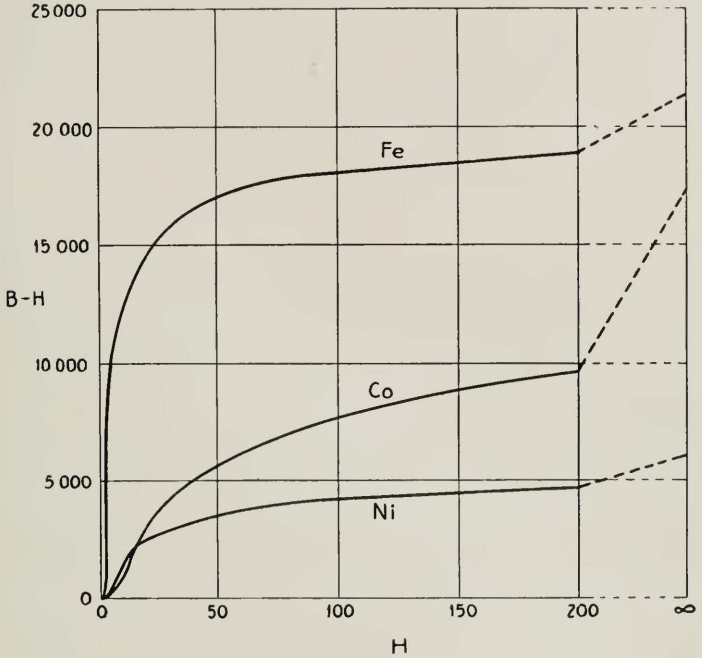


Fig. 2—Initial curves for annealed iron, nickel, and cobalt.
(After L. W. McKeehan.)

for iron to 19000; the best recorded values for nickel and for cobalt are very considerably lower. Certain alloys of iron, however, leave the pure metal far in the rear; by slight admixtures of silicon (between 0.15 per cent and 4 per cent) Yensen produced materials which, after being melted in vacuo, annealed at a high temperature and very slowly cooled, developed a value of $\mu_{max.}$ as high as 66500. These in their turn were surpassed by the permalloys of G. W. Elmen for which $\mu_{max.}$ ascended past 100000.

Other alloys of iron, and in fact nearly all of them, are much inferior to the pure metal in respect of $\mu_{max.}$. Carbon in particular is

pernicious; one per cent of this element mixed with iron brings the maximum permeability down to 350. A few per cent of manganese mixed into iron reduces I/H to the nearly-constant value of .03. Tempering, cold-working, forging, and drawing all tend to reduce the permeability. Since these processes render the metal harder in the literal sense of the word, the change which they imprint upon the magnetization-curve is called by association of ideas a "magnetic hardening." As $\mu_{\max.}$ is reduced by any of these processes, the contrast between the three segments of the I-vs.-H curve diminishes, and in some cases there is scarcely more than the point of inflection left to mark the passage from the initial to the final range of the curve.

The final approach to saturation conforms to the law

$$I = I_{\max.}\left(1 - \frac{c}{H}\right).$$

The value of the constant c is large for magnetically-hard materials, and small for the well-annealed samples for which the tripartite division of the I-vs.-H curve is obvious. In iron (I quote Weiss) saturation is approached within one promille at a field strength of 5500, in nickel at $H = 10000$. In permalloy it must be approached as closely with a field of a few dozens of gauss.

The saturation-intensity of magnetization, or *saturation* for short, is much more nearly independent of the present hardness and the past mechanical and thermal treatments of the material than the other features of the initial curve—much more nearly, therefore, a function of the chemical composition exclusively, than is any other single nameable magnetic quality. For this reason it is possible to present such a Table as the accompanying one with comparatively few qualifications. The first column of figures contains values of $I_{\max.}$ obtained near room-temperature; the second, values measured at the temperature of boiling hydrogen, inserted here for future reference.*

* These values may be described as the "saturation magnetization of a cubic centimetre" of the materials in question. Dividing each by the density ρ of the material, we obtain the "saturation magnetization per gramme." Multiplying this by A, the atomic weight of the element or molecular weight of the compound (if the material is of either sort) we get the "saturation magnetization per gramme-molecule." This last is the quantity most often tabulated, being sometimes expressed in "magnetons" (units equal to 1126 C. G. S. units; cf. page 353). It may be advisable to recall that an *isolated* cube containing one cubic centimetre, or one gramme, or one gramme-molecule of material would not acquire the magnetization in question at any finite field, since it could not be magnetized uniformly.

TABLE

	$I_{max.}$ (20° C.)	$I_{max.}$ (20° K.)
Iron	1706	1742
Nickel	479	505
Cobalt	1412	
Alloy Fe_3Co	1880	
Permalloy Ni 78.5 per cent, Fe 21.5 per cent	870	
Heusler alloy Cu 75 per cent, Mn 14 per cent, Al 10 per cent	222	
Magnetite	490	518
Pyrrhotine	62	

The dependence of the initial curve upon temperature and strain is great and important; but it is expedient to reserve discussion of these variations to later sections.

The Hysteresis-Loops

Any ferromagnetic material has an infinite variety of hysteresis-loops, almost any one of which may turn up in practice; but I will limit this discussion to those obtained by a particular procedure, thus: Commence by demagnetizing the sample—increase H gradually to any desired value, denote this by H_0—decrease H gradually to and through zero, reversing it and bringing it to the equal and oppositely-directed value $(-H_0)$—return gradually to $+H_0$—return to $(-H_0)$—and so over and over again, ten or twenty times at the least. The point representing I as function of H, or B as function of H, traces out at first an arc of the initial curve extending as far as H_0; thence-forward it travels in sweeping detours passing around and around the origin, successive ones becoming more and more closely alike, until at last it settles into a routine of tracing the same oddly-shaped loop over and over again. I have spoken of the "memory" of the magnetic material; this process recalls the consolidation of memory into habit. The final habitual loop thus attained is the particular and chief *hysteresis-loop associated with* H_0. Demagnetizing the sample afresh and repeating the process with a new value of H_0, one gets another loop*; and in this way a family of hysteresis-loops can be determined, one for every point along the initial curve.

So long as H_0 is so low that the initial curve does not depart appreciably from a straight line, the hysteresis also is inappreciable; the point representing I (or B) as function of H goes back and forth

* The demagnetization may be dispensed with, if the new value of H_0 is greater than the prior one.

through the origin along the line of slope a (or μ_0). For this reason, the sensibly-linear part of the initial curve is often called the "reversible part." When it passes over into the perceptibly-upward-turning part, the hysteresis-loop becomes perceptible. Over a certain range its area varies as the cube of H_0, and Weiss gives this formula, in which the coefficient b is used with the same meaning as heretofore:

$$Area\ of\ hysteresis\text{-}loop = \int HdI = \tfrac{3}{4}bH_0^3.$$

In the second segment of the initial curve, the loop swells out to its fullest amplitude. This forms one of the reasons for the division of that curve into three parts; the middle one is sometimes called the "irreversible portion" of the curve. There is no formula available in this region, except the oddly though not universally effective one discovered by Steinmetz, in which the area of the loop is related not to H_0 but to the maximum value B_0 attained in the cycle. This "law of Steinmetz" reads *

$$area\ of\ loop = \eta B_0^{1.6}.$$

Values of the constant η are frequently quoted in describing magnetic materials.

When H_0 is carried far into the third stage of the initial curve, so that in each cycle I approaches within a few per cent of $I_{max.}$, the hysteresis-curve assumes the form of a wide loop prolonged at its northeast and southwest corners (I use the analogy of a map) by long slender projections which narrow down into mere lines. So long as I is nearly equal to $I_{max.}$, the point tracing the I-vs.-H curve passes back and forth along nearly the same path. The final stage of the initial curve is therefore also called "reversible." The Steinmetz formula here becomes invalid.

The reason for laying so much stress on the areas is well known. When a piece of magnetizable metal is carried through a cycle of magnetization, for instance by varying the current through an encircling solenoid in a cyclic manner, the battery supplying the current is found to have expended an amount of energy $\int HdI$ per unit volume; and the metal is found to be warmed to a degree indicating that an equal amount of heat energy has appeared within it.

* The formula of Steinmetz is more general; it applies to hysteresis-loops executed between any two (not overly great) values of induction B_1 and B_2, and for these assumes the form

$$area\ of\ loop = \eta\left(\frac{B_1 - B_2}{2}\right)^{1.6}.$$

It is clear that B_1 and B_2 must be given opposite signs if directed in opposite senses.

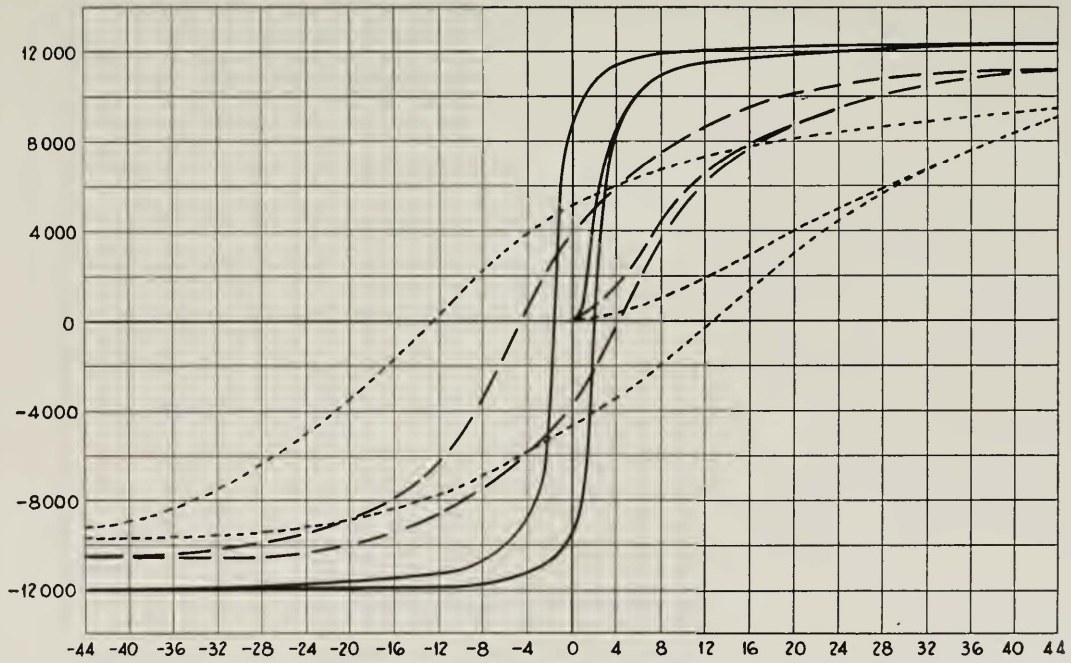


Fig. 3—Initial curves and hysteresis loops for annealed iron (continuous), harddrawn iron (dotted), and hard steel (dashed). (After F. Auerbach.)

We do not know how the transformation from electrical energy to thermal energy was effected; but we do know that so much electrical energy vanishes, and that so much thermal energy appears. A broad hysteresis-loop therefore signifies that there will be much dissipation of energy if the sample is exposed to cyclic magnetizing forces, as usually happens in electric machinery in which magnets play an important rôle; and the heat developed is not merely a sign of energy gone to waste, it is often detrimental to the material, and a bad contribution to that unforgettable history which the magnet is forever piling up. For these reasons the discovery of a new ferromagnetic material of low hysteresis is always welcome.

As a rule, narrow hysteresis-loops go with high values of initial permeability, and with initial curves easily divided into three stages, and with early saturation; and wide loops go with initial curves which rise slowly and bend upward slowly and display no sharply-marked second segment and approach very tardily to the saturation value. Magnetic hardening, in the sense which I earlier defined, accentuates hysteresis; and the agencies which bring it about—tempering and mechanical hardening and the admixture of certain elements in small quantities, such as carbon into iron—widen the loops and augment the generation of heat. These effects are frequently described by giving measurements of the heat W generated in a single cycle of magnetization in which B is carried back and forth between standard values $+ B_0$ and $- B_0$ of the induction-measurements which in their turn are cited by giving the corresponding values of $W/B_0^{1.6}$, the quantity η of the formula of Steinmetz.* The value of this quantity is only .00032 for very pure well-annealed iron, leaps to 0.015 when one per cent of carbon is added, leaps again to 0.034 when the so-constituted steel is tempered; while the addition of silicon to iron, the very process which raises $\mu_{max.}$ to values excelled only by permalloys, brings the value of η down to .00011. The permalloys themselves are still more eminent in this regard, some of them having hysteresis-loops only a sixteenth as great in area as those of pure annealed iron.

To give the area of a loop is not always sufficient; its shape and orientation are very important for theory and for practice. The agencies which harden a material not only widen its hysteresis-loops, but rotate them clockwise around the origin, as the figures show. This rotation tends to decrease the intercept of each loop upon the

* I should emphasize that for many materials the "law of Steinmetz" is not accurate, so that strictly one should plot the actual curve of hysteresis-loss-*vs*-B_0, instead of making a single measurement and using it to determine η by the assumption that the "law" is valid.

axis of I or B, and increase the intercept upon the axis of H.* The former of these intercepts, representing as it does the magnetization which the metal retains when the external field has been reduced to zero, is known under the names of *residual magnetization* and *remanence* and *retentiveness*. The last two of these words, and also *residual magnetism*, are used in a general sense, to denote the property of not losing magnetization altogether when the magnetizing field is withdrawn. The intercept on the axis of H, representing as it does the force which must be applied oppositely to the direction of the prior magnetization in order to annul it altogether, is known under the names of *coercive force* and *coercivity*.

Residual magnetism was the first of magnetic phenomena to come under human notice. If the pieces of magnetite (lodestone) in the fields of Asia Minor had not been able to retain the magnetization which they had acquired in past ages, the Greeks would never have observed nor produced a magnetized metal; if steel needles rubbed against pieces of magnetite or held parallel to the earth's field and "smartly tapped," as the English textbooks say, could not retain the magnetic moment they so acquire, there would have been no compass-needles; the discovery of magnetism would probably have waited upon that of electric currents. Residual magnetism is the property to which the intercept of the hysteresis-loops gives a definite and definable meaning.

The greatest remanence, usually called *retentivity*, is attained after the material is magnetized to saturation. It may be as much as three-quarters of $I_{\max.}$, or more.† Occasionally one finds samples of materials for which the ratio of the greatest remanence to $I_{\max.}$ lies close to some simple fraction—in nickel, for instance, to 1/2. The greatest *ratio* of remanence to previously-attained magnetization, however, is obtained by choosing H_0 somewhere in the second segment of the pristine curve. In fact, there may be a long range of the second segment over which the difference $(I_1 - I_2)$ between the ordinates of the initial curve for any values H_1 and H_2 of the magnetizing field is practically equal to the difference between the values of the remanence in the hysteresis-loops for which $H_0 = H_1$ and $H_0 = H_2$ respectively. In other words: along the second segment of the curve, whatever added magnetization is given to the metal by increasing

* This is not a universal rule; samples of electrolytic iron studies by E. Gumlich and W. Steinhaus (*E. T. Z.*, 36, pp. 675–677, 691–694; 1915) which had been annealed at constant temperatures and cooled at various rates displayed intercepts on the I-axis which were much lower when the cooling had been rapid than when it had been slow; but the intercepts on the H-axis remained nearly unchanged.

† Ewing records an instance of remanence 0.96 as great as prior magnetization (hardened nickel under strong compression).

the field is *kept* almost intact when the field is annulled. Near the beginning and near the end of the curve, the magnetization which is conferred upon the metal by the field departs with the field. Ewing's theory of magnetization is strengthened by this fact.

The greatest remanence, as I have intimated, occurs with magnetically-soft materials. Magnetic hardening tends to augment the coercive force at the expense of the remanence. It does not follow that in constructing a good strong permanent magnet one should take a piece of well-annealed iron or permalloy. A substance of low coercive force is liable to lose its magnetization not only when it is exposed to a weak counteracting field, but also when it is bumped or jarred. A magnet which ceases to be one when dropped on the ground is not of much use in the compass or the automobile.* Great coercive force is much sought after in designing permanent magnets, and the alloys developed for this purpose are at the opposite pole of the ferromagnetic world from the permalloys. The maximum coercive force is attained after the material is magnetized to saturation; for it the name *coercivity* is used and should be reserved. The coercivity of iron, which when the metal is very pure and well annealed may be as low as 0.5 gauss, is elevated past 50 by alloying with one per cent of carbon, past 60 by a few per cent of tungsten, past 80 by a few per cent of molybdenum, up to 370 by amalgamating the iron with mercury. For the permalloys the values drop below 0.05. These figures naturally relate to samples already magnetized to saturation.

Other I-*vs.*-H *curves*

Still other *I*-vs.-*II* curves are obtained in special ways, a few of which I will mention.

If during the measuring of an initial curve the sample is continually shaken, or if after each change in magnetizing field it is traversed by a damped alternating current before the magnetization is read, the *I*-vs.-*II* curve rises very swiftly from the origin; it seems as if the first segment had been suppressed, the second rendered steeper than for the undisturbed sample. Fantastically high values of the ratio I/II are sometimes obtained in this way. Such curves are sometimes called "ideal curves," owing to an impression that they represent the true law of magnetization undisguised by accidental (?) influences.

If in the process of measuring a hysteresis-loop the observer stops

* There is an additional reason for not making permanent magnets out of substances of low coercivity. Suppose an ellipsoid of such a substance magnetized to saturation by an external field H_e; let H_e be reduced gradually to zero; the field $II = H_e - H_i = H_e - NI$ passes through zero long before H_e does, and when H_e finally falls to zero the value of I has fallen far below the true remanence unless the *I*-vs.-*II* curve runs nearly parallel to the axis of H.

short after the first reduction in magnetizing field following the attainment of the maximum applied field (the one which I above called H_0), returns to H_0, and alternates the field several times between H_0 and the inferior value $H_0 - \Delta H$, he finds that the magnetization settles down to a routine of alternating between definite values I_0 and $I_0 - \Delta I$. The limiting value of the quotient $\Delta I/\Delta H$, for small values of ΔH, is known as *reversible susceptibility*. It is a function of I; that is to say, if we select any particular value of I we always get one and the same value of $\Delta I/\Delta H$ when we impart that value of I to the metal, whether by mounting to it along the initial curve or the "ideal" curve or coming to it along any hysteresis-loop.

If the hysteresis-loop is described very rapidly and continuously, it retains its shape surprisingly closely until the frequency is raised into the hundreds of thousands. The initial permeability is still more nearly unaffected by rapidity of variation of field, remaining sensibly unchanged until the range of radiofrequencies is reached and passed. In the range of light-frequencies, however, it is reduced to unity.*

Magnetostriction

"Magnetostriction" is the clumsy name given to the divers very inconspicuous strains in a magnetizable body, brought about by the process of magnetizing it. As they are exceedingly small—a variation of any linear dimension amounting to four parts in one hundred thousand would be ranked as a remarkably big one—and as magnetizable materials are usually investigated in the form of long thin rods, the change in the length of such a rod resulting from a magnetic field applied parallel to its axis ("Joule effect") is the only magnetostrictive change which is often mentioned. Changes in the dimension normal to the field do, however, take place; a rod which expands lengthwise in a longitudinal field will contract sidewise, and *vice versa*. It used to be thought that the change in length just compensates the change in thickness, so that the net change in volume would turn out to be nil; but this turned out too simple to be true. A wire exposed to a longitudinal field and traversed by an electric current will twist itself (the "Wiedemann effect"). This occurs because the impressed field and the circular field due to the current itself are compounded with one another into a resultant pointing slantwise to the axis, so that any particular "line of force" can be visualized as winding in a helix

* In certain materials there is said to be a "magnetic viscosity," because of which the magnetization continues to vary for an appreciable time after an alteration in field is made and ended. The observations upon this are much confused by eddy-currents, and the question is still under debate.

around the wire from top to bottom, like the frieze of the Vendome Column; the expansion (or contraction) of the material along this line of force requires the wire to twist.

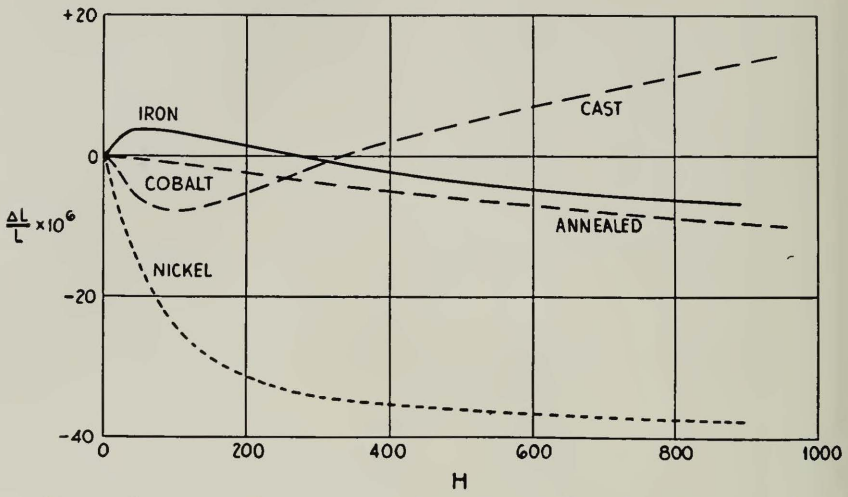


Fig. 4—Magnetostriction (Joule effect) in polycrystalline wires of iron, cobalt ("cast" and "annealed"), and nickel. (After K. Honda and S. Shimizu.)

It is customary to say that, in a gradually-increasing longitudinal magnetic field, nickel contracts continually; cobalt contracts at first, then returns to its original length, then expands; iron first expands, then returns to its original length, then contracts; the Heusler alloys expand continually. Unfortunately, some at least of these statements are valid only for samples which have been and are being treated in particular ways. One finds in the literature, for instance, the information that hard steel and very-well-annealed cobalt behave like nickel, shortening continually as the field is augmented. If the rules which I stated at first are really typical of the respective elements in standard states, then one may lay what emphasis he chooses on the fact that the four consecutive elements which are nickel, cobalt, iron and the manganese which is the essential element of the Heusler alloys are associated each with a different one of the four conceivable permutations of expansion and contraction.

The change in length, whichever its eventual sign, comes to an end when the material is magnetized to saturation. Intensity of magnetization is therefore the natural independent variable on which to consider magnetostriction as depending.

Quite the most exciting of the lately-discovered facts about magnetostriction is disclosed in Figure 4*a*, which consists of curves representing

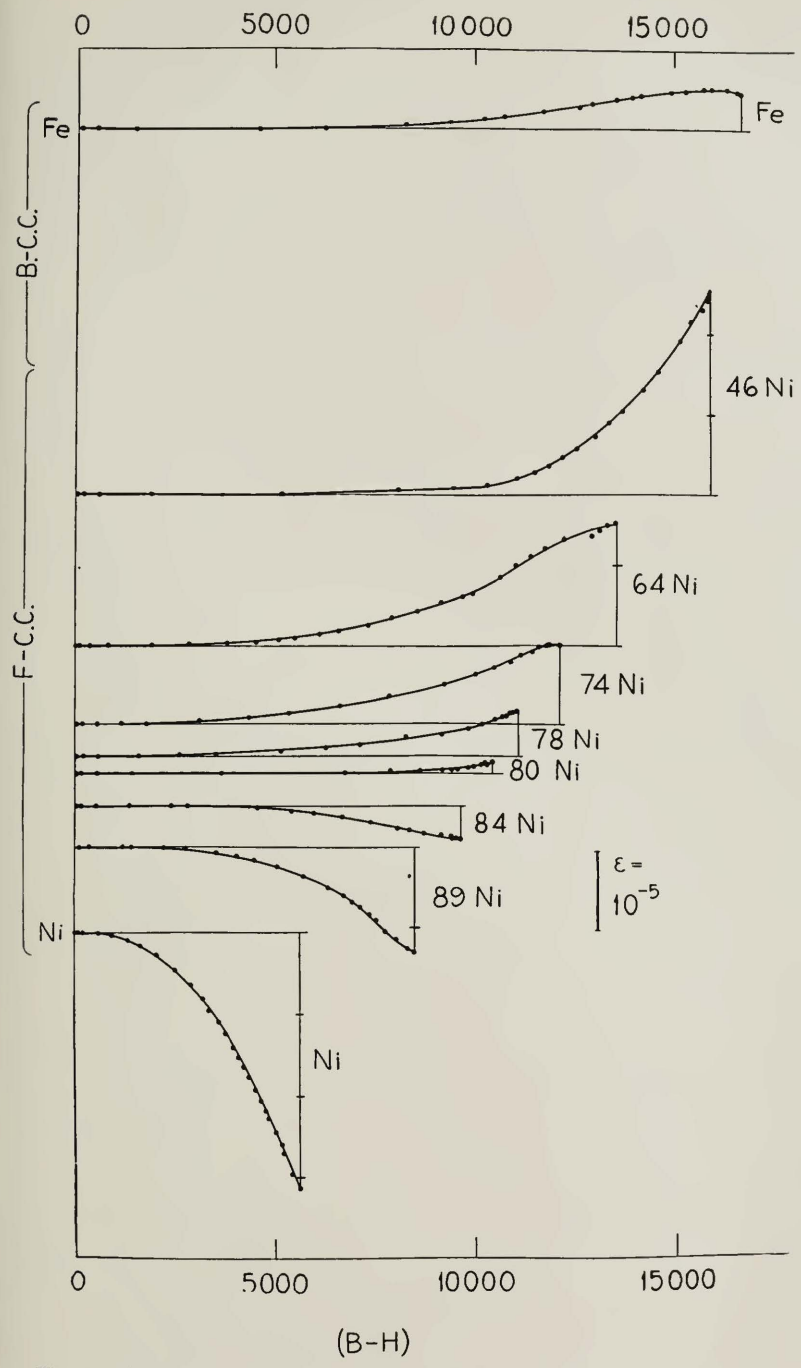


Fig. 4*a*—Magnetostriction in polycrystalline wires of annealed iron, nickel, and various permalloys. (After L. W. McKeehan and P. P. Cioffi.)

change-in-length for iron, nickel, and six permalloys in which the percentages of nickel are those indicated beside the curves. The term "permalloy," I recall, is applied to iron-nickel alloys containing more than 30 per cent of nickel, of which the initial permeability is remarkably high; the heat-treatments which these alloys had undergone conferred that quality on them, the nickel had been treated in quite and the iron in nearly the same way. The abscissa is intensity of magnetization, for the reason aforesaid; consequently the curves terminate when this reaches its saturation-value (not, however, attained in the experiments on iron and nickel).

These curves display the gradations from a steady lengthening reminding the onlooker of the initial lengthening of iron (not, however, followed by a contraction) to a steady contraction approaching the scale of that which nickel displays. Intermediate there lies an alloy which is influenced very little, indeed up to a high stage of magnetization it suffers no perceptible change at all; and this is precisely the alloy having the greatest permeability and the least hysteresis in the entire series. Upon this correlation McKeehan founded his theory of ferromagnetism.

This series of curves reveals other curious facts; for instance the extreme ineffectiveness of the first stages of magnetization in developing the strain—the 46 per cent-nickel alloy had expanded, by the time it was magnetized one third of the way to saturation, by less than one one-hundredth as much as it was destined to expand in acquiring the remaining two thirds of its final magnetization. This is significant; and more significant yet is the point, that when the magnetic field is applied to one of the permalloys containing less than 80 per cent of nickel and subject to a length-increasing longitudinal tension, the magnetostriction is much reduced—that is to say, the mechanical tension seems to have effected of itself a large part of that task of extension which else it would have been incumbent upon the magnetization to perform. It effects a great deal more, of course; the extension due to even a moderate load is vastly greater than the extension which even the greatest of magnetic fields could by itself ever cause; the point is, that the former extension seems to include the latter. Furthermore, elongation by tension is found to produce just as great and no greater an increase in the electrical resistance of a permalloy wire than the much smaller maximum elongation attending magnetization. Yet tension by itself does not magnetize; hence the change which it produces inside the wire does not entirely overlap the effect of magnetic field. It is also true, as one would expect, that tension acting upon a permalloy containing more

than 80 per cent of nickel so affects it that the magnetostriction is increased—the tension seems as it were to have undone something, which the magnetic field must restore before proceeding to the contraction which it operates upon the unstrained metal.

Effect of Tension upon Magnetization

The effects of strain upon magnetization are very complicated, and one would almost despair of ever being able to interpret them, were there not certain relations between them and the effects known collectively as "magnetostriction"—between, to take the simplest instance, the influence of magnetizing upon length and the influence of lengthening upon magnetization—which indicate that law and order reign even in this seemingly chaotic field.

I mention the simplest instance only. A nickel wire, as we have seen, shortens when magnetized parallel to its length; well, when such a wire is shortened by compression, it becomes more magnetizable, the value of I and the value of I/H produced by a continually-applied field H increase; when it is lengthened by stretching (a much easier, consequently a much oftener performed process!), its susceptibility falls off greatly. An iron wire is lengthened when magnetized a little, shortened when magnetized strongly; when it is lengthened by stretching, the magnetization which a weak field imparts to it is increased, that which a strong field imparts to it is diminished; the magnetization-vs.-field curves for different extensions intersect one another somewhere upon the "second segment." Again, a cobalt wire, when lengthened by stretching, has a lower susceptibility in weak fields and a higher susceptibility in strong fields than it does when untensed; this corresponds to the rule governing the magnetostriction of cobalt. To the Wiedemann effect there correspond a magnetization which occurs when a wire carrying a current is twisted, and a rush of current which occurs on twisting a wire already magnetized. The signs of these effects, and of various others, vary from one ferromagnetic metal to another, and vary in iron and cobalt when the magnetization is sufficiently varied, in the ways which may be deduced from the corresponding magnetostrictive effects.

The variations in magnetization produced by extension may be very much more striking than the variations in length produced by magnetization. In nickel, for instance, the susceptibility of a wire may be reduced to a tenth of its pristine value by stretching the wire, although the utmost change in length which can be brought about by magnetization is less than one part in ten thousand.

The relations between the influence of magnetization on strain, and the influence of strain on magnetization, have been derived from the laws of thermodynamics. It appears that each of the several effects agrees with the theory insofar as the sign is concerned (for instance, tension applied to a wire which shortens when magnetized should diminish its susceptibility, and does) but not always in magnitude. I have not heard of anyone renouncing the laws of thermodynamics on this account.

Hysteresis plays a great part in the effect of tension on magnetization; if a constant magnetizing field is applied to a wire while the tension is being cyclically varied, the magnetization when plotted as function of extension follows a hysteresis-loop. Also the first application of a load to a wire is likely to make a sudden and violent change in the value of I. Some avoid these troubles, or try to, by shaking the wire continually or by continually applying an alternating magnetic field during the measurements. These introduce further complications. In fact, if all the data that could be assembled concerning the effect of strains upon magnetization were to be sought out, I suspect that "the world itself could not contain the books that should be written."

The Barkhausen Effect

Imagine once more a piece of some ferromagnetic substance, encircled by a magnetizing coil, through which the current is being steadily increased; encircled also by a loop, which is connected to the voltage terminals of an oscillograph, or to some other device which moment by moment records the electromotive force impressed upon the loop by the changing magnetization. This electromotive force, as I have said, is proportional to the rate-of-change dB/dt of the induction, for which the changing of the magnetizing field is responsible. It is a measure of the rate of magnetization of the sample girdled by the loop. The magnetizing field is being increased continuously; were the magnetization also to rise continuously towards its saturation-value, as we should probably expect, the voltage-curve would be smooth. However, when a sensitive oscillograph is used, the curve is a succession of sharp teeth. The magnetization of the sample evidently proceeds by small but sudden jumps. These can be shown—in the most literal sense of the word "to show"—by connecting a telephone-receiver through an audion-amplifier to the loop. Listening at the receiver, one hears a rustling or a crackling sound; it has been compared with rain beating upon a tin roof, also with coal rattling down a chute. Barkhausen discovered the effect in this way.

By increasing the magnetizing field very slowly it is possible to space the peaks in the oscillographic curve, or the clicks in the receiver, so widely that the bigger can be counted. Listening to the separated clicks, van der Pol estimated that the process of magnetizing a cubic centimetre of iron or of an iron-nickel alloy involves several thousand of the jumps. It is also possible to measure the area under each of the larger peaks in a curve obtained with a good oscillograph, and calculate from it the magnetic moment of a magnet, the sudden creation of which within the substance would have resulted in just such a peak. One observed by E. P. T. Tyndall could have come about through the sudden creation of a magnet of moment .0027. The word "creation" must not be taken too literally; it might imply, for instance, that two adjacent magnets were at first pointed contrariwise to one another, and one of them was suddenly wheeled around by the field, so that they ceased to neutralize each other. Data such as that just cited from Tyndall would then indicate the sizes of the magnets preexisting in the substance; data such as those of van der Pol, their number. Both sets of data show that one cannot identify these magnets with individual atoms; they are too large (the moment .0027 is as great as that of a piece of saturated iron 0.12 mm. on a side) and too few. Neither can they be identified with individual crystals; a piece composed of a single crystal makes as much noise in the receiver, while being magnetized, as a fine-grained sample. The data suggest that ferromagnetic metals are built up out of magnetic units larger than atoms and smaller than crystals—a suggestion which to the theorists is often extremely acceptable. It is also a welcome fact, that the peaks and the crackling are associated with the steeply-sloping segments of the magnetization-curve, while the initial and final nearly-horizontal arcs of the curve are smooth and silent.*

Magnetization of Single Crystals

Ferromagnetic crystals large enough to be studied are only just ceasing to be a rarity. Only two sorts occur in Nature: those of magnetite (a modification of one of the oxides of iron, Fe_3O_4) and those of pyrrhotine (a sulphide of iron, Fe_7S_8). To procure single crystals of a metal or an alloy, it used to be necessary to wait on the hazards of the foundry, out of which there might arise at long intervals a single large uniformly-crystallized lump. This condition prevails

* Attention must be drawn to the possibility that the peaks in the curve, or the clicks in the sound, are due to fortuitous coincidences of events individually too insignificant to be perceived. Should this turn out to be the case, the Barkhausen effect would resemble the Schroteffect of thermionic emission, and the interpretation of the data would be changed.

no longer; there are methods for producing large single crystals of metals at will, whether by direct solidification from the melt or by suitable treatment of the masses of randomly-disposed minute crystals which blocks of metals usually are; and there are methods for determining the orientations of the axes of these crystals by means of X-rays. So lately have these methods been developed (they are outgrowths of researches of the last ten or fifteen years) that the first data concerning the ferromagnetic crystals, except for some relating to magnetite and pyrrhotine and a very few early measurements on iron, are only now appearing. One has at times a feeling that these are the first really significant data, the only suitable foundation for a theory of ferromagnetism; that the properties of a polycrystalline rod or wire or ellipsoid do not form a proper basis for theorizing, not being even a simple average of the properties of single crystals oriented in all directions, but a deformed and distorted average infected by the crowding and the cramping and the squeezing which the little crystals perpetually inflict on one another.

All but two of the well-known ferromagnetic substances crystallize in the cubic system. (The exceptions are pyrrhotine and one modification of cobalt, which conform to the hexagonal system). In cubic crystals, directions parallel to the edges of the cubes, to their diagonals, to the diagonals of their faces, are called the tetragonal, trigonal, digonal axes, or the quaternary, ternary, binary axes respectively; the planes to which these directions are perpendicular are called (100) planes, (111) planes, (110) planes respectively. This is as much of the technical language of crystal analysis as we shall require. Of the three lattices in which atoms may be arranged in a cubic crystal—simple cubic, body centred, face centred—iron adopts the second, nickel and cobalt the third. The iron-nickel alloys containing more than 30 per cent of nickel copy the nickel lattice (the permalloys belong to this class) while those containing less than 30 per cent of nickel imitate iron. Many other metals which are not ferromagnetic have cubic lattices of the second or third type, none at all a lattice of the first; it is therefore futile to look for any correlation between ferromagnetism and the arrangement of the atoms.

When a magnetic field is applied to a crystal, it produces a magnetization which is not parallel to the acting field—to the resultant, I mean to say, of the applied field and that due to the "demagnetizing effect of the poles"—unless this resultant is parallel either to a tetragonal or to a trigonal or to a digonal axis. If we apply a field parallel to the axis of an ellipsoid or a long rod, cut from a single crystal in such a way that this axis is parallel to one of the specified directions,

the I-vs.-H curve mounts much more swiftly than does the normal curve for polycrystalline iron; the first segment is very short, and the second passes into the third while the field is still low. The slope of the first part of the curve, that is to say the initial susceptibility, is greatest when the axis of the rod is a tetragonal axis of the crystal, less if it is a digonal, least if it is a trigonal axis; though the differences (Fig. 5) are not great. This is sometimes expressed by saying that iron is most easily magnetized along the tetragonal axis, less so along the digonal and least along the trigonal. Magnetization curves consisting of three or four straight lines meeting at sharp corners have been observed by two of the recent students of single crystals, but not by two others; I infer they are still debatable. The saturation value of I, whether it be

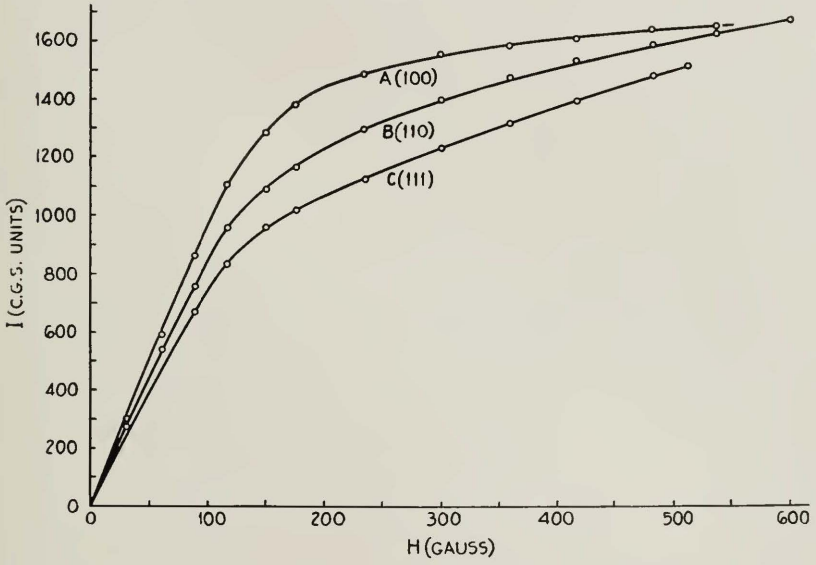


Fig. 5—Initial curves of a single crystal of iron, magnetized parallel to tetragonal (100), digonal (110), or trigonal (111) directions. (After W. L. Webster.)

attained soon or late, seems always to be about the same—another of the reasons for attaching a peculiar importance to it. Honda in fact obtained the value 1707, which he confronts with the 1706 given by Weiss for polycrystalline iron; but this is an agreement which looks too good to be true, or at least to be significant.

The hysteresis-loop for a single crystal is so exceedingly narrow that when it is plotted on any ordinary scale, its sides are too close to be distinguished. Measurements upon rods composed of many crystals, the average size of which varies from rod to rod, show that the area of the hysteresis-loop decreases quite steadily as this average

size of the "grains" is diminished. This is a potent argument against all theories in which hysteresis is attributed to an arrangement of atoms in a uniform space lattice.

When a magnetic field is applied to an iron crystal in any direction not parallel to one of the axes, the magnetization is not quite parallel to the acting field. This manifests itself, for instance, when one cuts a disc out of a crystal and exposes it to a magnetic field in its own

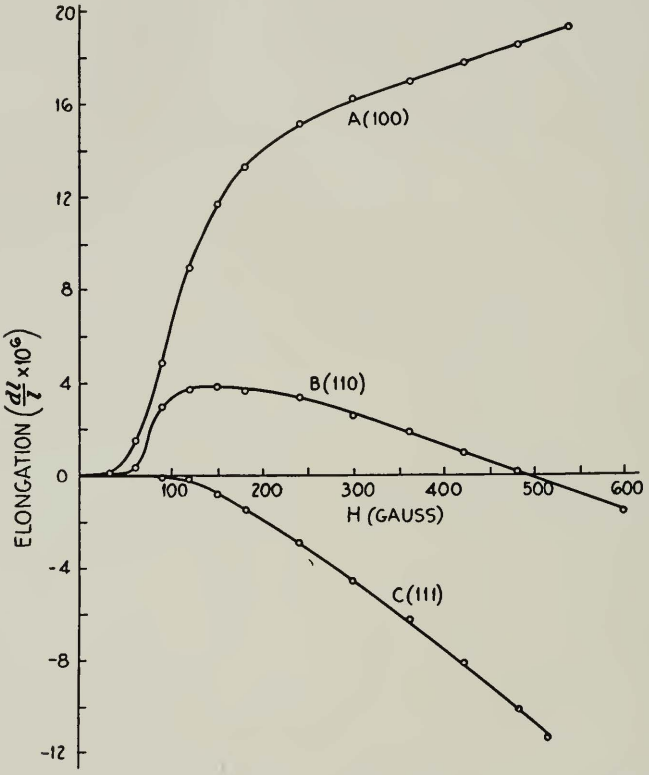


Fig. 6—Magnetostriction of a single crystal of iron, magnetized parallel to tetragonal, digonal, or trigonal axes. (After W. L. Webster.)

plane; it cannot rest in equilibrium until it has so turned itself that one of its three preferred directions lies parallel to the field, for otherwise there is a component of the magnetic moment which suffers a torque from the very field which evoked it. The angle between the vectors **I** and **H** seldom attains and never exceeds twelve degrees; when the field is kept constant in direction and varied in magnitude, this angle of deviation is less for very weak and less for strong fields than for some intermediate value of fieldstrength. In pyrrhotine,

however, the angle may be enormous—a field inclined at no more than five or ten degrees to the hexagonal axis produces a magnetization which, when investigated by delicate methods, seems to lie exactly in the plane perpendicular to the hexagonal axis, which consequently is known as the "plane of easy magnetization." A sphere of pyrrhotine to which a bar magnet is brought up from the direction in which its hexagonal axis points does not seem to realize that the magnet is there, but if the approaching magnet is displaced a little

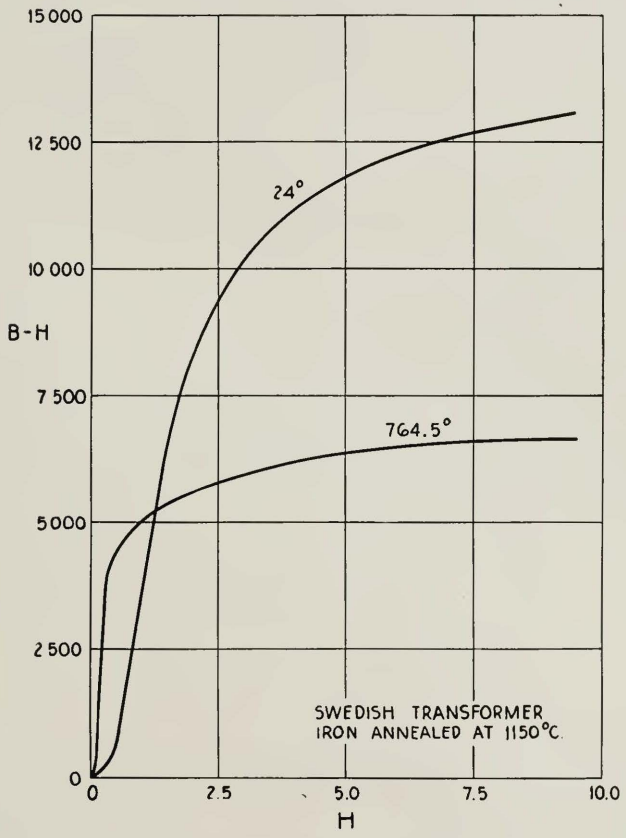


Fig. 7—Initial curves for Swedish transformer iron at two temperatures. (After D. K. Morris.)

sidewise the ball flies over to its surface at once. It will readily be seen what complications these facts introduce into the mathematics of predicting or describing the magnetization of an arbitrarily-shaped solid body—and in this connection it is well to remember that an ordinary polycrystalline mass of metal partakes as soon as it is strained, by pulling or rolling, of some of the properties of a single crystal.

Magnetostriction in single crystals has some very curious features. A crystal of iron unites in itself all the three modes of magnetostriction which have been supposed typical of iron, nickel and Heusler alloys respectively. A rod having a tetragonal axis along its length expands continually when exposed to a longitudinal magnetic field; a rod cut along the trigonal axis contracts continually; if cut along the digonal axis it first expands, then returns to its original length, finally contracts. The expansion in the first of these cases may attain twenty parts in a million—four or five times as great a value as one ever finds with a polycrystalline sample. This shows how great the extent to which the little crystals in an ordinary block of iron must interfere with one another when the block is magnetized.

Dependence of Magnetization on Temperature

As the temperature of a sample of iron is raised, its normal magnetization-curve varies in a manner suggesting the influence of tension;

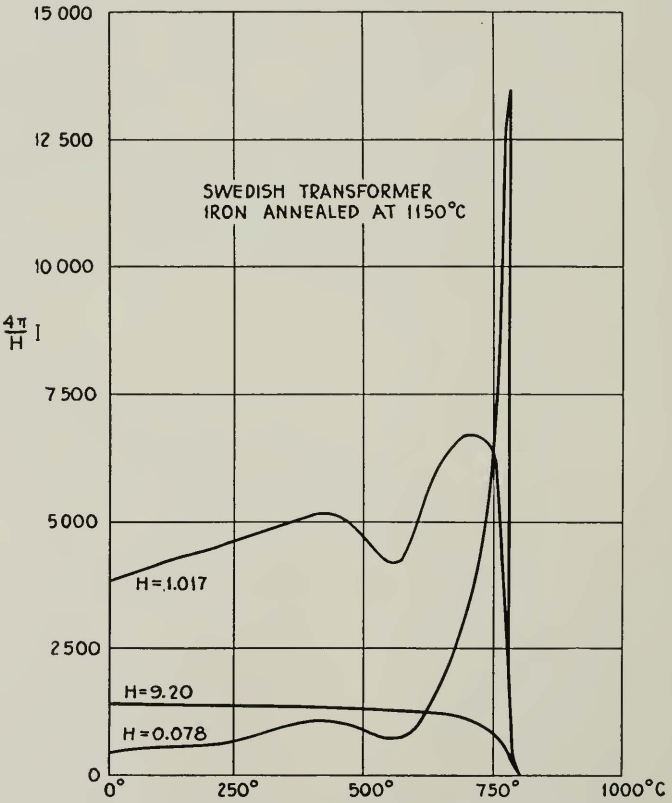


Fig. 8—Magnetization-vs.-temperature curves for Swedish transformer iron at three values of magnetizing field. (After D. K. Morris.)

the earlier part is exalted, the later part is depressed, so that the susceptibility increases in low fields and diminishes in high; curves obtained at different temperatures, not too far apart, intersect one another somewhere upon the "second segment" (Fig. 7). On plotting I or I/H for individual fieldstrengths as functions of temperature, one obtains curves which for very low fieldstrengths, such as 0.3 gauss for instance, are remarkably shaped (Fig. 8). The initial susceptibility rises to an enormous height at a temperature slightly above 700° C., and then precipitately falls almost to nothing—it does not quite vanish, but instruments of a much higher order of sensi-

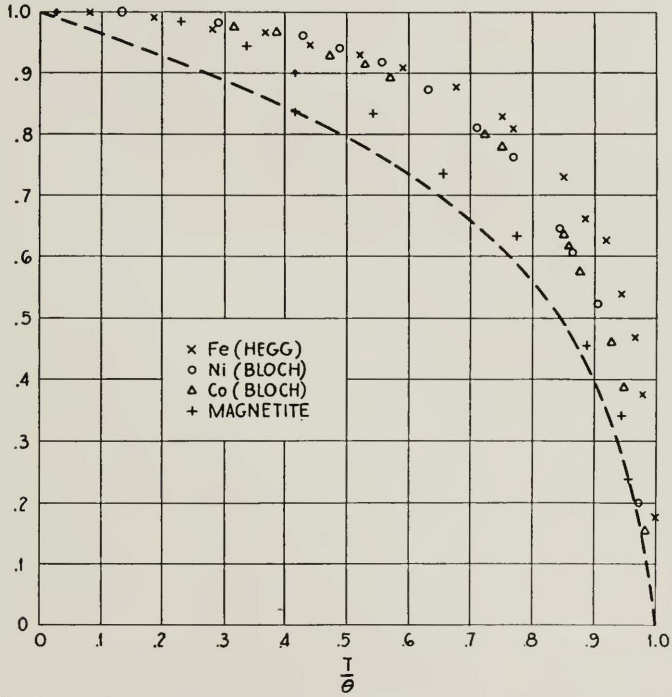


Fig. 9—Saturation-vs.-temperature data for iron, cobalt, nickel, and magnetite below their respective Curie-points, with a theoretical curve. Abscissa is ratio of absolute temperature to Curie-point temperature for each substance individually. (After P. Weiss.)

tiveness are required to detect or measure it beyond say 770°. At a somewhat higher fieldstrength, about 4 gauss, the I-vs.-T curve is nearly horizontal for a long way, and then declines gradually to the axis of H, which it reaches near 770°. At higher fieldstrengths the decline sets in progressively earlier (Fig. 8). At very high fields one obtains what is substantially the curve of $I_{max.}$ versus T (here the analogy with the effect of tension breaks down) which is shown in Figure 9.

The temperature at which I-vs.-T curves intersect the axis of T, or would intersect that axis were it not that they turn aside shortly before reaching it, is known as the *Curie-point*. For iron, values of Curie-point ranging from 768° to 790° are given; the differences seem to be due partly to uncertainties in deciding just where the I-vs.-T curves "would intersect the T-axis if they continued on downward without turning," partly to the indubitable fact that these intersection-points are not the same for I-vs.-T curves for different values of H, and partly to the use of other definitions of the Curie-point. For nickel, cobalt, and magnetite the Curie-points are in the neighborhood of 360°, 1130° and 550° respectively; and values are recorded for a considerable number of alloys.

The Curie-point is not the sign of what is properly designated as a "change of phase." Iron suffers changes of phase at temperatures near 900° and near 1400°, changes in which the atom-lattice goes over into an entirely different type, and a number of physical properties are sharply altered; but the Curie-point is not one of these, it is the locality of merely a rapid (though not absolutely sudden) change in magnetic properties and an evidently-correlated anomaly in specific heat.* As for the real changes of phase, they normally occur at temperatures so high that they do not influence the magnetization of iron below the Curie-point. Yet it is possible to bring one of the high-temperature modifications suddenly down into the low-temperature range, and then its magnetic properties are quite different from those of "ordinary" iron. In certain alloys this possibility is easy to realize; I will quote only the notorious case of a "nickel-steel" discovered by J. Hopkinson, which at 580° C. is merely one of the many non-ferromagnetic metals, remains so as it is cooled all the way down to zero, then turns suddenly into a modification which is strongly magnetizable and retains this state as it is being heated all the way back to 580° C. But indeed ferromagnetism of alloys is entangled with all the infinite complexities of the behavior and the internal changes of these complicated substances, and varies with all the variations of the more or less durable equilibria between their components.

Definition of Ferromagnetism

Ferromagnetism has sometimes been defined as "the kind of magnetism which iron exhibits"—an easy evasion, to which one is

* Contrary statements about iron are to be found in the early literature; but they are due partly to inaccurate experiments, and partly to the fact that the change-of-phase which in pure iron lies well above the Curie-point descends when carbon is progressively added to the iron, and before long comes into coincidence with the Curie-point; and if still more carbon is added, the "vanishing of ferromagnetism" takes place at the transition temperature.

sorely tempted to have recourse after the first few efforts to devise a better definition. Let us, nevertheless, at least take notice of a few of the alternative proposals.

Materials are classified into diamagnetic and paramagnetic and ferromagnetic. To distinguish those of the first sort is relatively easy, since in any of them the magnetization I called forth by an applied field H is antiparallel to H (in isotropic materials, at least; in crystals the angle between I and H lies between 90° and 180°). In materials of the second or of the third sort, the vectors I and H are parallel and point in the same sense, or at least are inclined to one another at angles smaller than 90°—provided, that is to say, that the material was demagnetized before H was applied. To distinguish between paramagnetic and ferromagnetic bodies, therefore, we must seek some other criterion.

The magnetization of iron, nickel, cobalt, certain of their alloys with one another and with other metals, and the Heusler alloys, may attain values enormously greater than those which can be impressed upon other substances with the highest possible fieldstrengths. One might therefore select some intermediate value for I, and say that all substances for which I may surpass this critical value are ferromagnetic, all others paramagnetic (or diamagnetic). In practice this is usually convenient, because of the great contrast between the substances which I just listed and practically all others. Among the elementary metals apart from the iron-cobalt-nickel triad, one of the most magnetizable is platinum, which shares a column of the periodic table with that triad; yet its susceptibility is only $2 \cdot 10^{-5}$, and an applied fieldstrength of 20000 gauss would impart to it a magnetization of less than one unit, which is utterly negligible compared to those which are easily imprinted even upon the less magnetizable of the substances which I named. The contrast is therefore great enough to be the basis for a useful definition. Yet it must be regarded as accidental, that in practice we are nearly always confronted with extreme cases of one sort or the other. If we travel along the iron-manganese or the nickel-chromium series of alloys (to take but two instances), or if we follow pure iron through a sufficient range of rising temperatures, we find a continuous series of intermediate stages between one extreme and the other; and in principle it is necessary to take account of these.

The magnetizations of iron, cobalt, nickel, certain of their alloys and the Heusler alloys increase, when the applied field is continuously increased, in the curious ways which I described above, attaining maximum limiting-values at fieldstrengths well within the practi-

cable range; while with nearly all other materials I is apparently proportional to H as far as the field can be carried. Here again there is a contrast so great that it can serve as the basis of a useful distinction. Yet all the intermediate stages between the two extremes are exhibited by iron at the various temperatures between 700° and 800° C. Furthermore, there is reason from theory (as we shall see) for supposing that the magnetization of any substance would cease to be proportional to H and would approach a limit, if we could force the field to high enough or the temperatures to low enough values. In fact, there is at least one of the substances conventionally called "paramagnetic" (it is gadolinium sulphate) for which I was found to approach a limit, when the applied fieldstrength was increased while the substance was maintained at the unprecedentedly low temperature of 1.°3 K. It is therefore evidently something of an accident that in practice we nearly always meet either with substances for which the ratio I/H is constant within the accuracy of measurement throughout the feasible range of the fieldstrengths, or else with substances for which that ratio varies greatly and unmistakably with the field.

Presence or absence of hysteresis is the third and last of the usual criteria. Iron and cobalt and nickel and some of their alloys and the Heusler alloys exhibit hysteresis-loops, and residual magnetism, and coercive force; and the normal magnetization curve must be distinguished from curves obtained by other procedures for varying the applied field, and one must bother with demagnetizings or else take account of the prior magnetization of whatever sample he is working with. Other substances are free from these complications. Here also it is probable that in iron all the measurable features of hysteresis dwindle off continuously to zero as the metal is heated. On the other hand, it appears that gadolinium sulphate, in spite of acquiring a curvature in its I-vs.-H curve at extremely low temperatures, does not acquire hysteresis and residual magnetism. Perhaps, then, it is better to take the presence of hysteresis rather than the inconstancy of the ratio I/H as the sign of that curious quality, whatever it may be, which makes iron notable among metals.

The general conclusion seems to be the same, as for many other classifications—that is to say: It is possible to draw distinctions between "ferromagnetic" and "paramagnetic" substances, valid for extreme cases of the two types, not sharply marked for intermediate cases; but it happens that for the time being the intermediate cases are in practice not conspicuous; and consequently the distinctions—any one of the three which I mentioned—are useful and worth the making.

C. Theories of Ferromagnetism

To devise a theory of ferromagnetism is not necessarily the same task as to make a theory of magnetism. In studying the properties of paramagnetic and those of diamagnetic bodies, one finds many indications that the ultimate atoms of the elements are magnets of definite and seldom-changing moments, or at least may profitably be so regarded. The theory of line-spectra reinforces this opinion, and it is confirmed by the observations of Gerlach and Stern upon the deflections undergone by free-flying streams of atoms traversing a strong magnetic field with a strong field-gradient.* Now, to say that atoms are magnets is scarcely tantamount to giving an explanation of magnetism. On the contrary, the problem is merely pushed a step further away, and must eventually be faced again and either be solved by explaining why atoms are magnets, or else be given up by conceding that magnetism is one of the fundamental properties of matter. Yet it is quite logical and sensible to aspire to construct a theory of *magnetization*—of the gradual magnetizing of a substance by an increasing applied field, of the shapes of the *I*-vs.-*H* curves, of hysteresis-loops—out of the assumption that the ultimate atoms are permanent magnets. To explain the gradual rise of an *I*-vs.-*H* curve by postulating atoms which are already magnetized to saturation, to explain hysteresis by postulating atoms which individually have no hysteresis—these would be triumphs not open to the objection made against many "explanations," that they are achieved by ascribing to the atoms the very properties to be explained.

We shall presently make the acquaintance of "elementary magnets" —hypothetical beings, of which each magnetizable substance is supposed to consist. To these we shall assign, for the time at least, definite and unchangeable magnetic moments. A magnetic field applied to an assemblage of such magnets could not change the moment of any. Yet it could change the net magnetic moment of the assemblage, which is the resultant of the moments of all the individuals; for it could, directly or indirectly, cause the elementary magnets to align themselves along its own direction. The assemblage, the substance, would be magnetized *not through magnetization of the individuals but through orientation of the individuals* which make it up.

That idea is an old one; but by itself it is nearly useless. We must think of some agency which could combat the tendency of the elementary magnets to align themselves along the field; for there must be such a one, as otherwise the weakest possible field would magnetize each substance to saturation; which is not the case. The most

* I refer for these to my *Introduction to Contemporary Physics*, pp. 48–50, 383–393.

celebrated theories of magnetization rest upon speculations about the nature of this agency which fights against the field.

On considering the unsurpassably simple system composed of *only two* elementary magnets close together, J. A. Ewing discovered that their interactions are such, that they can prevent each other from aligning themselves immediately along the field; one can almost say that they "interlock," and they interlock in such a way, that the pair of them displays a tripartite I-vs.-H curve, and the quality of hysteresis, though neither separately has any such properties. Systems comprising a dozen, a score, or a multitude of such magnets, arranged in chains or in a cubical array, may be devised to imitate actual initial curves and actual hysteresis loops with stunning accuracy (Fig. 11). Such close agreements need not be overstressed. The astonishing feature of Ewing's discovery is (I think) that although each individual magnet possesses neither the quality of gradual magnetization nor the quality of hysteresis, a pair of them put close together possesses both. So great a result is attained from so simple an apparatus, that it seems very unlikely that any radically different explanation of either quality will ever be put forth. Whatever may be added to Ewing's model, its central idea will probably never be supplanted.

P. Langevin, devising a theory for paramagnetism, supposed that the agency which combats the aligning influence of the field is the thermal agitation of the magnetic atoms. Contrary to one's first impression, this theory is not easily visualized; but it establishes a union between paramagnetism on the one hand, and the great general principles of thermodynamics and equipartition of energy on the other. In the form in which Langevin put it forth, it does not account for hysteresis.

P. Weiss supplemented Langevin's theory by supposing that the actual magnetizing field prevailing inside a magnetizable substance is not that sum of the applied field H_e and the "demagnetizing field" H_i which I defined in Section A, but a combination of this sum with another term depending on the magnetization. As I stressed in Section A, experience teaches us nothing about the value of the true field inside a magnet; Weiss' assumption was therefore a perfectly legitimate choice, to be justified (if at all) by its fruits. One of these is, that it accounts for the presence of hysteresis at low temperatures and its absence at high.

Ewing's Theory

Ewing conceived a piece of iron as an assemblage of tiny bar-magnets, each endowed with a fixed and constant magnetic moment, and wheeling about a pivot under the combined influence of the

impressed magnetic field and the magnetic attractions and repulsions of its neighbors.

Imagine a chain of long slender bar-magnets end to end, the positive pole of each almost touching the negative pole of the next—that is the equilibrium position which they would naturally assume, so long as no external field affects them. By preference, build such a chain out of pivoted magnets; for Ewing's model enjoys the singular merit, that it can be made out of actual magnets and exhibited to the eye. Now there is a remarkable feature of this chain: if a magnetic field is applied to it in some oblique direction, then so long as the fieldstrength is quite small the individual magnets incline themselves toward it slightly, each setting itself at the same angle to the direction of the chain which was originally the common direction of them all; and when the fieldstrength is gradually increased the angle increases gradually, but only up to a certain point—for suddenly, at a critical moment, all the bar-magnets very suddenly capsize, and set themselves in nearly the direction of the field. I use the word *capsize* to invoke the too-familiar analogy of the upsetting boat. As weights are piled upon one side of a boat, it responds at first by tilting gradually sidewise and downward; to each slight increment of the load it accommodates itself by finding an equilibrium-slant a little farther over; but eventually there comes a moment when balance and compromise are no longer possible; the boat cannot find a position of equilibrium except by overturning, and this it does, suddenly and irrevocably. Such is the behavior of a chain of bar magnets; and this is the property which adapts it for representing the general shape of an initial magnetization-curve such as I showed in Fig. 1, with its first slowly rising arc followed by the rapid uprush and the final slow adjustment to saturation.

The overturned boat will not right itself even when the load which upset it is removed; will the chain of bar-magnets be equally unforgiving? The analogy is not perfect, except in one very particular case: if the angle between the direction of the chain (defined as the direction in which the north poles of all the magnets originally pointed) and the direction of the field is 180°, the capsizing will result in a right-about-face of each magnet and a reversal of the so-defined direction of the chain, and this reversal will persist after the field is annulled.

Suppose however that there is a multitude of chains oriented at random, so that half of them are inclined at less than 90° and half at more than 90° to the direction of any strong field which we choose to apply. The field will cause all the bar-magnets to capsize (except those belonging to the few chains to which it is almost parallel);

and thereupon, those which belong to the chains originally inclined at more than 90° to the field are more than halfway turned around, and when the field is nullified they will realign themselves with their first associates, but every one will be reversed. Originally the net magnetization of the assemblage of chains was nil, for half neutralized the other half; now it is considerable, for half have been inverted. Its ratio to the total magnetization of all the chains when parallel is, in fact, one half. This consequently would be an adequate model for a substance of which the remanence is one half of the saturation-intensity.

Other values than one half for the ratio of remanence to saturation can be derived from Ewing's picture by choosing a suitable arrangement for the elementary magnets. Suppose, for another and a final example, that they are arranged in a cubic lattice, so that each has the choice (as it were) of orienting itself along any one of the directions parallel to the cube-edges. Chains of magnets may then form themselves along any one of six possible directions (counting the two opposite senses of any line parallel to a cube-edge as two distinct directions). In a demagnetized crystal, one may imagine that the elementary magnets in the lattice fall into groups or "complexes," within each of which all the chains are parallel, while from one complex to the next they change over from one to another of the six specified possible directions. In a demagnetized piece of metal composed of many small crystals oriented quite at random, there will be chains of magnets pointing in all directions. To such a piece of metal let a field be applied, increased to so great an amount that it saturates the material, and reduced gradually to zero. Whatever the direction of the field, it will be inclined at 45° or less to one or more of the six possible directions for the magnet-chains in every crystal. As the field is varied in the manner which I have described, the magnets in each crystal will be wheeled into parallelism, and subsequently will relapse into chains pointed in that direction (or those directions) which makes the least angle with the field. The ratio of remanence to saturation for a polycrystalline sample, resulting from this model, should then be 0.893.

By adjusting the disposable constants, Ewing's model may be made to predict not only the general shape of the I-vs.-H curve, but the values of fieldstrength and magnetization at which the first segment of the curve should pass into the second. Apparently no very pleasing agreements between experiment and theory have yet been attained in this way. Nevertheless I will show how the attempt is made; by doing so, I can at least bring out the influence of the various disposable constants upon the result.

The simplest form of Ewing's model * is composed of linear chains of elementary magnets. To analyze this it suffices to consider a system composed of two identical magnets, each so long and slender that it may be visualized as a pair of poles of equal polestrength M separated by the length L of the magnet, and both of them pivoted around their centres at points distant from one another by a spacing S which is only slightly greater than L (Fig. 10). If there is no external field, they come to an equilibrium, in which state both point in the same sense along the line of centres. If there is an applied field oblique to the line of centres, they come to an equilibrium in which both are deflected through equal angles from that line. Denote their angles of deflection by θ, the angle between the field and the line of centres by α, the fieldstrength by H, the distance between the adjacent unlike poles of the magnets by R. The distance R is equal to $(S - L)$ when θ is zero, and in general is given by the equation:

$$R^2 = L^2 + S^2 - 2LS \cos \theta. \qquad (1)$$

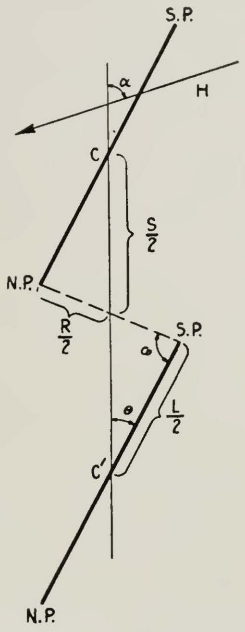


Fig. 10 — Illustrating the elementary magnet-pair of Ewing's theory.

The adjacent poles attract one another with forces M^2/R^2 directed along R, which result in torques T' upon each magnet:

$$T' = \frac{M^2 L}{R^2} \sin \varphi = \frac{M^2 LS \sin \theta}{2R^3}. \qquad (2)$$

The remote poles likewise exert forces upon the adjacent poles and upon one another, and torques upon the magnets; but it will be necessary to reduce these to relative insignificance by supposing the

* The next four pages resulted from an attempt to formulate what I take to be Ewing's objection to his own early model, which he phrases in these words: "Now it is known that in ordinary iron barely one per cent of the whole magnetism of saturation is acquired in the quasi-elastic stage before the effects of hysteresis set in. To conform to this condition the magnets of the model must have only a very narrow range of stable deflexion, and consequently they have to be set very near together with the result that in the old model their mutual control became excessive. A calculation of the force required to break up rows of pivoted magnets, of atomic dimensions, when set near enough together to satisfy the above condition, showed it to be many thousands of times greater than the force which is actually required, in iron to reach the steep part of the curve."

"clearance" $(S - L)$ between the adjacent ends of the magnets to be extremely small by comparison with S and L, and by considering only values of θ which are so small that R itself remains small by comparison with L; otherwise the equations will be hopelessly intricate, and they are more than bad enough even with this restriction. Happily the model possesses some of the required properties even when limited by this restriction.

The torque T exerted by the field H upon either magnet is given by

$$T = MHL \sin (\alpha - \theta). \tag{3}$$

The general condition for equilibrium is

$$T - T' = 0. \tag{4}$$

The special condition for "neutral" or "labile" equilibrium, i.e. for the state of incipient capsizal, is

$$d(T - T')/d\theta = 0. \tag{5}$$

The values of H and θ, obtained by solving (4) and (5) as simultaneous equations, are the fieldstrength just sufficient to produce capsizal and the angle of deflection attained just before the overturn; they are obtained as functions of the variable α, and of the constants M, L, S which are features of the model.

Solving these equations, however, is easier said than done; they prove surprisingly intractable. Only in one particular case is the solution easy: we must choose values of α so near to 90°, and suppose the clearance and consequently the deflections so small, that the cosine of $(\alpha - \theta)$ may be set equal to zero. In this case equation (5) is reduced to the form

$$dT'/d\theta = \text{const. } d(\sin \theta/R^3)/d\theta = 0, \tag{6}$$

which, if we write a for S/L, is found equivalent to

$$(a^2 + 1 - 2a \cos \theta) \cos \theta = 3a \sin^2 \theta. \tag{7}$$

Putting $a = 1 + \epsilon$—so that ϵ stands for the quantity $(S - L)/L$, which by hypothesis is small—and neglecting powers of ϵ higher than the second, we arrive at the equations:

$$\cos \theta_c = 1 - \tfrac{1}{4}\epsilon^2; \qquad \sin \theta_c = \epsilon/\sqrt{2} = (S - L)/L\sqrt{2}, \tag{8}$$

for the value θ_c of the deflection just at the verge of capsizal; and putting this into equation (4), we get

$$H_c = \frac{M}{3\sqrt{3}(S - L)^2}. \tag{9}$$

Equation (9) gives the fieldstrength H_c which effects capsizal of a pair of magnets initially transverse to the field, and having a clearance $(S - L)$ extremely small compared with their lengths L. For a chain of magnets the value given for H_c would need only to be doubled; for any number of chains lying in the plane normal to the field, that double value of H_c would remain valid. It is, we see, proportional directly to M and inversely to the square of the clearance.

Multiplying the expression given in (8) for sin θ_c by ML, we get the component along the field-direction of the moment of any magnet belonging to such a pair or to such a chain. If there were N magnets grouped in pairs or chains in the plane normal to the applied field, the magnetization I of the entire assemblage, parallel to the field-direction, would be $NML \sin \theta$. The magnetization I_c at the verge of capsizal

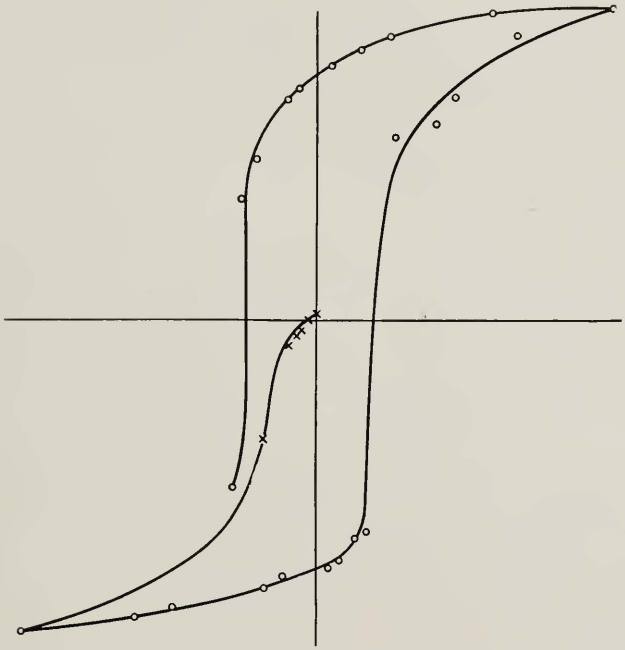


Fig. 11—Initial curve and hysteresis loop of an Ewing model composed of 24 pivoted magnets. (After F. A. Ewing.)

would be $NML \sin \theta_c$. The saturation-value of magnetization, $I_{\max}$, would be NML. Thus we arrive at the equation for I_c:

$$I_c = I_{\max} \frac{S - L}{L\sqrt{2}}, \tag{10}$$

beside which we may place equation (9), reshaped and with allowance for the doubling:

$$H_c = \frac{2I_{\max}}{3\sqrt{3}NL(S-L)^2}. \tag{11}$$

It is now obvious that, in the case of a material for which $I_{\max}$ is known, we have apparently three disposable constants N, L, and $(S - L)$. However, the ratio of $(S - L)$ to L must be a very small fraction; otherwise the assumptions from which the equations were deduced would not be valid. This ratio is determined by the ratio $I_c/I_{\max}$. If we take for I_c the value of magnetization somewhere near the division between the first and second segments of the initial curve, we find $I_c/I_{\max} = .01$ for soft iron (I quote Ewing) or about .05 for the permalloy of which the curve is exhibited in Fig. 1. Now we have the ratio of $(S - L)$ to L fixed, and ostensibly two disposable constants left. However, if we assume that each elementary magnet is an atom and each atom an elementary magnet, both of these are determined by the crystal lattice of the metal. Nothing remains adjustable; a definite value is imposed by the theory upon H_c. This value is enormously too great.

It is clear that the situation could be saved by dropping the assumption that every atom is a magnet, so that the constants N and L might again become freely disposable. Ewing proposed another way of escape—a modification of the model involving the introduction of a fourth constant. He invented a system composed of three magnets with their centres in a line, the two outer ones fixed and pointing in opposite senses along the line of centres, the middle one free to revolve. The polestrengths of the outer magnets, M' and M'', are supposed to differ slightly; then, when no outside force is acting, the middle one comes to an equilibrium in which it points in the same sense as the stronger of its neighbors. When a field is applied in a direction inclined at α to the line of centres and steadily increased, capsizal occurs at a certain value of fieldstrength H and the corresponding value of deflection θ. When the clearances are small and α is very nearly 90°, the equation for θ is equation (8) with an unimportant change in numerical factor; while the equation for H is changed, in that $(M' - M'')$ now stands in the place of M. This is the new constant introduced into the model.

Ewing supposed that the pivoted magnet of his model might be the analogue of an internal electron-orbit of the iron atom, while the fixed neighbors might correspond to external distributions of whirling electrons, in the periphery of the same atom or in neighboring

ones. This notion is endangered by the discovery that a single crystal of iron displays only a slight degree of hysteresis, much less than a polycrystalline mass—a discovery which likewise weakens the force of calculations of remanence based upon the assumption of a cubic lattice, such as I gave earlier. In fact, it seems quite probable that in the course of assimilating the newly-acquired data concerning single crystals, all of the numerical agreements hitherto derived from Ewing's and other theories of ferromagnetism may be swept away.

Nevertheless the basis of Ewing's theory is likely to persist; for it has two great advantages which are nearly independent of numerical agreements. Hysteresis is derived from an atom-model in which nothing of the nature of hysteresis is introduced by postulate; and the general effect of mechanical and electrical jerkings, bumpings and jogglings is explained in a way which seems most natural and plausible to our mechanical instincts. As for the first point: to explain hysteresis by the mutual interactions of magnets which in themselves are constant and do not possess it is so eminently satisfactory a solution that any theory or model in which hysteresis is introduced *ab initio* or derived from some extra assumption will start under a great handicap. As for the second: to take one illustration, it is well known that a demagnetized piece of iron exposed to a weak field, and endowed thereby with the moderate magnetization corresponding to some point or other on the first segment of the initial curve, becomes enormously more intensely magnetized when it is jolted or jerked. Visualized by Ewing's model, this seems the most natural thing in the world: the elementary magnets which were on or close to the verge of capsizing are pushed over that verge by the mechanical shock. Equally natural seem the annulment of the residual magnetism of a piece of iron, by mechanical shocks and jerks; the like effect of rapidly-alternating magnetic fields; the tendency of a current along an iron wire to favor magnetization of the wire; and the fact to which I alluded earlier, that in a very strong rotating magnetic field a piece of iron does not grow hot, and consequently there can be no hysteresis-loss. This last-named feature may be visualized by supposing that the chains, having been once completely broken up, do not form again as the magnets are whirled round and round. It seems natural also to expect that as the temperature is raised, the chains will be broken up by thermal agitation, and the reversible first segment of the initial curve will mount more sharply and continue longer. In trying to deal with the effect of temperature, however, we soon reach the limits to which Ewing's theory can be forced; and another method of attacking the problem of ferromagnetism recommends itself.

Weiss' Theory

There is another theory of magnetization, built upon an entirely different basis from Ewing's—a basis involving the notion and in fact the definition of temperature. To import temperature into theories of magnetism is clearly most desirable, considering how great is the influence of that variable upon the I-vs.-H curves; an influence so great, indeed, that when a sample of any ferromagnetic substance is made sufficiently hot, all the distinctive features of ferromagnetism depart from it. In developing Ewing's model, it is easy to say that as the temperature is raised the little magnets are more vigorously agitated, the bonds which are responsible for remanence and coercivity are more frequently ruptured; but such statements, though plausible, lack precision and hold out no promise of numerical agreements between theory and experience. That being the case, it seems unreasonable to expect numerical agreements from a theory offering a much less definite and specific picture of the interior of a ferromagnetic body than even Ewing's. Such agreements, nevertheless, emerge from the theory of Langevin and Weiss.

Langevin took as his point of departure the theory of temperature developed by the great savants Maxwell and Boltzmann (the same from which, by the way, the quantum-theory arose through the modifications made by Plânck). To introduce as much, or as little, of this theory of temperature as is required for our present purpose, we envisage a sample of oxygen gas, N molecules per unit volume, in thermal equilibrium at absolute temperature T. Let each molecule be visualized as a rigid body of mass m, having three principal axes of rotation and corresponding moments of inertia I_1, I_2, I_3. The molecules are darting to and fro, with translatory velocities which may be specified by giving the three components u, v, w of each in some coordinate-frame. They are likewise revolving, with angular velocities which may be specified by giving the three components r, s, t of each along the principal axes of the molecule in question. The kinetic energy of the molecule is given by

$$\begin{aligned} K &= \tfrac{1}{2}mu^2 + \tfrac{1}{2}mv^2 + \tfrac{1}{2}mw^2 + \tfrac{1}{2}I_1r^2 + \tfrac{1}{2}I_2s^2 + \tfrac{1}{2}I_3t^2 \\ &= K_u + K_v + K_w + K_r + K_s + K_t, \end{aligned} \tag{1}$$

each of which six terms may be regarded as the kinetic energy associated with the variable which its subscript denotes. We will further suppose that each molecule is a magnet of moment M. When the gas is pervaded by a magnetic field H, each molecular magnet has a potential energy V_θ given in terms of the variable θ, the angle which its axis makes with the field, by the equation

$$V_\theta = -MH\cos\theta. \tag{2}$$

I propose now to show that Langevin's theory of magnetization is obtained by applying to the potential-energy term V_θ the same mode of reasoning as is customarily and familiarly applied to the kinetic-energy terms $K_u \cdots K_t$.

It is well known that the average kinetic energy of translation, the average of the sum of the terms K_u and K_v and K_w, taken over all the molecules of a gas of absolute temperature T, is proportional to T; it is, in fact, given by the equation

$$\overline{K_u + K_v + K_w} = \frac{3}{2}kT, \tag{3}$$

in which k stands for the ratio of the gas-constant R to the Loschmidt number N_0 (number of molecules per gramme-molecule).* The average of each of these three terms separately is equal to $\frac{1}{2}kT$; and this result was generalized by Maxwell and by Boltzmann to the three rotational terms in the expression for K, so that

$$\bar{K}_u = \bar{K}_v = \bar{K}_w = \bar{K}_r = \bar{K}_s = \bar{K}_t = \tfrac{1}{2}kT. \tag{4}$$

We go one step further in the analysis of the motion of the molecules. Consider the distribution-function for any one of these six variables, u for instance; it is given by Maxwell's formula:

$$dN = NC_u \exp\left(-\tfrac{1}{2}mu^2/kT\right)du, \tag{5}$$

in which dN stands for the number of molecules (among the N molecules occupying unit volume) for which the velocity-components along the x-axis lie between the values u and $u + du$. The constant C_u is so adjusted that the integral of dN over the entire range of values of u shall be equal to N; on being computed it turns out to be $\sqrt{m/2\pi kT}$. The quantity $\frac{1}{2}mu^2$ is the one hitherto designated as K_u. For the distribution-function with respect to u, which is the coefficient of du in (5), and may be denoted by $F(u)$, we therefore have:

$$F(u) = N\cdot\sqrt{m/2\pi kT}\cdot \exp.\ (-K_u/kT) \tag{6}$$

* The primitive way of deriving (3), reproduced in all elementary texts, is as follows: Imagine a cubical vessel one cm. along each edge containing N molecules; suppose that $N/3$ molecules are moving in lines parallel to each edge, with uniform speed v; each face is then struck with $Nv/6$ impacts per second, and in each impact an amount of momentum $2mv$ is communicated to the face, so that the average pressure upon the surface is $p = Nmv^2/3$. According to the well-known gas-law, $p = \rho RT/M$ (ρ standing for the density, M for the molecular weight of the gas); hence $Nmv^2/3 = \rho RT/M$, and recalling that $\rho = Nm$ and that $M/m = N_0$ and that $\frac{1}{2}mv^2$ is the kinetic energy K of a molecule, we have $K = 3RT/2N_0 = 3kT/2$. The same result is reached by more sophisticated methods of averaging.

and the distribution-functions with respect to v, w, r, s, and t differ only by the substitution of the appropriate kinetic-energy term for K_u, and (if necessary) of I_1 or I_2 or I_3 for m.

For the distribution-function with respect to θ, we shall write an equation copied after (5), as follows:

$$\begin{aligned} dN &= NC_\theta \, exp \, (-V_\theta/kT) \sin \theta d\theta \\ &= NC_\theta \, exp \, (MII \cos \theta/kT) \sin \theta d\theta. \end{aligned} \tag{7}$$

The constant C_θ is to be so adjusted that the integral of dN over the entire range of values of θ (which extends from 0 to π) shall be equal to N. It turns out that

$$C_\theta = a/(e^a - e^{-a}) = a/2 \sinh a \tag{8}$$

in terms of the parameter

$$a = MII/kT, \tag{9}$$

which we shall use often enough to justify the special symbol for it. The factor $\sin \theta$ in equation (7) requires comment. Imagine all the molecular magnets brought together at a point P, and their axes prolonged until these intersect a sphere of unit radius traced around P as center. The locus, upon this sphere, of the points of intersection of lines associated with magnets inclined at angles between θ and $d\theta$ to the field is a belt or collar of area $2\pi \sin \theta d\theta$. There are dN of these points, and they are distributed over this belt with surface-density $dN/2\pi \sin \theta d\theta$. By making dN proportional to the product of $\sin \theta$ into an exponential function, we make that surface-density, which is the density-in-solid-angle of the directions of the magnetic axes, proportional to the exponential function itself; and this is what is done.

We proceed to calculate the net magnetic moment of the assemblage of N molecular magnets. Resolving the moment of each, we find $M \cos \theta$ for its component parallel to the field-direction (with the perpendicular component we are not concerned, since the average of its values for all the molecules is obviously zero). Summing the values of these parallel components for all the molecules, we have:

$$I = \int_0^\pi M \cos \theta dN, \tag{10}$$

the symbol I being used for the sum of the parallel components, since this sum is precisely the intensity of magnetization per unit volume defined near the beginning of this article. Remembering (7) and (8), and performing the integration, we arrive at

$$\frac{I}{I_{\max.}} = \coth a - \frac{1}{a} = L(a), \tag{11}$$

the symbol $I_{\max.}$ being used for NM, the total magnetic moment which the assemblage of N molecules would have if they were all directed perfectly parallel to the field.

This function $L(a)$ is represented by the curve of Fig. 12, which departs from the origin with slope $a/3 = MH/3kT$, and bends over

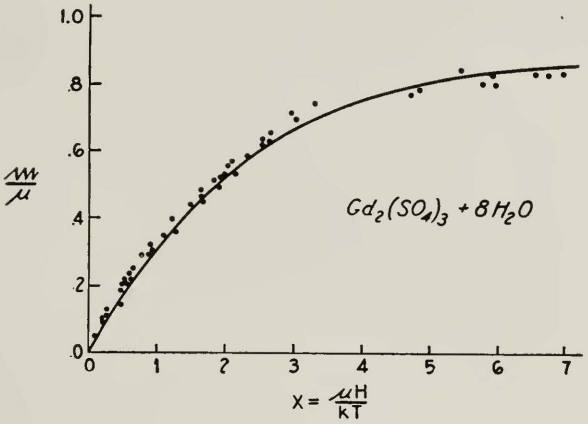


Fig. 12—The I-vs.-a curve of Langevin's theory of paramagnetism, and the data for gadolinium sulphate. (After P. Debye.)

toward its asymptote $L = 1$ without passing through any point of inflection. It has thus a resemblance to the initial curve; but one must not be misled by this, for Langevin's theory is not a theory of ferromagnetism. It is based on assumptions appropriate to a gas, and gases are not ferromagnetic; it gives no account of remanence, and remanence is an essential feature of ferromagnetic bodies. With a gas of which the molecules are permanent magnets, we should expect I to vary with H in the manner indicated by the curve.

Now as a matter of fact, in oxygen and other paramagnetic gases I is apparently proportional to H, up to the greatest fieldstrengths which can be applied:

$$I = \sigma_0 H. \tag{12}$$

This however does not necessarily mean that equation (11) is not valid; it may mean simply that the greatest available fields (some tens of thousands of gauss) are not great enough to pass beyond the sensibly-straight initial portion of the curve. If so, then

$$\sigma_0 = I_{\max.} M/3kT = NM^2/3kT \tag{13}$$

and σ_0, the susceptibility of the material, should vary inversely as the absolute temperature. This, as Curie found, is true for the paramagnetic gases. It is true also for a number of salts in dilute solutions, and even for a certain number of solid substances, although for these the underlying assumptions would scarcely be expected to remain valid; one has the feeling that the data are left floating in the air by the withdrawal of the logical basis for the theory with which they agree.

Suppose nevertheless that the theory remains valid; then, for any substance of which the susceptibility σ_0 varies inversely as T, one can calculate the moment M of its molecular magnets from (13); for k is a known constant, and N is knowable at least when one is dealing with a gas of known density or a solution of known concentration (with solids there may be doubt as to the number of atoms grouped together to form an "elementary magnet"). Multitudes of such values have been computed; their orders-of-magnitude are 10^{-18} to 10^{-20}. Commonly they are expressed as multiples of a certain unit, the "Weiss magneton," which is equal to $1126/N_0$ or about $1.858 \cdot 10^{-21}$. Many of them are nearly integer multiples of this unit.*

On taking any observed value of M, and multiplying it by the corresponding value of N to obtain the "theoretical" value of $I_{\max.}$ for the substance in question, we find that as a rule this last is so much larger than the highest value of I attained with practicable fields that there is no contradiction between the theory and the fact that I is sensibly proportional to H all through the feasible range of fieldstrengths. There is only one substance (gadolinium sulphate) for which $I_{\max.}$ can be approached and this only at extremely low temperatures, below 5° absolute; in Fig. 12 the data are displayed; it is evident that the Langevin curve, drawn with the initial slope best suiting the points near the origin, fits fairly well to all the other points.

I pass now to the assumption whereby Weiss so extended Langevin's theory that it became competent to describe not only these simplest cases of paramagnetism in which $1/\sigma$ is proportional to T, but also the much more numerous cases of paramagnetic substances conforming to a more general law, and certain aspects of ferromagnetism also.

Formally the extension amounts to this, that in the expression for the parameter a which figures in equation (9), the fieldstrength H is replaced by a linear function of H and I:

* To enter into the long and fiercely debated questions about the meaning and even the reality of the Weiss magneton would lead me too far afield; but it is so frequently used as a unit in stating data of experiment that one must know at least its value.

$$a = M(H + nI)/kT, \tag{14}$$

which is transported bodily into the function $L(a)$ of equation (11). This is a very abstract way of putting the fact; but the more concrete ways have not been satisfying. One may say that the true field acting within the material is not H, but $(H + nI)$—that the actual though unverifiable field acting at any point in the inaccessible interior of the magnet is the sum of the field H_e due to objects in the external world, and the field H_i due to the "demagnetizing effect of the poles," *and an additional term proportional to the intensity of magnetization at the point in question.* The suggestion of Weiss, then, is tantamount to making a new assumption concerning this tantalizing internal field. The natural next step is, to visualize or explain the agent of the extra force, the "molecular field" as Weiss calls it; that is the step which no one has yet succeeded in making, not at least with general assent.

Making the expression in (14) the argument of $L(a)$, we see that the fundamental equation (11) now has the variable I in both its members, and must be solved for I. The resulting function is one of the infinitely many which have neither names nor well-known features, and most of those who write on this subject recommend the high-school expedient of plotting the curves representing the two functions

$$I = (kT/nM)a - H/n, \tag{15a}$$

$$I = I_{\max.}L(a) = NML(a), \tag{15b}$$

in a coordinate-plane with I as ordinate and a as abscissa, and looking for the point or points of intersections between the two curves. These, which I shall designate for a few paragraphs as "the line" and "the

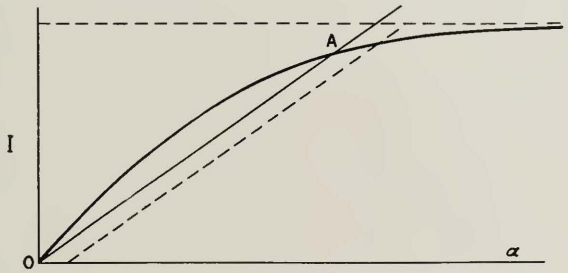


Fig. 13—The "curve" and the "line" of the Langevin-Weiss theory of ferromagnetism

curve," are shown in Fig. 13. It is easy to see that, when T is held constant and H increases, the line slides from left to right and the intersection-point mounts along the curve; when H is held constant

and T increases, the line wheels counterclockwise around the point where it cuts the axis of abscissæ, and its intersection with the curve descends along the latter.

There is a valuable approximation, which is more nearly valid, the higher the temperature and the lower the field. At the origin, the tangent to the curve ascends with slope $NM/3$ (as I have said) and so long as a is not greater than unity, the ordinate of the curve agrees within six per cent with the ordinate of the tangent. If H is so small and T so great that the crossing of the line and the curve occurs within this range, the problem of locating it may be translated for all practical purposes into the algebraic problem of solving the simultaneous equations

$$I = (kT/nM)a - H/n, \qquad I = NMa/3, \tag{16}$$

achieving which, one obtains

$$I/H = \sigma = \frac{C}{T - \Theta}, \qquad \begin{cases} C = NM^2/3k \\ \Theta = nC. \end{cases} \tag{17}$$

The susceptibility of an assemblage of elementary magnets, in thermal equilibrium under the influence of an applied field on which there is superposed an extra field proportional to the magnetization of the assemblage, should then depend on temperature approximately

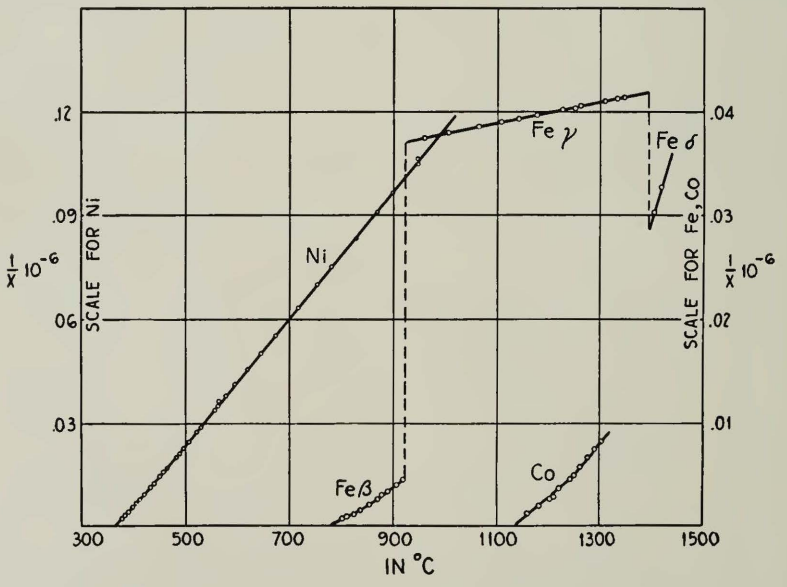


Fig. 14—Susceptibility-vs.-temperature curves for iron, cobalt, and nickel above their respective Curie-points. (After P. Debye.)

according to (17); the approximation being closer, the higher the temperature and the lower the field.*

Now there is a very large class of paramagnetic substances of which the susceptibilities at low fieldstrengths conform, over wide ranges of temperature, to equations like (17); and what renders the theory important for our present purposes is, that the ferromagnetic metals at high temperatures enter into that class. To make the test for any substance it is best to plot $1/\sigma$ as a function of absolute (or Centigrade) temperature. On doing this for nickel beyond the Curie-point (near 360° C.) one finds a curve which at first is somewhat bent, but beyond 410° passes into a beautiful straight line which continues undeflected to 900°. This line is shown in Fig. 14, together with data for iron beyond its Curie-point at 775°; among these, the points for temperatures between 920° and 1395° lie along a straight line which is sharply broken off at each end of that interval, being followed beyond 1395° by what seems to be the beginnings of an entirely different line, and preceded before 920° by a series of points which are well fitted by a pair of straight lines connected with each other at 828°. The data for cobalt beyond its Curie-point at 1130° likewise conform to a pair of connecting straight lines.

For each of these straight lines one may compute the values of the constants called C and Θ; and from these, if one accepts the theory, the values of the moment M of the elementary magnets and the coefficient n of the postulated extra force. In calculating M it is necessary to make an assumption about the number of elementary magnets per unit volume of the metal; assuming that there are as many such as there are atoms, and expressing M in Weiss magnetons, Weiss obtained the values 20.9, 17.4, 28.2 and 7.05 for the four straight lines of iron (in order of increasing temperature); 15.9 and 14.55 for those of cobalt; 8 for the solitary straight line of nickel. All these are of the orders of magnitude customarily found in dealing with paramagnetic gases and salts and solutions. The corresponding values of Θ are 1047, 1063, $-$ 1340, 1543; 1404, 1422; and 645. The corresponding values of n (which is the quotient of Θ by C) are of the order of several thousands. The postulated extra field must therefore be supposed enormously greater than the field H, and even the induction is quite insignificant by comparison with it. In one of these cases (and in many others among the paramagnetic salts,

* I should state that formulae of the same type as (17) may be derived without assuming that there is a molecular field, provided that we suppose that the distribution-in-energy of the atoms in thermal equilibrium is governed not by the equipartition-law, but by a quantum-law involving a zero point energy.

and in that of liquid oxygen) it must even be supposed *antiparallel* to the field H; for Θ is negative, and consequently so is n. Necessities such as these make it hard to accommodate the "molecular field" to what is known or conjectured about the interior of solid bodies.

Since it is necessary to assign several distinct values to the coefficient M in order to explain the behavior of iron over various ranges of temperature, one cannot maintain that the iron atom possesses a constant and characteristic magnetic moment which is the source of ferromagnetism. Any such notion, of course, would have been destroyed by facts already mentioned; but it is useful to know these in addition. Changes in M sometimes coincide with great and striking changes in the condition of the metal; at 920° iron exchanges its body-centered lattice (spacing 2.88A) for a face-centered lattice (spacing 3.60A) which it retains as the temperature rises until 1395° is attained, whereupon it returns to the body-centered-cubic arrangement. These alterations in atom-lattice are attended by changes in the physical properties of the metal, so great that three separate "modifications" of iron were distinguished and named before ever the atom-lattices were known or suspected: β-iron normally existing from the Curie-point to 920°, γ-iron from 920° to 1395°, δ-iron from 1395° upward. By certain processes these modifications may be enabled to survive in temperature-ranges not appropriate to them, but that is too long a story for these pages. Changes in M sometimes occur quite unaccompanied, so far as can be made out, by changes in atom-lattice or other physical features. The variation occurring at 828° in iron is of this type, and so is a mysterious change in nickel which in occasional samples brings about values of M near 9, instead of the usual 8 Weiss magnetons.

We turn to residual magnetism, on its explanation of which every theory of ferromagnetism must stand or fall. It is the supreme merit of the theory of Weiss that residual magnetism figures as a property which substances paramagnetic at high temperatures naturally and gradually acquire, when they are cooled below a certain critical point. We shall see this best by returning to Fig. 13. Begin by imagining the line corresponding to a particular pair of values of H and T; leave T constant, reduce H steadily to zero; the intersection of curve and line slides down the curve, *reaching the origin if the slope of the line is greater, stopping short of the origin if the slope of the line is less, than the slope of the tangent to the curve at the origin.*

The slope of the line is kT/nM; the slope of the tangent is $NM/3$; the critical condition is, that these be equal, and this occurs when

$$T = nNM^2/3k = \Theta,$$

i.e., when the temperature assumes the value of that constant Θ which previously entered into our equations. If T is greater than Θ, there should be no residual magnetism. If T is adjusted to be equal to Θ and then reduced gradually to zero absolute, the residual magnetization given from the theory—the ordinate of the point where the curve is intersected by the line of slope kT/nM passing through the origin—increases continuously from zero to its limiting value NM, following the curve traced in Fig. 9. That is the central idea of Weiss' theory of ferromagnetism.

The first of the predictions from the theory which can be put to test is the equality between the temperature at which residual magnetism disappears—the Curie-point—and the constant Θ in the equation (17) for the paramagnetism of the substance beyond the Curie-point. For nickel, the agreement is good: 633° against 645° absolute. For cobalt and for iron, the first short straight line out of the sets of two and four respectively, which are given for these metals in Fig. 14, is so adjusted that Θ agrees perfectly with the Curie-point; its aptness to the plotted data supports the theory.

The next question to be asked is whether the curve of Fig. 9 corresponds to experience. In analyzing this question, one makes the discomfiting discovery that the quantity which was defined as residual magnetization in the theory cannot be identified with the quantity defined as remanence in describing the experimental hysteresis-loops. This results from an imperfection, or at least an incompleteness, in the theory. There is nothing in it to account for the initial curve; there is nothing to account for the gradual increase in I produced by applying a gradually-increasing field to an initially-demagnetized piece of iron, and in fact there is nothing to account for the existence of demagnetized pieces of iron at all—every block of iron at a temperature below Θ should possess, whenever it is not under the influence of an external field, the residual magnetization calculated from the intersection-point of the curve $NML(a)$ and the line of slope kT/nM which passes through the origin.

On grasping this situation, one is likely to feel that the theory has collapsed. The situation can be saved, however, by supposing that the "demagnetized" metal subdivides itself into a vast number of little regions, zones, or filaments, each of which possesses the full residual magnetism of the theory, while in direction their magnetic moments are oriented quite at random. It is not possible to identify these with individual crystals, nor with any other discernible granulations of the metal. Perhaps they are to be identified with the chains of elementary magnets once postulated by Ewing; it would be grati-

fying to make a connection between the theories of Ewing and Weiss. Perhaps they are the units from which arise the separate clicks which constitute the Barkhausen effect. As for the initial curve, attempts must be made to explain it either by supposing that the increasing field wheels the magnetic moments of the several zones gradually into parallelism with itself, or—what is more probable—that the field abruptly reverses, one after the other, all the magnetic moments which initially are inclined to it at angles superior to 90°. By suitably combining these two images, one may copy almost any possible form of initial curve. I cannot enter into these questions, except to answer as far as possible what I designated as the second question to be asked in testing the theory: what observable quantity is to be compared with the "residual magnetization" predicted from the theory of Weiss?

A piece of iron brought to saturation by a large applied field is supposed to consist of these magnetized zones, their moments all directed either parallel or at least at inclinations of less than 90° to the field. The applied fieldstrength should elevate the magnetization of each to a value somewhat greater (corresponding to an intersection-point somewhat farther along the "curve" of Fig. 13) than the predicted "residual magnetization"; but the values of n and I and hence their product are so enormous that the addition is only slight. The saturation intensity of magnetization of the iron, $I_{\max.}$, should then be very nearly equal to the predicted residual magnetism, if all the magnetic moments are parallel; or to one half of the predicted residual magnetization, if the magnetic moments are distributed at random over the directions inclined at less than 90° to the applied field. In the former case, the variation of $I_{\max.}$ with T should follow the curve of Fig. 9; in the latter case, a curve of the same form. The actual observations upon iron, nickel, cobalt and magnetite are shown in that figure, and the reader may judge of the agreement for himself.

Comparison of Ewing's Theory with that of Langevin and Weiss

At first glance the Ewing model and the Langevin-Weiss conception of a ferromagnetic substance seem extremely different; contradictory, in fact. In Ewing's view, the perpetual effort of the applied field to align the elementary magnets is hindered by the forces which these exert on one another. In Langevin's theory, the antaonigst of the applied field is the thermal agitation. Now Langevin's theory is competent to deal with paramagnetic substances which are difficult to magnetize, but not with iron and the like which are strongly mag-

netized by weak fields. This means that the thermal agitation is too strong an antagonist to the applied field. Weiss therefore provided the latter with a powerful ally, in the form of an intense molecular field parallel to it and proportional to the magnetization. The applied field and its ally together are able to overpower the thermal agitation and bring about saturation in cold iron. Now to say "molecular field" is merely to use a different phrase for "influence of the atoms on one another." In the theory of Weiss, this influence of the atoms on one another helps the field to align them; in Ewing's theory, it hinders the field. How do away with this arrant contradiction?

Perhaps a partial union may be effected, in this wise. According to Langevin and Weiss, a piece of cold iron consists of a multitude of small zones or regions of atom-groups, each magnetized to a high degree, their directions of magnetization dispersed at random; an applied field acts primarily by wheeling these magnetizations into line. According to Ewing, a piece of cold iron consists of a multitude of chains or pairs of systems of elementary magnets, which an applied field upsets, perhaps only to re-weld them anew into more favourably oriented chains. Weiss deals with the state of affairs inside the atom-groups; Ewing deals with the effect of the applied field in breaking up and rebuilding the atom-groups. Might one say that Weiss explains the conditions, under which the elementary magnets form themselves into groups or chains such as Ewing preassumed? that Ewing describes the action of the external field upon these groups, an action which Weiss left imprecise? so that the two theories, when properly revised, will complement each other? It seems possible. At all events, each of the theories has so many successes to its credit, that there can be no thought of discarding either for the sake of the other. Those who are weary of trying to reconcile waves and quanta might refresh themselves by reflecting on this problem.

McKeehan's Theory

In the theory of McKeehan, magnetostriction is promoted to the dominant role. The distortion which a metal undergoes when it is magnetized is held responsible for hysteresis, and for the fact that the rise of the *I*-vs.-*II* curve is gradual, not sudden. This view was suggested by the fact which I have mentioned already: that, in the series of the permalloys, the permeability reaches a surprisingly high maximum value and the hysteresis a surprisingly low minimum value, just at that alloy of which the magnetostriction is indetectably small until saturation is nearly attained—the alloy intermediate between

those which lengthen and those which shorten when magnetization commences. The alloy which is most rapidly magnetized when the field is gradually increased from zero, and which dissipates the smallest amount of energy when the field is varied in cyclic fashion, is also precisely the one which suffers the least deformation. From this McKeehan drew the inference, that were it not for the deformation inseparable from the act of magnetizing, the initial curve for every metal would rise swiftly from the origin to saturation, and the sides of the hysteresis-loop would fall together.

D. The Atomic Magnets

Had I announced at the beginning of this article that some sixty pages would be spent over the data of ferromagnetism and the theories of the influence of elementary or atomic magnets on one another, and only a few closing paragraphs over the atoms which are supposedly responsible for the whole affair, the plan might have seemed most ill-adjusted to the relative interest of these divisions. Now, I hope, it will seem less perverse. The truth is, that we do not understand ferromagnetism well enough to draw from it any reliable conclusions concerning the atomic magnets. For these, we must consult the behavior of paramagnetic substances, and line-spectra, and the observations of Gerlach and Stern and their followers upon streams of atoms flying through magnetic fields.

In the apparatus of Gerlach and Stern, the atoms are probably as nearly free from mutual forces as atoms in the laboratory can ever be; having issued from a small hole in the wall of a furnace full of hot vapor, they rush swiftly across a high vacuum while they are being examined. In the mapping of absorption-spectra, the atoms are those of a rarefied gas, and are "free" in the sense in which atoms of gases are free—that is to say, they are influenced only by those agencies which establish and maintain thermal equilibrium, agencies which we commonly conceive as short, sharp collisions between atom and atom. Some paramagnetic gases behave toward an applied magnetic field as though their molecules, some salt-solutions behave as though their ions, were magnets of fixed permanent moment on which the field can act, but otherwise were free in the foregoing sense. Other gases and salt-solutions behave as though their molecules or ions were permanent magnets, influenced by the applied magnetic field and by an extra field proportional to the magnetization of the assemblage, and otherwise free except for the agencies which establish thermal equilibrium and maintain it.

In all the foregoing cases of atoms or molecules or ions enjoying

variously close approximations to perfect freedom, the theories are good enough to make it possible to bring about quantitative agreement between theory and experiment, simply by choosing appropriate values for the magnetic moments of these particles. The values so determined nearly always lie between 10^{-18} and 10^{-20} C.G.S. units.

Ferromagnetic substances are solids, and we need not be surprised that the mutual influence of the atoms becomes so great as to make the task of devising a theory much more difficult. Ewing, it is true, did show that elementary magnets of a particular shape and crowded close together would form systems displaying the peculiar features (hysteresis, and a crooked magnetization-curve) of ferromagnetics. Weiss did show that atomic magnets, subject to the agencies which bring about thermal equilibrium and maintain it, and in addition to a field proportional to the magnetization of the assemblage and enormously great, would form systems displaying residual magnetism below a certain temperature, and paramagnetic above. Dazzling as these achievements are, the theories are not so good that they can be brought into complete accord with the data, simply by choosing appropriate values for the moments of the imagined elementary magnets.

Can we at least assign a value of the order familiar among paramagnetics, 10^{-19} for instance, to the magnetic moment of (say) the iron atom—that is to say, the atoms of a piece of solid pure iron, since iron is not in all conditions ferromagnetic—without definitely contradicting any fact of experience? Probably we can. In fact, the saturation-values of the magnetizations of iron, nickel, and cobalt support this idea. If saturation signifies that all the atomic magnets are parallel, then the magnetic moment of each must be the quotient of $I_{max.}$ by the number of atoms in unit volume; at all events, the magnetic moment of the atom cannot be less than the quotient, by that number of atoms, of the highest value of I ever observed. Now the highest values of I are observed at the lowest temperatures; extrapolating from the data (shown in Figure 9) to zero absolute, Weiss obtained values of the quotient which are indeed of the order 10^{-19}—eleven "magnetons" for iron and three for nickel, and probably eight for cobalt. This concordance with the values of magnetic moment to which we are accustomed among free atoms is evidently important. However, as Ewing found, we cannot take the natural next step of supposing that each atom is a long slender magnet having its ends very close to the ends of the adjacent magnets; for then the I-vs.-H curve of the assemblage would not agree with the initial curves observed in practice.

Everyone now agrees with the idea, proposed more than a century ago by Ampère, that atoms are magnets because of the circulating charges which they contain. The estimates of atomic moments deduced from line-spectra are based on this assumption, and the verified correctness of these estimates sustains it. Now, if a magnetic atom is a whirl of electricity, it possesses angular momentum as well as magnetic moment. If so, the process of magnetizing an iron wire involves the bringing-into-parallelism of myriads of spinning-tops, of which the angular momenta when all aligned combine into a respectable sum. If this goes on inside a wire during magnetization, there should be a "recoil" somewhere, comparable to the recoil of a gun when a shell is fired—the suspension of the wire should receive an opposite angular momentum, experience a torque. Conversely, the process of twisting an unmagnetized wire should impress a lateral torque upon myriads of spinning-tops of which the axes point in directions scattered at random; each of these should be urged to set itself more nearly parallel to the axis of the twist, which is the axis of the wire; and the twisting should therefore magnetize the wire.

Both of these effects, which jointly are called the "gyromagnetic effect," have been detected and measured. From the measurements (thus far performed upon iron, nickel, cobalt, magnetite and a Heusler alloy), it results that the ratio of the angular momentum P to the magnetic moment M of an elementary magnet conforms to the equation:

$$P/M = mc/e,$$

in which m stands for the mass of the electron and e/c for its charge measured in electromagnetic units. *This is the value which would be expected for the ratio, if the elementary magnet is an electron spinning upon itself.*

Now there are weighty reasons for supposing that the conception of a "spinning electron," possessing a fixed characteristic angular momentum and a permanent magnetic moment e/mc times as great, may be what is required to complete the theory of line-spectra of free atoms which Bohr began. The gyromagnetic effect of the ferromagnetic solids therefore indicates that the elementary magnets scattered through these are the same as the elementary magnets located in free atoms—they are electrons, or groups of electrons suitably linked together. The test cannot be made upon paramagnetics, for they cannot be (or at least have not yet been) strongly enough magnetized. Ferromagnetic substances are the only ones which in a feasible field acquire so great a magnetization that the

recoil from the spinning electrons is detectable. This seems to be as yet the only contribution of ferromagnetism to contemporary atomic theory.

Yet even if we take it for settled that the elementary magnets within the atoms of a solid piece of iron are spinning electrons, the real problem of ferromagnetism remains unsolved. If the elementary magnets in iron are just like those in all other atoms, how does it happen that iron and two other elements alone may be ferromagnetic? that even iron may cease to be ferromagnetic, if mixed with a little manganese? that manganese and copper and aluminium can become ferromagnetic when and only when alloyed together? Since apparently we must not suppose that each atom of iron is distinguished from all those of never-ferromagnetic substances through having a peculiar kind of magnet inside it, we must suppose that something strange in the arrangement of the electron-magnets of the iron atom permits it to be so distorted, and so to distort its neighbors, that on occasion its neighbors and itself jointly develop ferromagnetism. There is something extraordinary about the systems of 26 and 27 and 28 electrons about a nucleus, which iron and nickel and cobalt atoms are. Their individual electrons are not unique; by themselves, or as ions in a solution, they show nothing unique; but they turn into something unique when they are rightly compounded together into a solid. The theories of ferromagnetism and the gyromagnetic effect have limited without solving the fundamental problem of ferromagnetism: what is it that makes the difference between the ferromagnetic substances, and all the rest?

Acknowledgments and References

The foregoing article is based largely upon the books of J. A. Ewing (*Magnetic Induction in Iron and Other Metals;* Electrician, 1900), P. Weiss and E. Foex (*Le Magnétisme;* Colin, 1926) and E. C. Stoner (*Magnetism and Atomic Structure;* Methuen, 1926); the articles by S. Bidwell in the eleventh edition of the Encyclopædia Britannica, by P. Debye in volume 6 of the *Handbuch der Radiologie,* by E. Gumlich and R. Gans in *Die Kultur der Gegenwart,* by K. Honda in the *Dictionary of Applied Physics;* and the articles of L. W. McKeehan on ferromagnetism (*Journ. Franklin Inst.* **197**, pp. 583–602, 757–786; 1924), magnetostriction (*ibid.* **202**, pp. 737–773; 1926) and the permalloys (*Phys. Rev.* (*2*) **28**, pp. 146–166; 1926, and others there cited).

Some of the very recent papers upon magnetization of single crystals are those of W. L. Webster (*Proc. Roy. Soc.* **A107**, pp. 496–509; 1925); K. Honda and S. Kaya (*Tohoku Univ. Sci. Rep.* **15**, pp. 721–753; 1926); W. Gerlach (*Z.S. f. Phys.* **38**, pp. 828–840; 1926). For magnetostriction of single crystals, see W. L. Webster (*Proc. Roy. Soc.* **A109**, pp. 570–584; 1925) and K. Honda and Y. Mashiyama (*Tohoku Univ. Sci. Rep.* **15**, pp. 755–776; 1926). For the data concerning permalloys see,

in addition to the papers already cited, that of H. D. Arnold and G. W. Elmen (*Journ. Franklin Inst.* **195**, pp. 621–632; 1923). The gyromagnetic effect, the data and the theories of paramagnetic substances, and diamagnetism are treated very fully in the above-cited book of Stoner; paramagnetism, and the interesting and important magneto-caloric effects which I had not space to discuss, in the book of Weiss and Foex; diamagnetism in a late article by E. S. Bieler (*Journ. Franklin Inst.* **203**, pp. 211–242; 1927).

I am much indebted for the comments and counsel and information abundantly given by Dr. L. W. McKeehan during the preparation of this article, and for the opportunity of using several cuts prepared for his article "Ferromagnetism"; to Dr. O. E. Buckley for reading and commenting upon a great part of the manuscript; and to Mr. L. A. MacColl for much collaboration in studying the equations of Ewing's model.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

The Crystal Structure of Magnesium Platinocyanide Heptahydrate.[1] RICHARD M. BOZORTH and F. E. HAWORTH. Positions of the Mg and Pt atoms in crystals of $MgPt(CN)_4 7H_2O$. These have been definitely determined by means of x-ray oscillating-crystal photographs and Laue photographs, using the theory of space-groups. Because the other atoms are too light in comparison with the metal atoms, especialy Pt, their positions could not be determined. The Pt atoms are located at 0 0 0 and $\frac{1}{2}$ $\frac{1}{2}$ $\frac{1}{2}$, the Mg atoms at 0 0 $\frac{1}{2}$ and $\frac{1}{2}$ $\frac{1}{2}$ 0, in a tetragonal unit of structure 14.6A × 14.6A × 3.13A. Two units of structure are shown in the figure. The peculiar optical properties are believed to be associated with the unusual arrangement of the heavier atoms in widely spaced rows parallel to the tetragonal axis. In these rows Mg atoms alternate with Pt atoms, and the distance between any two adjacent atom-centers is 1.57A. The shortest distance between rows, however, is 10.3A, 6.6 times the distance between atoms in the same row. The atomic radii of Mg and Pt as determined by Bragg from other crystal data do not agree with the observed distance between these atoms, the calculated value being 2.7A, the observed distance 1.57A. The observed distance, however, is consistent with that calculated by the method of Davey, who assumes that the radius of an ionized atom differs much from the radius of the same atom un-ionized, and that the radii of Cs^+ and I^- are substantially equal in crystals of CsI.

Photoelectric Emission as a Function of Composition in Sodium-Potassium Alloys.[2] HERBERT E. IVES and G. R. STILWELL. The entire series of alloys of sodium and potassium have been investigated with respect to the relative values of the photoelectric currents produced by light polarized with the electric vector in and at right angles to the plane of incidence. The pure metals when molten exhibit values below three for the ratio of the two emissions; the alloys show three maxima at compositions approximately 20, 50, and 90 atomic per cent of sodium, with values from 10 to 30 for the ratio; the minima between show low values approximating those for the pure metals. The maxima and minima of the ratio of emissions are due to complicated variations in magnitude of the two emissions compared.

[1] *Physical Review*, Feb. 1927, Vol. 29, No. 2, p. 223.

[2] *The Physical Review*, Feb. 1927, Vol. 29, No. 2, p. 252.

Submarine Insulation with Special Reference to the Use of Rubber.[3] R. R. Williams and A. R. Kemp. (1) Soft vulcanized rubber, though not well adapted to some of the processes of manufacture of submarine cable, can be so made as to be mechanically and electrically suitable and to withstand the action of sea water in a manner comparable with that of gutta percha over a period of a few years. Whether such rubber will retain these characteristics for decades remains to be demonstrated, but it seems probable that it will.

(2) The principal factor to be controlled in producing this result is the amount of water absorbed by the rubber.

(3) Osmotic pressure of internal and external fluids is of prime importance in governing the in-flow of water into rubber and gutta percha.

(4) Lowered water absorption is achieved by removal of water-soluble matter from the rubber, the choice of an insoluble, non-reactive filler of suitable particle size and having a minimum of adsorbed gases or other contamination on its surfaces.

(5) The electrical characteristics of rubber compounds and of gutta percha are clearly related to their water content but are not simple functions of the water content.

(6) It appears that the mode of distribution of water is also extremely important.

(7) Most fillers for rubber compounds are not suitable for submarine insulation, either because of undesirable intrinsic electrical properties or because they are conducive to changes incident to water absorption. Hard rubber dust, silica and zinc oxide are the best fillers from these standpoints so far as known.

An Efficient Apparatus for Measuring the Diffusion of Gases and Vapors through Membranes.[4] Earle E. Schumacher and Lawrence Ferguson. An efficient diffusion measuring apparatus, embodying a mechanical clamp and a mercury seal, is described. This apparatus can be used for measuring the rate of diffusion of gases and vapors through materials such as rubber, waxes, leathers and certain types of paper.

Investigation of the Thermionic Properties of the Rare Earth Elements.[5] Earle E. Schumacher and James E. Harris. Thermionic emission measurements over a range of temperatures were made on samples of pure Ce, La, Pr, Nd, Sa and the aluminum alloys of Yt, Eu, Ga,

[3] *Jr. of the Franklin Inst.*, Jan. 1927, Vol. 203, pp. 35–61.

[4] *J. Amer. Chem. Soc.*, Vol. 49, 427 (1927).

[5] *J. Amer. Chem. Soc.*, Vol. 48, pp. 3108–3117 (1926).

Tb, Dy, Ho, Er, Th, Yb and Lu. These measurements showed the rare earth elements, without exception, to be more active thermionically than the commonly occurring metals. At 1800° C. all of these metals gave emissions of more than 10^5 that of clean tungsten at the same temperature.

The Solidus Line in the Lead Antimony System.[6] EARLE E. SCHUMACHER and FOSTER C. NIX. An investigation of the solidus line above the solid solution field for the lead antimony system was made by the quenching test procedure. Three points were determined between the melting point of pure lead and the end of the eutectic horizontal. The position of the solidus line has been precisely fixed.

Production Control.[7] C. G. STOLL. This paper treats the subject of production control from the practical rather than the theoretical point of view. It is confined largely to a description of the generally accepted principles of production control as applied in the Manufacturing Department of the Western Electric Company. This plant employs approximately 30,000 people and produces annually over $150,000,000 worth of manufactured products. These products are comprised of some 13,000 kinds of apparatus containing over 110,000 different parts.

The paper discusses the organization of the factory, which is set up along functional lines, and also the extensive system of records and charts used to facilitate the work of the organization and to assist in production control.

The Significance of the Dielectric Constant of a Mixture.[8] HOMER H. LOWRY. It is pointed out that in many cases it would be of great value to be able to calculate either the dielectric constant of a mixture of substances of known dielectric constants or, knowing the dielectric constants of a mixture of two components and that of one of the components, to calculate the dielectric constant of the other. A review of the literature, however, shows that this can be rarely accomplished. This is due mainly to the inadequacy of the theories of dielectrics, all of which are insufficiently developed to include the dielectric behavior of mixtures. Nevertheless, as is shown, many attempts have been made to develop formulæ of theoretical significance for application to mixtures. Inspection of the derivation of these formulæ shows that those with the best theoretical background are limited to such special cases that they are of practically no value.

[6] A. I. M. E. Pamphlet No. 1636-E, Feb. 19, 1927.

[7] *Mechanical Engineering*, Vol. 49, p. 201, 1927.

[8] *Jr. of the Franklin Institute*, 203, 413–439, 1927.

A brief review of these formulæ is given together with a brief account of the results of experimental investigations on the dielectric behavior of mixtures. A rather extended bibliography is given.

The Effect of Moisture on the Electrical Properties of Insulating Waxes, Resins and Bitumens.[9] J. A. Lee and Homer H. Lowry. The results of measurements of dielectric constant and effective conductivity at 1,000 cycles and resistivity are reported for 31 waxes, resins and bitumens, including not only naturally occurring products but also commercial dielectrics and mixtures. The measurements were made on the materials initially in a thoroughly dry condition, after six months' immersion in a salt solution corresponding qualitatively to exposure to 98 per cent relative humidity, and after having been redried. All the insulating materials studied absorbed water under the conditions of experiment. The absorption was least with the hydrocarbons and greatest with shellac and bayberry wax. In general, the greatest increase in capacity and conductivity and the greatest decrease in resistivity were shown by the materials which absorbed the most water. The percentage change was much greater in the conductivity and resistivity than in the dielectric constant, as was to be expected.

The Mechanism of the Absorption of Water by Rubber.[10] H. H. Lowry and G. T. Kohman. Data are reported which show the influence of the various factors which determine the amount of water absorbed by any given sample of rubber. From a consideration of the results obtained, it was concluded that, at a given temperature, the most important external factor determining the amount of water absorbed by a given sample of rubber is the vapor pressure of water with which it is in equilibrium. The data show further that the water-soluble constituents within the rubber are responsible for most of the water absorbed at high humidities, that increasing the rigidity of a rubber compound decreases greatly the amount of water absorbed, and that aging increases the water absorption. It is pointed out that all the experimental facts are consistent with the view that the absorption of water by rubber consists of two processes: the formation of a true solution of water in rubber and the formation of solutions internal to the rubber of the water-soluble constituents of the rubber which can be removed by washing.

[9] *Jr. of Industrial and Engineering Chem., 19,* 302–306, 1927.

[10] *Jr. of Physical Chemistry, 31,* 23–57, 1927.

Rapid Evaluation of Baked Japan Finishes.[11] E. M. HONAN and R. E. WATERMAN. The service life of a japan film baked on metal can be evaluated by determining the rate of decomposition of the film when it is placed in an 8.5 per cent phenol-water solution. The effect of the time and temperature of baking the film and the cleanness of the metal previous to applying the japan can also be evaluated. The 8.5 per cent phenol solution is a desirable testing solution because its composition is quite constant at ordinary room temperatures and is not changed by the evaporation of the water.

Magnetostriction. L. W. MCKEEHAN.[1] This paper contains the principal part of three lectures given at the Franklin Institute in April 1926. The history of investigations on the changes in dimensions which accompany magnetization and the changes in magnetization which accompany forcible changes in dimensions is sketched and a classification of the rather complicated cases which have been examined is offered. The bearing of magnetostriction on theories of ferromagnetism is emphasized and a number of new experimental results are described. A representative bibliography is appended.

[11] *Ind. and Eng. Chem.*, Oct. 1926, Vol. 18, pp. 1066–1068.

[1] *Journal of the Franklin Institute* **202**, 737–775 (1926).

Contributors to this Issue

J. R. Shea, B.S. in E.E., University of Wisconsin, 1909; Manufacturing Department, Western Electric Company, 1909–. Mr. Shea is now Assistant Superintendent of the Manufacturing Development Branch covering metal manufacturing and metallurgical work.

S. McMullan, M.A., McMaster University, Toronto, 1912, LL.B., Chicago Kent College of Law, 1924; Manufacturing Department, Western Electric Company, 1916–. Mr. McMullan is now Development Engineer in the Manufacturing Development Branch on the manufacture of copper rod and wire.

C. R. Moore, B.S. in Mechanical and Electrical Engineering, Purdue, 1907; E.E., Purdue, 1910; Instructor and Assistant Professor Electrical Engineering, Purdue, 1907–13; Manager of La Fayette Electric and Mfg. Co., 1913–14; Associate in Electrical Engineering, University of Illinois, 1914–16; Engineering Department of the Western Electric Co., 1916–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Moore for several years has been associated with transmitter development work and has contributed important inventions relating to telephone instruments and acoustic devices.

A. S. Curtis, Ph.B., 1913, E.E., 1919, Sheffield Scientific School; Instructor in Electrical Engineering, Yale University, 1913–17; Engineering Department, Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Curtis' work has been connected with the development of telephone instruments.

A. G. Landeen, E.E., University of Minnesota, 1910; incandescent lamp development and manufacturing, General Electric Company, 1910–19; Engineering Department, Western Electric Company, 1919–24; Bell Telephone Laboratories, 1925–.

Ralph Bown, M.E., 1913, M.M.E., 1915, PhD., 1917, Cornell University; Captain Signal Corps, U. S. Army, 1917–19; Department of Development and Research, American Telephone and Telegraph Company, 1919–. Mr. Bown has been in charge of work relating to radio transmission development problems and recently has given particular attention to the quantitative and engineering side of transatlantic telephony.

W. P. Mason, B.S., University of Kansas, 1921; M.A., Columbia, 1924; Engineering Department, Western Electric Company, 1921–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Mason has been associated with work on carrier current communication, and more recently with the development of filters and subscribers' sets.

Karl K. Darrow, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D. in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

The Bell System Technical Journal

July, 1927

Measurement of Inductance by the Shielded Owen Bridge

By J. G. FERGUSON

Synopsis: The study described in this paper shows that the Owen bridge is well adapted to the accurate measurement of inductance and effective resistance to above 3,000 cycles. The construction of a shielded bridge for audio frequencies is described and a theoretical discussion is also given. It was found possible to measure inductances ranging from 0.1 to 3 henrys with an error of measurement less than 0.1 per cent, and for 10 henrys the accuracy is better than 0.25 per cent. As a means of measuring effective resistance the bridge shows an accuracy of about 2 per cent. The sources of error and method of eliminating or correcting them are discussed.

Introduction

THE accurate measurement of inductance and capacitance is essential to the correct design of practically all precision electrical apparatus. Particularly is this so in the field of electrical communication where the successful introduction of new circuits and equipment, such as the carrier telephone and the telephone repeater, depends largely on the accuracy with which the elements can be adjusted to the nominal values, this accuracy in turn depending on the accuracy with which the electrical measurements can be made.

Owing principally to the ease with which a telephone receiver may be used to indicate a balance at audio frequencies, bridge measurements are very generally used for the measurement of capacitance and inductance in telephone work. The simplest type of bridge and the one used most for the comparison of like impedances is the equal ratio arm bridge described by Shackelton.[1] This bridge requires standards of the same kind and magnitude as the impedances which are to be measured. The calibration of these standards is a separate problem, for which a distinct type of bridge is required.

Either capacitance or inductance may be measured by a bridge method in terms of time and resistance, both of which are fundamental quantities. However, since condensers may be obtained with very low losses and small changes with frequency, this type of measurement is usually made with capacitance,[2] inductance measurements being

[1] W. J. Shackelton, "A Shielded Bridge for Inductive Impedance Measurements," *Bell System Technical Journal*, January, 1927.

[2] J. Clerk Maxwell, *Electricity and Magnetism*, Vol. 2, pp. 776–7.

made by comparison with capacitance and resistance or with capacitance and frequency. The resonant method is adapted to the comparison of inductance with capacitance and frequency. However, this method demands an accurate measurement of the frequency used, which is not always convenient. It is therefore evident that a bridge which furnishes a comparison of inductance with capacitance and resistance serves a very useful purpose in the calibration of standards of inductance for use in simple comparison bridges.

A bridge circuit due to Owen [3] furnishes a very good example of this type, the balance conditions being independent of frequency and the equations of balance giving a relation between inductance, capacitance, and resistance. The circuit is shown in Fig. 1. It consists of a fixed resistance r_1 in the arm BC, a fixed capacitance C_3 in the arm AB, a fixed capacitance C_4 in series with a variable resistance R in

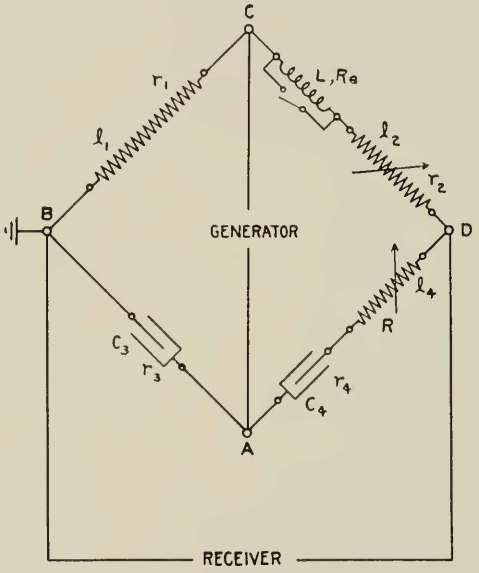


Fig. 1

AD, and a variable resistance r_2 in series with the inductance to be measured in CD. The adjustments for balance are made with R and r_2. These two adjustments are independent of each other. The relations between the quantities at balance, as will be shown later, are such that the bridge may readily be made direct reading for

[3] D. Owen, "A Bridge for the Measurement of Self Induction," *Proceedings of Physical Society of London*, Oct. 1, 1914.

inductance, and these advantages make this bridge superior to practically all other bridges for this type of comparison.

This paper contains a discussion of the theoretical relations of this bridge circuit, its possibilities and limitations for the accurate measurement of inductance and effective resistance, and the sources of error and methods of eliminating them. A shielded bridge, constructed for use in calibrating inductance standards, is described and sufficient measurements are given to show the accuracy of which it is capable.

The maximum frequency at which measurements were given by Owen is 530 cycles. For the measurement of telephone apparatus considerably higher frequencies are used, and it is desirable that the bridge be capable of measurements up to 3,000 cycles without loss of accuracy. It is in the upper part of this range that the greatest difficulties are encountered, requiring special precautions not so necessary for the lower frequency measurements.

While in the following discussion the maximum frequency considered is 3,000 cycles, this is not meant to indicate a maximum limit to this type of bridge.

Equations of Balance

Taking into consideration the phase angle of the resistances and the loss in the condensers, the complete network is shown in Fig. 1, the reactive component of the resistances being shown as series inductance, and the condenser losses as series resistance. Let

L and R_e = Inductance and effective resistance of coil to be measured,
r_1 and l_1 = Total resistance and inductance in arm BC,
r_2 and l_2 = Resistance and inductance in CD exclusive of R_e and L,
R and l_4 = Total resistance and inductance in AD including the equivalent series resistance of C_4,
r_3 = Equivalent series resistance of C_3.

The inductance in the arm AB may readily be reduced to a negligible amount and will not be considered.

We may now balance the bridge with the inductance terminals short circuited, that is, take a zero reading, and then balance again with the inductance inserted.

Writing the equations of balance in each case, subtracting one from the other, and separating reals from imaginaries, we get the following equations:

$$C_3r_1(R - R') = L + (l_2 - l_2') + C_3r_3(r_2 + R_e - r_2') + p^2C_3l_1(l_4 - l_4') \quad (1)$$

and

$$\frac{r_2' - r_2 - R_e}{C_3} = p^2 l_1(R - R') + p^2 r_1(l_4 - l_4') - p^2 r_3(L + l_2 - l_2'), \quad (2)$$

where l_2', r_2', l_4', and R' are the values of l_2, r_2, l_4, and R at balance with L short circuited, and p is 2π times the frequency. These are practically identical with Owen's equations (10) and (12).

In equation (1), each of the third and fourth terms contains two factors of second order, namely r_3 and $(r_2 + R_e - r_2')$, and l_1 and $(l_4 - l_4')$ respectively.

We may therefore write

$$C_3 r_1(R - R') = L + (l_2 - l_2'). \quad (3)$$

In equation (2), let

$$\frac{1}{pC_3} = -X_3, \qquad pl_1 = x_1, \qquad pl_4 = x_4,$$
$$pL = X, \qquad pl_2 = x_2, \quad \text{and} \quad pl_2' = x_2'.$$

Then we may write

$$-(r_2' - r_2 - R_e)pX_3 = px_1(R - R') + pr_1(x_4 - x_4') - pr_3(X + x_2 - x_2'). \quad (4)$$

But from (3)

$$R - R' = \frac{L + l_2 - l_2'}{C_3 r_1} = \frac{-(X + x_2 - x_2')X_3}{r_1}.$$

Substituting in (4),

$$r_2' - r_2 - R_e = \frac{(X + x_2 - x_2')x_1}{r_1} + (x_4 - x_4')\frac{(X + x_2 - x_2')}{R - R'} + \frac{r_3(X + x_2 - x_2')}{X_3}$$

$$= (X + x_2 - x_2')\left[\frac{x_1}{r_1} + \frac{x_4 - x_4'}{R - R'} + \frac{r_3}{X_3}\right]$$

and

$$R_e = r_2' - r_2 - (X + x_2 - x_2')\left(q_1 + q_4 + \frac{1}{Q_3}\right), \quad (5)$$

where q_1 = ratio of reactance to resistance of arm BC,
q_4 = ratio of reactance to resistance of *change* in arm AD,
Q_3 = ratio of reactance to resistance in arm AB.

From equation (3) we see that, if we take a zero reading first, the inductance is given by the expression

$$L = C_3 r_1 (R - R'), \tag{6}$$

the percentage error due to neglecting $l_2 - l_2'$ being

$$\frac{100(l_2 - l_2')}{L}. \tag{7}$$

From equation (5), the effective resistance of L is given by

$$R_e = r_2' - r_2, \tag{8}$$

the percentage error due to neglecting corrections being

$$\frac{100(X + x_2 - x_2')}{R_e}\left(q_1 + q_4 + \frac{1}{Q_3}\right). \tag{9}$$

The error in L is approximately, from equations (7) and (8),

$$\frac{x_2 - x_2'}{R_e} \cdot \frac{R_e}{X} = \frac{q_2}{Q},$$

where q_2 = ratio of reactance to resistance of change in arm CD, and
Q = ratio of reactance to resistance of the inductance being measured.

This error is usually negligible and may be approximately corrected for when appreciable. Dr. Owen has pointed out that this type of error is not peculiar to the Owen bridge, but is present in practically all methods of inductance measurement.

The error in R_e is a function of the Q of the coil measured, and of q_1, q_4 and Q_3. It is greatest for coils of high Q.

It is possible to make $q_1 = -\frac{1}{Q_3}$ for a given frequency, in which case the error reduces to approximately Qq_4 and the two errors are of the same order of magnitude for $Q = 1$,—the error in R_e becoming greater, and in L less as Q is increased.

However, in the general case we cannot cancel q_1 against $\frac{1}{Q_3}$ over any appreciable range of frequencies, and they are normally additive. Also for ordinary inductance coils Q is considerably greater than one, sometimes as large as 100. For such cases the error in R_e becomes large and difficult to determine without an accurate knowledge of the reactances of the resistances used and the losses in the condenser.

From the above relations we see that a method of this type is capable of measuring inductance with a high degree of accuracy and may be made to measure effective resistance with fair accuracy,

provided that there is no coupling between any of the four arms nor any between them and the input and output circuits. This is in practice a difficult result to realize, and this difficulty in obtaining a simple but adequate system of shielding is one of the most serious limitations to the bridge.

Shielding

Since the bridge contains no inductances of appreciable magnitude, it is a comparatively simple matter to eliminate electromagnetic coupling by using input and output transformers in toroidal form, the input transformer being so designed that the core will not be saturated when using the maximum input to the bridge.

The elimination of the electrostatic coupling is not so simple, as any electrostatic shielding introduced adds capacitance which, unless due care is taken, will involve errors in the bridge. This means that such capacitances must be limited to the corners BD and AC where

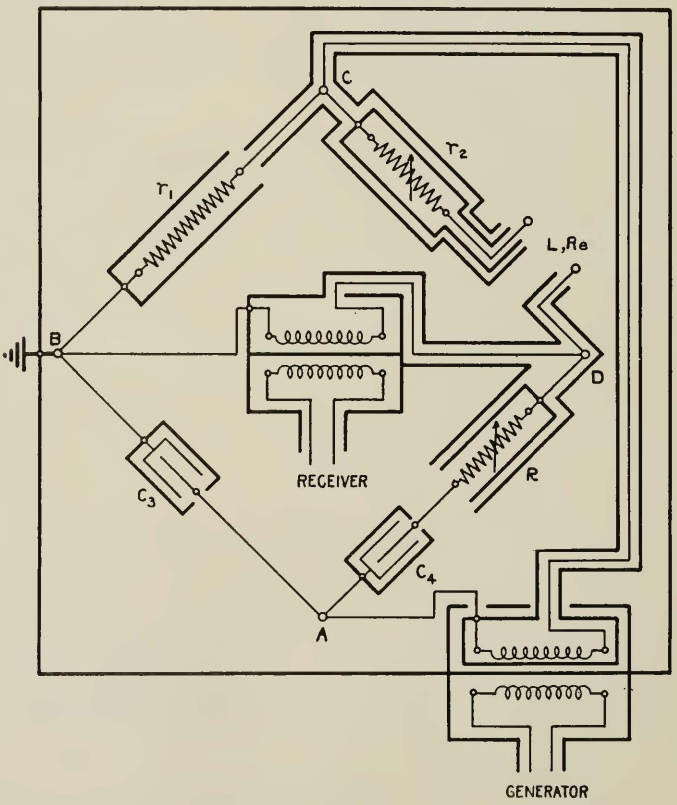


Fig. 2

they simply shunt the input and output circuits, and to AB where they shunt the capacitance C_3 and may be included in the assumed value of C_3. If such shielding is not used, the balance of the bridge will be affected by external conditions such as body capacitance, and the position of the bridge arms with respect to each other and to other apparatus, with the result that accurate results can be obtained only by the use of the greatest precautions.

A shielding scheme which satisfies the above requirements is shown in Fig. 2. In this system all capacitance between shields is limited to the diagonal corners of the bridge and the arm AB. However, this system of shielding, while about as simple as can be designed where complete shielding is required, is rather difficult to carry out in any practical bridge construction.

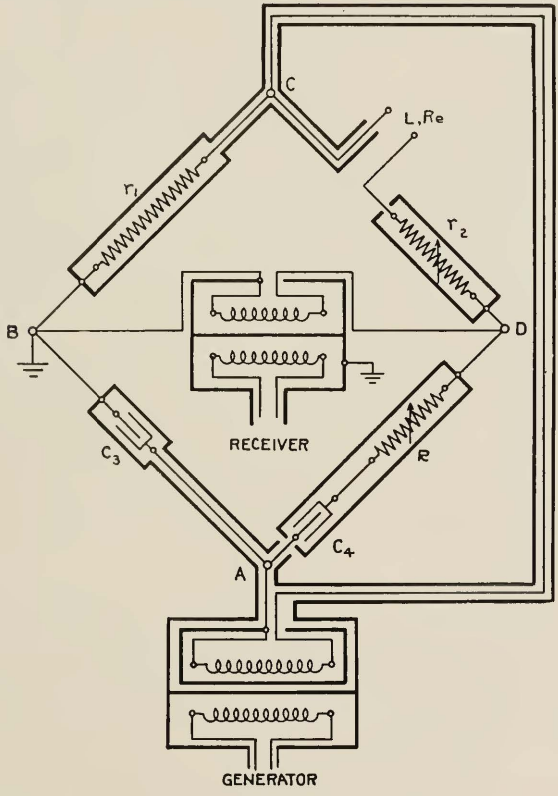


Fig. 3

The question of reducing the amount of shielding and still retaining a high degree of accuracy has been investigated and the modified scheme shown in Fig. 3 has been developed. In this circuit the

shielding is complete insofar as it limits the electrostatic coupling to specific points in the bridge, and eliminates coupling between the bridge and the input and output circuits. However, in addition to capacitance across the diagonal corners and across arm AB, capacitances are introduced across r_1, across R, and across arm AD. The capacitance across AD may be made small enough to neglect since it consists of the capacitance of one condenser lead to the shield. Capacitances across r_1 and across R do not enter as first degree errors in the value of L but do directly affect the measurement of R_e. However, where the bridge is used primarily for the accurate measurement of inductance this compromise is justified. Even for the measurement of effective resistance, although the corrections may be larger due to the presence of the shielding, the bridge will give more consistent results and the corrections may be fairly well estimated.

The method of shielding shown requires one transformer having two shields between the windings and one transformer with a single shield between windings. It is essential that these shields be as perfect as possible. The other shielding shown is comparatively simple, no equipment requiring more than a single shield. The ground is shown at the point B simply because grounding at this point results in the simplest shielding. It would be desirable to have the ground at C in order that one terminal of the coil under test would be grounded, but at the time of balance the points B and D are at the same potential, and provided that r_2 is only a small fraction of the total impedance of the coil under test we may consider that one terminal of the coil is practically at ground potential. However, it should be noted that for a coil having a considerable capacitance from intermediate points in its winding to ground, a ground at B cannot be considered exactly equivalent to a ground at D. This difficulty is only appreciable in the case of very large inductances of large physical size when measured at high frequencies, and in such cases the effective inductance will be dependent on external conditions, whatever bridge circuit it is measured in. In the case of shielded coils, the ground should in all cases be connected to D rather than to B. In spite of the slight disadvantages noted, this method of shielding appears to be the most satisfactory, and a bridge has been constructed in accordance with it.

Construction of the Bridge

From the equation giving the value of L, it is seen that we may obtain an additional range for the inductance by having either r_1, C_3, or both, variable in steps. In the present bridge we have

used two steps for C_3 and five steps for r_1. It is possible by choosing the correct values for r_1 to make the bridge direct reading for inductance. The actual values used for the capacitance were .6 mf and .06 mf. The values used for r_1 were 1,000/.6 or 1,667 ohms and multiples or submultiples of this value. In this way the bridge was made direct reading in millihenrys.

The capacitance C_4 has only one requirement to meet. It must be small enough so that the ratio of resistance to reactance of arm AD shall always be less than the ratio of reactance to resistance of the coil.

Taking 3,000 cycles as the maximum frequency, 10,000 ohms as the maximum resistance in arm AD, and 200 as a maximum value for the Q of the coil measured, then

$$2\pi fC < 200,$$

and

$$C < 1 \text{ mf}.$$

We have accordingly used a value of .6 mf in this arm to correspond with the value of C_3.

Resistances R and r_2 are dial type completely shielded resistance boxes which can be varied from 0 to 10,000 ohms in .01 ohm steps. The resistances are all of the reversed layer type, wound on impregnated wood spools and designed to give low phase angle and high stability.

The condensers are of the paraffine impregnated mica type, about ten years old, thus ensuring high stability, and having temperature coefficients less than .003 per cent per degree C., over the ordinary range of working temperatures.

The transformers are of a special type described by Shackelton.[1]

Accuracy—Measurement of Inductance

As previously stated the shielding, while increasing the stability of the bridge, introduces capacitances across R and r_1 which increase the corrections necessary in computing the effective resistance and may also require corrections in the measurement of inductance if sufficiently large. Accordingly, measurements were made on the bridge to determine the magnitude of this error. By shunting R and r_1 respectively, it was readily shown that capacitances as high as 200 mmf would not change the indicated inductance reading by as much as .01 per cent for all settings of r_1, for the whole range of R, over the whole audio frequency range. This conclusion is in accordance with equation 1. Since the shielding introduced capacitances

across these points of the order of 25 to 50 mmf, this source of error may be neglected in the measurement of inductance.

Table I gives the exact values for C_3 and r_1, and the corresponding constant K by which the indicated value of R must be multipled to give the true inductance. This table shows how accurately the resistance r_1 has been adjusted to make the bridge direct reading. K is a simple number within .02 per cent in all cases when using the large condenser. The two condensers might have been made to have a ratio more nearly 10 to 1 by adding an auxiliary condenser to the larger one.

TABLE I

$K = C_3 \times r_1$ = Millihenrys per Ohm

r_1 (Ohms)	82.785	165.59	828.04	1656.1	8280.9
C_3 (mf)					
.60381	.049987	.099985	.49998	.99998	5.000
.06052	.0050103	.010022	.050113	.10023	.50117

A check was next made on a single inductance having a nominal value of .1 henry to determine the relative accuracy of different values of K at different frequencies. These values are given in Table II. It will be noticed that the value of L obtained is approximately

TABLE II

Comparison of Different Values of K Using a Single Inductance

Nominal Inductance, Millihenrys	K	Frequency, Cycles	R Ohms	R' Ohms	$L = K(R - R')$ Millihenrys
100	.099985	1,000	1,006.64	.03	100.65
"	.49998	"	201.34	.00	100.66
"	.050113	"	2,009.4	.45	100.67
"	.049987	"	2,013.4	.09	100.64
"	.099985	3,000	1,022.0	.03	102.18
"	.49998	"	204.40	.00	102.20
"	.050113	"	2,040.3	.00	102.24
"	.049987	"	2,044.1	.09	102.17

independent of K but the highest value obtained is for the value of K corresponding to the highest value of r_1. Since the reactance of this coil is only approximately 600 ohms at 1,000 cycles and the largest value of r_1 used was 828 ohms, it is evident that the potential of the coil with respect to ground varies considerably for different values of K. This is sufficient to account for the increased inductance value obtained for values of K using $r_1 = 828$ ohms. Keeping this

in mind the different values of K agree with each other very closely. It has already been stated that r_1 should be small compared with X and therefore the values of K using $r_1 = 828$ ohms would not normally have been used for the measurement of this coil.

Table III gives a comparison of the inductance of several coils as measured on the Owen bridge and by a resonant method, the last column giving the difference between the two methods in per cent.

TABLE III

Comparison of Owen Bridge with Resonance Bridge

Nominal Inductance, Henrys	Frequency, Cycles	Measured Inductance		Difference, Per Cent
		Owen Bridge, Henrys	Resonance, Henrys	
.1	1,000	.10065	.10066	− .01
.1	2,000	.10124	.10118	+ .06
.15	1,000	.15072	.15082	− .04
.15	2,000	.15112	.15111	+ .01
1.0	2,000	1.0143	1.0144	− .01
2.9	1,000	2.918	2.918	.0
2.9	2,000	2.976	2.974	+ .07
10.0	2,000	11.295	11.27	+ .22

The resonant method was a highly accurate one in which frequency errors were negligible. The accuracy was probably of the same order as the measurements on the Owen bridge. The agreement between these two methods does not in itself indicate the accuracy of either method. However, the resonant measurements were made on a completely shielded equal ratio-arm bridge,[1] in terms of frequency and capacitance, using entirely different equipment from the Owen bridge in which the inductance is measured in terms of resistance and capacitance. Accordingly it is very improbable that these two methods had any errors in common and we may assume that the agreement obtained is a fair measure of the combined error of the two methods. Consequently from this table we see that for a range of .1 to 3 henrys and for frequencies up to 2,000 cycles the error in the measurement of the inductance by the shielded Owen bridge is less than .1 per cent and for 10 henrys is less than 1/4 per cent.

Accuracy—Measurement of Resistance

The measurement of effective resistance in the case of an impedance of low reactance practically consists of the substitution of the unknown for the known resistance. In this case the accuracy of the measure-

ment is high. However, the usual case we have to consider is the measurement of the effective resistance of coils of high Q. It is in such measurements that the greatest corrections are necessary, and it is also in such measurements that the greatest errors in effective resistance are produced by incomplete shielding in the bridge. Consequently it is in the measurement of effective resistance that shielding is most essential, and although this shielding may introduce a necessity for larger corrections due to the capacitance it introduces, these corrections may be made with a certain degree of precision and having made them the value obtained will be more reliable than in the case of a complete absence of shielding.

Table IV gives the figures for the measurement of effective resistance of three coils having a high Q. Referring to equation 5 we see that q_1 and q_4 are positive when the reactance is inductive and that Q_3 is

TABLE IV

Inductance, Henrys	Frequency, Cycles	r_2' Ohms	r_2 Ohms	q_1	q_4	$\frac{1}{Q_3}$	$X\left(q_1+q_4+\frac{1}{Q_3}\right)$	R_e Ohms	R_e' Ohms	Diff., %
1	1,000	1,670.66	1,358.30	− .0004	.0000	+ .0023	− 17	329.4	326	1
1	3,000	1,670.15	1,373.14	− .0013	"	"	− 68	365	378	3
.1	1,000	1,670.66	1,641.46	− .0004	"	"	− 1.7	30.9	30.1	3
.1	3,000	1,670.15	1,643.34	− .0013	"	"	− 6.8	33.6	32.5	3
.02	1,000	1,670.66	1,659.30	− .0004	+ .0003	"	− .30	11.66	11.47	2
.02	3,000	1,670.15	1,659.15	− .0013	+ .0011	"	− .94	11.94	11.8	1

always negative. The column headed R_e is obtained from equation 5. The column headed R_e' is obtained from a resonant method of measurement which has the same order of accuracy as the present method. Consequently the last column of differences gives the combined error in the two methods. In these measurements covering the most used range of inductance and a frequency range of 1,000 to 3,000 cycles, the largest difference between the two methods is 3 per cent. The total corrections to be made are in some cases extremely large, especially for the higher inductances and frequencies. This correction may amount to 30, or 40 per cent in some cases, and this means that an effective resistance obtained by the Owen bridge when not corrected may be in error by this amount. However, after allowing for the necessary corrections we can say that the bridge is capable of an accuracy for the measurement of effective resistance of about 2 per cent over the greater range of inductance and frequency.

Determination of Electrical Characteristics of Loaded Telegraph Cables

By J. J. GILBERT

Synopsis: The use of permalloy for continuous loading has introduced a number of new factors of importance in the study of transmission of signals over long submarine telegraph cables. Data to check the theoretical assumptions that are used in the design of permalloy loaded cables can be obtained by measuring on such cables the attenuation and time of propagation of sinusoidal currents of various frequencies in the telegraph range. By combining the results of these measurements with data obtained on the cable during process of manufacture, the resistance, inductance, capacity and leakance of the cables can be determined.

This paper describes the experiments that were performed on three laid cables and discusses in a general way the methods of computing the cable parameters.

WITHIN the last few years the art of telegraphing over submarine cables of transoceanic length has been revolutionized by the development of effective means of applying to such cables the principle of inductive loading. By surrounding the copper conductor of the cable with a thin layer of permalloy, a material of high magnetic permeability, the range of signal speeds attainable over cables of the order of 2,000 n.m. in length has been multiplied eight to ten times.[1] In place of the low frequency band extending from zero to about 15 c.p.s., which represents the range of frequencies which can be efficiently transmitted over the usual type of non-loaded cable, we are concerned in the case of the loaded cable with a transmission band extending from zero to about 120 c.p.s. Largely because of this comparatively high speed of operation, a number of factors, which were of negligible influence in the case of non-loaded cables, have become of primary importance in affecting the speed of signalling, and it has been found necessary, in order to establish a definite basis of estimating the performance of loaded cables, to make a thorough study of these factors by theoretical analysis supplemented by experimental work in the laboratory, and by measurements on laid cables.

Principles of Cable Transmission

The theory of transmission of signals over submarine telegraph cables [2] and the principles governing the design of permalloy loaded

[1] O. E. Buckley, *B. S. T. J.*, Vol. IV, No. 3, July 1925; Electrical Communication, Vol. 4, No. 1, July 1925; *Jour. A. I. E. E.*, Vol. XLIV, No. 8, August 1925.

[2] H. W. Malcolm, "The Theory of the Submarine Telegraph and Telephone Cable," London, 1917.

J. W. Milnor, *Jour. A. I. E. E.*, Vol. 41, p. 118, 1922.

cables [1] have been fully discussed elsewhere and only a brief summary will be given here for the purpose of indicating the importance of the measurements that will be described. On account of the fact that for a given value of sending voltage the amplitude of the signals received over a submarine cable diminishes rapidly as the speed of signalling is increased, there is a practical limit to the speed of operation of any cable. This limit depends on the electrical characteristics of the cable and the magnitude of extraneous interference encountered at the receiving terminal. The criterion for legibility of signals is, in general, that the attenuation constant of the cable at a value of frequency which may be termed the critical frequency shall not exceed a given value, the attenuation constant αs being defined by the relation

$$\frac{|V_R|}{|V_S|} = \epsilon^{-\alpha s}, \tag{1}$$

where $|V_R|$ is the amplitude of voltage arriving at one end of the cable when a sinusoidal voltage of amplitude $|V_S|$ is impressed at the other terminal. The value of this critical frequency depends mainly upon the method of operation, and it usually lies somewhere between the signal frequency and one and one half times the signal frequency.

Given the values of the four fundamental parameters of the cable, resistance (R), inductance (L), capacity (C) and leakance (G), the attenuation constant at the frequency $p/2\pi$ can be computed by means of the formula

$$2\alpha^2 = \sqrt{(R^2 + p^2L^2)(G^2 + p^2C^2)} + RG - p^2LC, \tag{2}$$

which to a close approximation reduces to the form

$$\alpha = \sqrt{\pi f CR} \tag{3}$$

in the case of a non-loaded cable, where R is large compared with $2\pi fL$, and to the form

$$\alpha = \frac{1}{2}\left(R + \frac{G}{C}L\right)\sqrt{\frac{C}{L}} \tag{4}$$

in the case of the loaded cable, where R is small compared with $2\pi fL$ at the critical frequency. In all cases it is assumed that G is very small compared with $2\pi fC$, which is strictly true for the insulating materials employed on submarine cables.

The manner in which the attenuation constant varies with frequency

for typical loaded and non-loaded cables is shown in Fig. 1, the signal frequencies at which they are designed to operate being as indicated. In the case of the non-loaded cable the resistance and capacity are practically constant over the frequency range and the attenuation

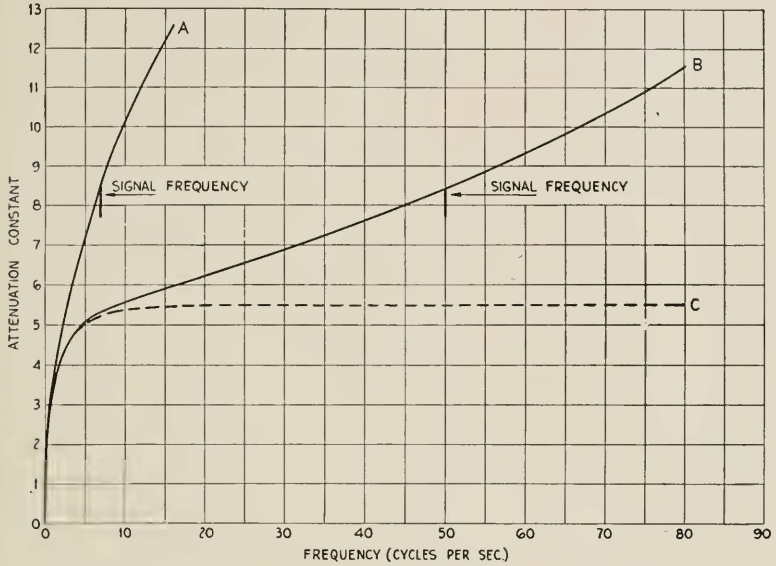


Fig. 1—*A*—non-loaded cable; *B*—actual loaded cable; *C*—ideal loaded cable

curve is approximately a parabola as indicated by formula (3). The curve for the loaded cable for small values of frequency is similar to the curve for the non-loaded cable, since for such frequencies the loading inductance has very little effect upon transmission. As soon as $2\pi fL$ becomes appreciable compared with R the beneficial effect of the inductance becomes apparent and the attenuation constant increases at a less rapid rate. If the cable parameters were constant throughout the frequency range, as in the case of the ideal cable, the attenuation constant would, at a value of frequency considerably below the signal frequency, attain a constant value, as represented by the dotted curve. On account of the fact, however, that R and G increase rather rapidly with frequency, the attenuation-frequency characteristic of an actual cable merely inflects, then increases, and at some frequency will actually cross the attenuation curve of the non-loaded cable.

To insure that legible signals will be obtained at the desired signal frequency the amplitude of the extraneous interference must be

accurately determined. If, for example, the interference in the case of the cable having the attenuation-frequency characteristic shown in curve *B* were found to be twice as great as had been anticipated, the amplitude of received signal would likewise have to be doubled, which would mean a reduction of 0.7 in the allowable attenuation constant. This, as can be seen from curve *B*, would correspond to a reduction in speed of 8 to 10 c.p.s. Also since the value of attenuation constant is considerably affected by variations of the electrical parameters, it is desirable that the values of these parameters in the laid cable be capable of predetermination to a degree of accuracy comparable with that obtained in the case of non-loaded cables. Methods of estimating the value of extraneous interference to be expected at the terminals of a projected cable have been described in a previous paper.[3] The present paper will be devoted to a discussion of methods of predetermining the electrical parameters of cables.

Measurements during Manufacture

In the case of a non-loaded cable the attenuation constant, as indicated by formula (3), is determined solely by the dielectric capacity and the conductor resistance. For the values of frequency involved in the operation of such cables, the latter consists almost entirely of the direct current resistance of the copper conductor. The values of capacity and copper resistance of a considerable part of the cable can be measured during the process of manufacture, and, by reducing these values to sea bottom conditions, an accurate estimate of the resistance and capacity of the laid cable is obtained.

In the case of the loaded cable the problem of predetermining the electrical parameters of the laid cable is much more difficult, since a number of the quantities involved in computing the attenuation are influenced by conditions which are not entirely known and which are difficult to simulate in laboratory experiments. The dielectric leakance, for example, is affected by pressure as well as by temperature, and since the hydrostatic pressure to which the cable is subjected may be as high as 10,000 pounds per square inch, it is evident that measurements of this characteristic of the cable, on any but a very small scale basis, will be very difficult and costly. The permeability of the loading material and consequently the inductance of the cable may be affected by mechanical strain and by superposed magnetic fields. An estimate of the average inductance of the laid cable can be obtained by bridge measurements in the factory on pieces of core about 1

[3] J. J. Gilbert, *B. S. T. J.*, Vol. 5, p. 404, and *Electrician*, Vol. 97, August 6, 1926.

nautical mile in length, selected at intervals during manufacture, the effect of strains and of superposed fields being estimated by means of experiments on short lengths of cable. However, there are ordinarily small unavoidable variations in electrical characteristics from point to point along the cable and it is not entirely certain that the average inductance obtained from measurements on a fraction of the core lengths entering into the cable structure will represent the average inductance of the entire cable. The resistance of the laid cable is likewise difficult to estimate. This parameter comprises, in addition to the copper resistance, the resistance of the return conductor consisting of the armor wires and sea water in parallel, components resulting from eddy current and hysteresis losses in the loading material and other components of lesser importance, the nature of which will be discussed later. The losses in the loading material depend upon the average permeability obtained in the laid cable, and their predetermination from factory measurements may be uncertain for reasons that have been pointed out. As regards the sea return resistance, rigorous methods of computation are available,[4] but there is some uncertainty regarding the conditions that should be assumed as existing at the ocean bottom.

Measurements on Laid Cables

For the purpose of placing the design of loaded cables upon a definite basis, it has appeared desirable to measure the parameters of a number of cables of this type that have been laid, and to compare the values so obtained with the estimates based on analytical methods and upon factory measurements. In order to simplify the problem, attention will be devoted mainly to determining the values of the parameters corresponding to a very small value of current in the cable conductor. Under these conditions the hysteresis component of resistance is negligible and the inductance and eddy current resistance can be considered constant at any frequency. This is entirely consistent with the method employed in the design of loaded cables, in which the attenuation constant is computed, first on the assumption that the current is very small throughout the cable, and then corrected for "head end losses" due to the effect of hysteresis losses which are present under actual conditions of operation.

The usual method of determining the parameters of a transmission system consists in measuring the propagation constant, Γ, per unit

[4] J. R. Carson and J. J. Gilbert, *Jour. Franklin Institute*, Vol. 192, p. 705, 1921; *Electrician*, Vol. 88, p. 499, 1922; *B. S. T. J.*, Vol. 1, No. 1, p. 88.

length and the characteristic impedance, K, which quantities are defined at the frequency $p/2\pi$ by the formulas

$$\Gamma = \sqrt{(R + jpL)(G + jpC)}, \tag{5}$$

$$K = \sqrt{\frac{R + jpL}{G + jpC}}. \tag{6}$$

Knowing these two quantities at any frequency, the values of the four parameters can be readily computed.

The propagation constant and the characteristic impedance of *telephone* cables 100 miles or less in length have been determined by measuring the input impedance of the cable with the distant end in turn insulated and grounded. These two impedances are determined for a cable of length s by the formulas

$$Z_I = K \coth \Gamma s$$
$$Z_G = K \tanh \Gamma s,$$

and given the values of Z_I and Z_G it is an easy matter to compute the corresponding values of propagation constant and characteristic impedance, the accuracy of this determination depending upon the difference between Z_I and Z_G. In the case of a submarine telegraph cable of the order of 2000 miles in length, the value of Γs is so large that Z_I and Z_G differ by less than one part in 10,000 in the frequency range in which we are interested. This means physically that the remote parts of the cable have little effect upon the terminal impedance of the cable and the values of input impedance are determined almost entirely by the parameters of the 400 or 500 miles of cable adjacent to the terminal. It is true that by going to extremely low frequencies, perhaps fractional cycles per second, the method above described could be used to determine the characteristic impedance and the propagation constant of long cables, but at such frequencies these quantities are determined almost entirely by the d.c. resistance and capacity of the cable and no information regarding the quantities in which we are particularly interested would be obtained.

The method that has actually been employed to determine the parameters of several of the continuously loaded cables which have recently been laid is to measure separately at a number of frequencies the real and imaginary parts of the propagation constant, the capacity of the cable at various frequencies being determined by correlating the results of laboratory tests with d.c. measurements of capacity made on the laid cable.

As can be seen from formula (4), the real part of the propagation constant, αs, the attenuation constant of the cable, involves all four of the cable parameters, but on account of the fact that the inductance, leakance and the various components of the effective resistance predominate in influence at different points in the frequency range it is possible, by methods of successive approximations, to obtain a reasonably good set of values of these quantities.

The imaginary part of the propagation constant, βs, is to a close approximation, given by

$$\beta s = sp\sqrt{CL}. \tag{7}$$

From this it follows that the time of propagation of a sinusoidal wave of voltage or current over the cable is given by

$$T = s\sqrt{CL}, \tag{8}$$

and knowing the time of propagation and the capacity at any frequency the inductance of the cable at this frequency can be easily computed. Since the resistance and leakance have only a slight effect upon the time of propagation, this is the most direct method of determining the average inductance of the cable.

Measurement of Attenuation

The attenuation constant of the cable is determined by measuring the values of voltage received at one end of the cable, due to various values of voltage of constant frequency impressed at the other end. The impressed voltage may be either sinusoidal or square top in shape, the latter being preferable for the reason that, at the low frequencies and high voltages required, it is difficult to obtain a wave form from an oscillator sufficiently free from harmonics to enable an accurate determination of the fundamental component to be made. Square top reversals of any frequency and amplitude can be easily obtained by means of a relay actuated by an oscillator, and the amplitude of the fundamental component can be accurately computed.

At the receiving end, for the frequencies of particular interest, the arriving voltage is practically sinusoidal, since the harmonic components are eliminated by the higher attenuation of the cable for such frequencies. This voltage is measured by terminating the cable in an impedance which is very large compared to the characteristic impedance of the cable, and measuring the potential drop across all or part of this impedance by means of a vacuum tube amplifier in the output of which is a thermocouple and meter. The advantages of the high impedance termination are, first, that by reflection it

doubles the amplitude of the arriving voltage, thus giving larger quantities to work with, and second, that it eliminates the necessity of taking into account the characteristic impedance of the cable and the impedance of the balanced type of sea earth which is usually employed as the earth connection of the amplifier. By means of a string oscillograph in the output of the amplifier, the wave shape of the received voltage and the nature of the extraneous interference can be determined. The amplifier is calibrated by impressing on it a measured voltage of the same frequency as that of the received voltage.

Knowing the values of received voltage and the corresponding transmitted voltage, the values of attenuation constant can be readily computed. By plotting the values of attenuation constant corresponding to various values of frequency and transmitted voltage as functions of the latter quantity and extending these curves to the axis of zero transmitted voltage, the values of attenuation constant corresponding to a very small current in the conductor can be obtained for various frequencies.

Assuming that all the parameters have been accurately predetermined, there are three sources of error which might possibly cause a difference between the measured value of attenuation constant and that computed from the average values of the cable parameters by means of formula (2). In the first place, the parameters are not uniform throughout the cable as is assumed in deriving this formula. In particular, the inductance may vary from point to point. At each point where the capacity or inductance changes value reflections of voltage and current will take place and the effect of these reflections should be to increase the attenuation constant of the cable. For variations of the parameters of the order that is to be expected in loaded cables, the increase in attenuation constant is quite small, and the magnitude of this increase can be computed approximately by a method due to Carson.[5] Another source of error is the presence of extraneous interference superposed on the received voltage. This factor is usually troublesome only at the highest frequencies and lowest voltages employed, and in this case measurements of the oscillograms of received voltage and of calibrating voltage will give a value of the received voltage independent of interference. The third source of error is due to the presence in the transmitted voltage of harmonics of the fundamental frequency. These harmonics are attenuated in transmission over the cable to a much greater degree than is the fundamental, so that they constitute only a small per-

[5] *Electrician*, Vol. 86, p. 272, 1921.

centage of the received voltage and are practically negligible in their effect upon the thermocouple.

Measurement of the Time of Propagation

The time of propagation of a steady state sinusoidal voltage over a loaded cable of transatlantic length is of the order of 0.3 second. It is measured by means of the circuit shown in Fig. 2, which is

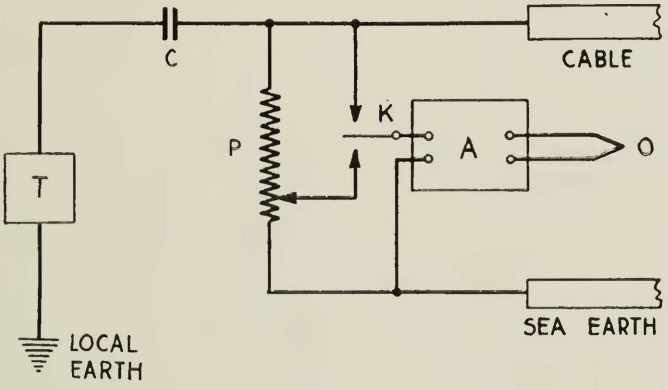


Fig. 2.

operated simultaneously at both ends of the cable. At each end a perforated tape is prepared which when inserted in the high speed transmitter T will cause a train of about ten reversals to be sent out over the cable. The potentiometer P is adjusted so that a measurable record of either transmitted or arriving trains, depending upon the position of the key k, will be obtained on the string oscillograph O after amplification by the vacuum tube amplifier A. The condenser C is inserted between the cable and transmitter in order to remove the low frequency components of the transient part of the train, which would otherwise overwhelm the steady state component at the distant end of the cable. The oscillograph, shown in Fig. 3, gives a continuous record of the current in a fine wire, which is free to respond to the interaction between the current and the strong magnetic field in which the wire is placed. The displacement of the wire, and hence the amplitude of current in it, is recorded on a long strip of sensitized paper, which is developed and fixed within the camera by a continuous process immediately after exposure. By this means it is possible to obtain a continuous record, over a period of several minutes, of voltages transmitted and received over the cable. A second wire can be used to give simultaneously a record of any other current which

may be desired for comparison. An arrangement is provided for superposing on the records vertical timing lines at intervals of one hundredth of a second. Short pieces of record are shown in Fig. 4.

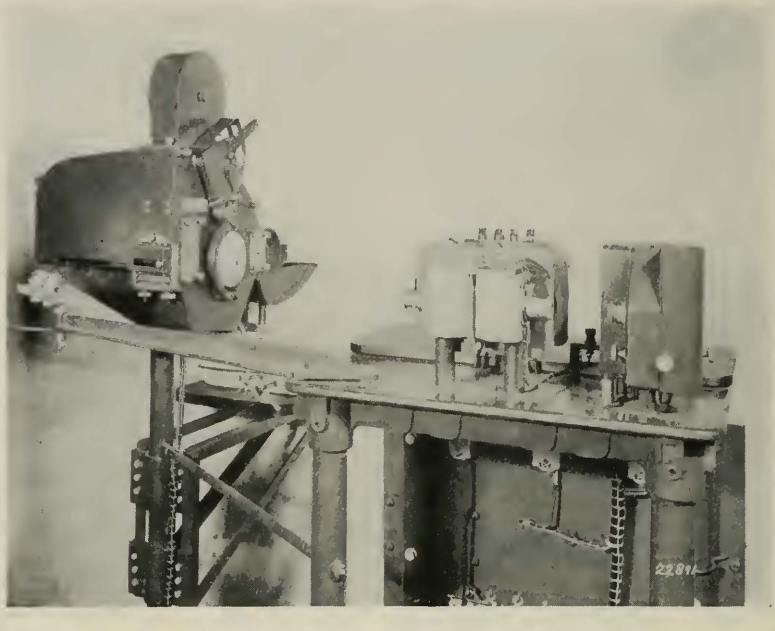


FIG. 3.

At a prearranged time the oscillographs at both ends are started. A train of reversals is transmitted from one end, a record being taken on the oscillograph at that end, and received at the distant end, where a record is also taken. Both stations quickly change potentiometer connections from send to receive or vice versa, and the distant station transmits a train of reversals, recording it on the same tape as was used for reception. Similarly at the first station the arriving train is

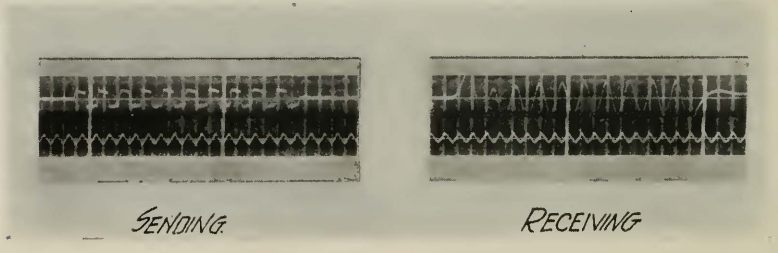


FIG. 4.

recorded on the strip containing the record of their transmitted train. Station 1 measures on its oscillogram the time elapsing between its transmitted train and its received train, and at Station 2 the time elapsing between the received train and the transmitted train is measured. After making suitable corrections, which will be described later, the difference between the interval measured at Station 1 and that measured at Station 2 will be equal to twice the time of transmission of the train of reversals over the cable.

A typical record such as would be obtained at Station 1 is shown in Fig. 4. It will be observed that in the record of received voltage the first few cycles are somewhat distorted because of the fact that the steady state has not yet been reached. Because of this fact the time of arrival or departure of a train is referred to a later cycle in the series, say the fifth. The times of departure and arrival of the various zeros following this cycle are measured, and the average of the values so obtained is defined as the time of arrival or departure of the train. In this way the possible errors due to interference or to distortion in the sent record due to improper functioning of the transmitter are eliminated.

It will be observed that, mainly on account of the presence of the condenser C, the voltage reversals impressed on the cable are not flat-topped and the zero phase of the fundamental component which we are measuring occurs somewhat ahead of the point in the transmitted voltage which we have used as the zero of reference in measuring the oscillograms. Since we are interested in the time elapsing between zero phase of the fundamental frequency in the transmitted voltage and the zero phase of the corresponding cycle in the received train, it is necessary to compute this interval, either by graphical analysis of the oscillogram or by computation from the constants of the circuit, and add the corresponding time to the time which has been measured.

Although the mechanical arrangement by which the timing lines are obtained on the oscillogram is adjusted as accurately as possible so that the interval between lines is very nearly one hundredth of a second, the very slight variations which occur in such a system are apt to introduce considerable error into the measurement of time of propagation. This is due to the fact that the time of propagation is obtained from the difference of two intervals each of which may be as much as ten times the time of propagation. An error in either interval will therefore result in a tenfold error in the final result. To guard against this condition a record is taken during the experiment of a periodic voltage obtained from a standard oscillator or fork, and the peaks of this oscillation serve as a check on the timing lines. As

a final check, records similar to Fig. 4 are taken with various times elapsing between reception and transmission at the second station. If an error exists in the timing arrangement its effect on the time of propagation will be greater the greater the interval between receiving and sending, and the time of propagation corresponding to negligible error in the timing system can be easily obtained by graphical methods. The error of measurement of the time of propagation is probably less than 1 per cent.

The inductance of a loaded conductor is an increasing function of current for the range of current values used in cable practice because of the increase of permeability of the permalloy, and since with finite transmitting voltage the current at the sending end may be quite large, the inductance of this portion of the cable under such conditions will be larger than the value it would have for very small current in the conductor. Accordingly the time of propagation at a given frequency will be a function of voltage. The value of inductance corresponding to very small current in the conductor can be derived from the time of propagation corresponding to zero transmitted voltage, which is obtained by extrapolation from measurements of the time of propagation at several values of transmitted voltage.

Measurement of Capacity

The dielectric capacity of submarine cables in the telegraph range of frequencies is in general comparatively insensitive to changes in temperature and hydrostatic pressure, so that it is possible to estimate this quantity rather accurately at various frequencies by means of measurements made in the factory, the factors required to reduce the results of the measurements to sea bottom conditions being relatively easy of determination. In order to check these values, however, the d.c. capacity of the laid cable is measured by the method of mixtures, employing a charging time of 10 seconds or more and a mixing time of equal duration.

Computation of Cable Parameters

The inductance of the cable can be computed at any frequency from the measured values of capacity and time of propagation by means of equation (8), proper allowance being made for the rather small effect of resistance.

Having computed the inductance and the capacity of the cable, only the resistance and the leakance remain undetermined. The direct current resistance can be computed from factory measurements and checked by measurement on the cable. The resistance component

due to eddy currents in the loading material can be computed from the resistance measurements obtained in the factory in the process of determining the inductance of sample core lengths. The eddy current resistance is proportional to the square of the product of frequency and permeability, and corresponding reduction factors must be employed in computing the eddy current resistance of the laid cable from the factory measurements. Since we are dealing with values of the parameters corresponding to very small current in the cable conductor, the hysteresis resistance is zero. In addition to the losses in the loading material there are other losses peculiar to continuously loaded cables due to currents induced in the cable structure. The loading material is ordinarily applied to the conductor in the form of a tape or wire of finite width, so that it has a definite lay, and since the magnetic flux in the loading material tends to follow the convolutions of the latter there is a component of this flux parallel to the axis of the central conductor. Consequently as the flux changes with signal current, electromotive forces are induced in those portions of the cable structure which link with it—the teredo tape and armor wires, for example. The resulting energy loss has in most practical cases comparatively small effect on the performance of the cable, and the magnitude of the corresponding resistance component can be estimated by theoretical methods and by measurements in the factory. The various components of resistance having been estimated, the total resistance at any frequency can be computed. Likewise the value of dielectric leakance of the laid cable at any frequency can be estimated from tests made during manufacture. These values of resistance and dielectric leakance should be considered merely as first approximations, since they are based in part on assumptions that cannot be directly verified.

Formula (2) is then employed to determine the effect upon the attenuation constant of departures from the approximate values of resistance and leakance, and by comparing these results with the measured values of attenuation constant, mutually consistent sets of values of resistance and of dielectric leakance can be computed at various frequencies. A choice of the best sets of values can then be made, due weight being given to the evidence available from computations and laboratory measurements regarding the manner in which these quantities vary with frequency.

From the curves relating the values of measured attenuation constant and the transmitted voltage, a check can be made of the method of computing the increase in attenuation due to hysteresis and to variation of inductance with current.[6] Since this method employs

[6] See Buckley, loc. cit., and U. Meyer, *E. N. T.*, Vol. 3, No. 1, 1926.

the inductance-current and resistance-current characteristics of the loaded conductor, as determined in the factory, the attenuation measurements also afford a check on these characteristics.

Conclusions

Measurements of attenuation, time of propagation and dielectric capacity of the laid cable at various frequencies, supplemented by measurements of eddy current resistance in the factory and by information regarding the manner in which sea return resistance and dielectric leakance vary with frequency are sufficient for determining the values of the four parameters of a loaded cable and for dividing the resistance into its component parts. A quantitative comparison of the results so obtained with the values of parameters that would be predicted from factory measurements alone would require a detailed discussion of the methods involved in such measurements, and is outside the scope of the present paper. A general conclusion that can be drawn from the results of measurements made on three cables of somewhat different characteristics is that the method of estimating the characteristics of laid cables from measurements made on short lengths of core during process of manufacture is capable of considerable accuracy. The values of inductance and dielectric leakance obtained from factory measurements are close enough to the actual values in the laid cable to give a value of attenuation constant within a few per cent of the actual value. The value of resistance obtained from the cable measurements appears to be about three to five per cent higher than the estimated value. This may in part be due to latent errors in measurement or in the method of allowing for the effect of reflections along the cable.

The greater part of the discrepancy between the estimated and measured values of resistance is perhaps due to erroneous assumptions involved in computing the value of sea return resistance employing the method described in the paper by Carson and Gilbert. In this work it was assumed that the cable is surrounded by a homogeneous medium, the sea water. For values of frequency higher than the telegraphic range this assumption appears to be sufficiently close to the truth, since only a comparatively small region around the cable plays any part in the phenomena. In the telegraph range, however, the return current is distributed through a comparatively large cross-section and more exact specification of the electrical characteristics of this region is required. To determine by rigorous methods the sea return impedance in the case where the cable lies in a plane separating

two different media is a problem of considerable difficulty. An approximate method, which gives results which are sufficiently accurate for purposes of cable design, consists in computing the combined impedance of the three parallel conductors, namely, the armor wires, the sea water, and the earth, the impedances of the latter two conductors being determined by the methods outlined in the aforementioned paper. The physical interpretation of the sea return resistance as obtained by this method is that the high value of reactance of the sea water and earth, due to the large cross-section of the conducting area, forces the return current to flow in the armor wires even though the resistance of this path is much higher than that of the paths through the sea water and earth. It appears probable that the electrical conductivity of the earth is very much less than that of sea water which would result in a larger cross-section of conducting area external to the armor wires and larger inductance of this path. This leads to higher values of sea return resistance than are obtained on the assumption that the cable is surrounded on all sides by sea water and thus gives a result more nearly consistent with the observed facts.

Automatic Printing Equipment For Long Loaded Submarine Telegraph Cables

By A. A. CLOKEY

SYNOPSIS: The introduction of the permalloy loaded submarine cable has presented the possibility of telegraph transmission at speeds several times those obtainable on non-loaded cables and has made practicable the operation of printer telegraph equipment. The present paper presents the various factors which affect the design of operating equipment and describes the apparatus which has been developed and used for a considerable period of time under service conditions. The transmission speed attained may exceed 2,400 letters per minute. To a certain extent, the detailed design of the terminal apparatus is controlled by the electrical characteristics of the particular cable to which it is to be applied and this type of equipment cannot, therefore, be completely standardized.

GENERAL

AT the time the development of the loaded submarine telegraph cable was undertaken, non-loaded cables were generally being operated duplex at signalling speeds ranging from 5 to 8 cycles per second (160 to 260 letters per minute) in each direction. The transmitting apparatus consisted of transmitters of the reciprocating contact type controlled by perforated tapes and the signals were received and recorded by the delicate moving coil type of amplifiers (generally referred to as magnifiers), relays and siphon recorders which produced a received signal record of such a character as to require the employment of highly skilled operators to translate and type the messages in final form. Except for a few trials, automatic printers had not been applied commercially to the operation of submarine cables, although the highly successful results which had been previously obtained with multiplex printing telegraph equipment on land lines coupled with the increasing demands made upon the cable systems as a result of the World War had directed the attention of telegraph and cable engineers to the need for applying automatic printing telegraph methods to submarine cables.

Preliminary studies of the characteristics of permalloy as a loading material for long telegraph cables indicated that, through its use, transmission speeds many times that of non-loaded cables could be readily attained. As the then existing apparatus was incapable of operation at the high speeds thus obtainable and the operating methods in use were not suited to handling the greatly increased volume of traffic over a single cable, it became apparent that new operating methods and equipment would have to be developed if the full ad-

vantage afforded by the use of permalloy loading [1] was to be realized. The development of the permalloy loaded cable was, therefore, paralleled by a study of the newly presented operating requirements and the development of suitable operating methods for high speed loaded cables. It is the purpose of this paper to present the various factors which affect the design of operating equipment for use on long loaded cables and to describe the apparatus and principles of operation which have been developed. The system which is to be described is similar to the multiplex system now in use on American land lines [2] but has been modified in several important respects in order to adapt it to the requirements of cable transmission.

The Cable and Amplifier as a Transmitting System

Submarine cables have heretofore been thought of as transmitting media which greatly distorted the signals and so reduced them in amplitude as to require the use of a sensitive siphon recorder for reception. The effects of signal distortion were, to a certain degree, compensated for by the addition of sending and receiving condensers, magnetic shunts at the receiving end, and, to a slight degree, by the inherent characteristics of the siphon recorder itself. A separate instrument termed a cable magnifier, of which there are several different types, was inserted between the cable and the siphon recorder to "magnify" the signal delivered to the recorder and thus partially offset the effects of attenuation. No two cables are identical as regards the distortion and attenuation of the signals and the means which will effectively provide for the correction and amplification of the signals on one cable will not necessarily be suitable for use on another cable of different length or construction. The apparatus provided for the correction of distortion in and amplification of the received signals is therefore an essential part of a signal transmitting system which includes the cable, and, except for the necessary switching arrangements, is independent of the means employed for impressing the signalling impulses upon the cable and for producing a permanent record of the corrected signals. Thus the development of terminal equipment for loaded cables comprised two separate and distinct developments, viz. the study of signal distortion and design of suitable signal shaping amplifiers (described in a separate paper by Mr. A. M. Curtis which appears elsewhere in this issue), and the development of apparatus for delivering signals to the system comprising the cable

[1] O. E. Buckley, "The Loaded Submarine Telegraph Cable," *Jour. A. I. E. E.*, June 26, 1925.

[2] J. H. Bell, "Printing Telegraph Systems," *Trans. A. I. E. E.*, Vol. 39, Part 1, 1920.

and amplifier and for converting the signals delivered by that system into a permanent printed record.

With the combination of cable and signal shaping amplifier, signals which are transmitted into the sending shaping network and the cable as square topped impulses as shown in Fig. 1 emerge from the amplifier

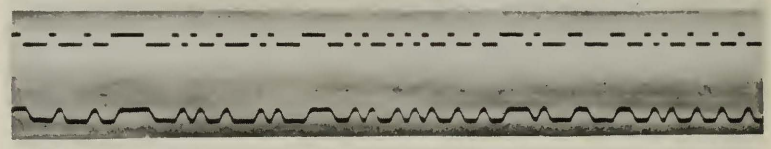


Fig. 1

as rounded impulses from which the high frequency components have been removed as a result of the attenuating effect of the cable. The receiving and printing system must therefore be capable of accurately translating rounded signals of this nature into printed characters.

Requirements of Operating System for Loaded Cables

The outstanding characteristic of the loaded cable is the enormously increased speed of transmission which may be as high as 2,400 letters per minute or more. For practical utilization of such high speeds the operating system must include some means for dividing the line time to provide a number of traffic channels. This is necessary in order to facilitate the distribution of the work of preparing the perforated transmitting tapes and checking the received message records among the required number of operators. The system must also provide for efficient two-way operation to avoid delay in the transmission of traffic from either terminal and should be capable of being joined with other cables or land lines through automatic repeaters to avoid the delay and expense introduced by manual methods of repetition.

As the shape of the received signal is determined by the cable-amplifier system and also by the character and amount of interference present which cannot be eliminated by the distortion correction networks, the operating system must be able to take the partly corrected signals delivered to it by the amplifier and accurately restore them to the form in which they were originally transmitted before using them to control the final recording mechanism.

The apparatus associated with loaded cables will in practically all cases be installed in the same offices as the equipment in use on non-loaded cables and, in order to avoid the necessity for duplicating the operating and maintenance staffs, it should be of such a nature as to permit of its being operated and maintained by men familiar with the operation of apparatus in use on land lines and ordinary cables.

Codes

The signalling speed attainable on any telegraph circuit, the effect of interference upon the received signals, and the design of the operating equipment depend to a certain extent upon the telegraph signal code used. A great variety of codes have been devised from time to time with a view to effecting greater economy of line time or greater freedom from the effects of interference, but only a few of them have been generally adopted in commercial practice. These may be divided into two general groups: the two-element codes which are composed of various combinations of positive and negative current impulses, of which the continental Morse and the Baudot codes are well-known examples, and the three-element codes in which a zero or no-current interval is employed to separate individual pulses of a group or as a third element in the combinations. The cable code and three-unit code are examples of the three-element type.

The codes in each of these two groups may be subdivided into two classes, those known as uniform codes in which all characters are composed of the same number of equal time units and those known as non-uniform codes in which the impulses forming the characters vary in length, number or both. The non-uniform codes are well adapted for use where the received signals are translated manually, but are not so well suited to automatic translation as the uniform codes on account of the mechanical and electrical complications introduced by the necessity for distinguishing between signal combinations of varying length.

Of the uniform codes which have been used in automatic printing telegraph systems, the Baudot or two-element five-unit type of code possesses advantages over the three-element three-unit type of code which make it much better suited for automatic operation of submarine cables. The three-element three-unit code employs, as does also the cable code, a zero interval of unit length in forming the signal combinations representing each character or letter and the shape of the received signals must be sufficiently refined to make this zero interval easily distinguishable (see *A*, Fig. 2) in order to prevent confusion in translation. Even with the best shaping obtained to date

on long non-loaded cables operated at high speeds the presence of this zero interval may be indicated only by a difference in slope of the recorded curve as shown in *B*, Fig. 2. Interference currents, which are present to some extent in all cables, are superposed upon the

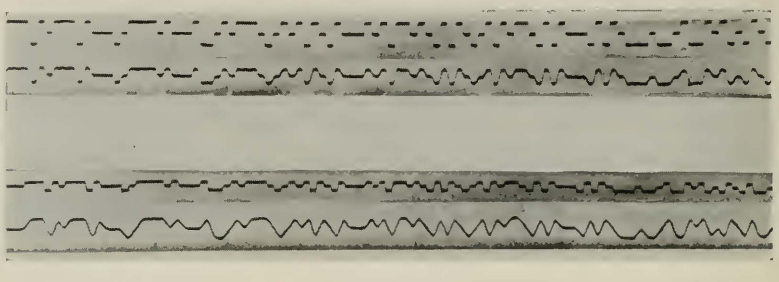


Fig. 2

received signals and cause troublesome distortion in the zero intervals and the length of the sustained pulses. The absence of these zero intervals of unit length in the two-element five-unit code, combined with the fact that only the middle portion of each received signalling impulse is used to operate the selecting mechanism of the printers, considerably reduces the effect of interference upon the accuracy of translation and makes it unnecessary to secure such refined signal shape.

The accurate evaluation, in terms of transmission speed, of the relative merits of the various telegraph codes is a highly complex problem which does not readily lend itself to solution through purely theoretical methods since it involves consideration not only of the total number of separate combinations which must be provided to represent the letters of the alphabet and all other characters to be transmitted, the frequency of occurrence in traffic of the various characters, and the average number of unit impulses required to form the combinations, but also depends upon the characteristics of the line or cable and the nature and distribution of the interference encountered and its effect upon the shape and definition of the received signals. The application of a code to any specific case also involves the more practical considerations of the type and operating characteristics of the apparatus employed. Practical experience therefore probably forms the best guide to the choice of a code.

In consideration of the conditions referred to above and the experience previously gained through the extensive use of the Baudot type of code on automatic telegraph circuits both in the United States and Europe, it was concluded at an early stage in the development that

the multiplex code used on American land line multiplex circuits would be the most suitable for high speed automatic submarine cable transmission. Subsequent experience has indicated that the original conclusion was amply justified and has shown that the Baudot type of code, when used in connection with terminal apparatus of suitable design, is probably faster than any of the other types of codes which have been considered for high speed loaded cable operation.

Outline of System

The multiplex system [3] used on land telegraph lines was in many respects well suited to the requirements of loaded cable operation. It was capable of operation at high transmission speeds, was more economical of line time than other methods which were considered, and provided for the division of traffic between a number of traffic channels in a manner which afforded great flexibility in the handling and routing of traffic and permitted the channel speeds to be fixed at values which would allow the operating staff to work at maximum efficiency. Its long continued use on land telegraph lines had resulted in bringing the apparatus, operating methods and routines to a high degree of perfection and the development of a thoroughly trained staff skilled in the operation and maintenance of the equipment, all of

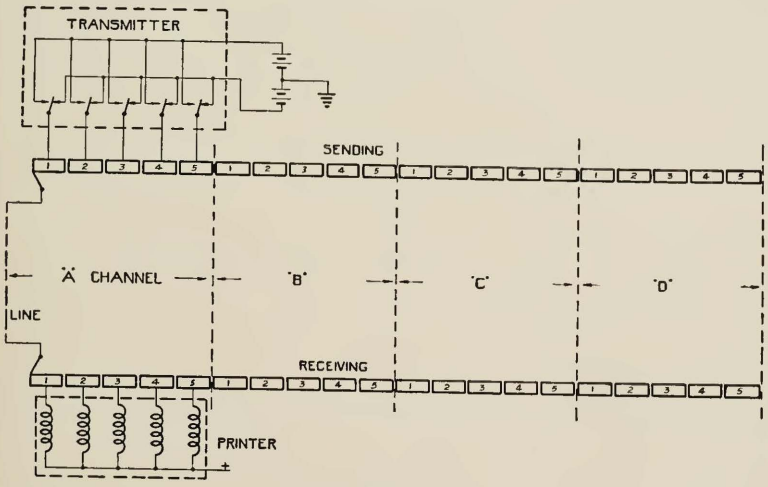


Fig. 3

which was of inestimable importance in the successful application and operation of printing telegraph methods to submarine cables.

The multiplex system provides for associating the line at the sending

[3] J. H. Bell, loc. cit.

end with each one of a number of transmitters in rotation by means of a rotating brush which passes over a segmented commutator to which the transmitters are connected as illustrated in Fig. 3. In this figure the commutator segments are shown developed for sake of simplicity. At the receiving end, the line is similarly associated in rotation with each one of a corresponding number of printers by means of the receiving brush and commutator. The commutator brushes at the two ends of the line are maintained in nearly exact synchronism by short correcting impulses which are derived from reversals in the received signalling currents, and their phase relation is such that each of the five segments connected with the "*A*" channel transmitter will be connected in rotation through the line to the corresponding segments of the "*A*" channel printer once during each revolution of the brushes, and the impulse transmitted from any one sending segment will pass through the corresponding receiving segment and operate the printer selector magnet which is connected to it. Similarly the transmitter on each of the other channels will be connected to its corresponding printer once during each revolution of the brushes. The commutators and the associated brushes together with the mechanism provided for correcting the phase relation of the brushes are usually referred to as distributors.

In the operation of this system a transmitting tape is prepared in which the characters to be transmitted are represented by combinations of holes perforated in the tape by means of a keyboard perforator which resembles a typewriter. The tape thus prepared is drawn through a transmitter which is arranged to apply to its associated distributor segments, positive and negative battery in the proper combination to form the five-unit impulses corresponding to the perforations in the tape. The received signal combinations control the operation of an automatic telegraph typewriter or printer which converts the signals into printed characters. Detailed descriptions of the perforating, transmitting and printing apparatus and the various methods for maintaining synchronism used in the multiplex system are given in the paper by Mr. J. H. Bell, previously referred to, and also in an excellent book by Mr. H. H. Harrison entitled "Printing Telegraph Systems and Mechanisms."

On account of the many advantages which this system embodied, it was chosen in principle as a basis for the development of the new system, although in several important respects much of the apparatus and operating methods employed were entirely unsuitable for loaded cable operation. The multiplex had been employed almost entirely in the operation of duplexed circuits and therefore was not applicable to

simplex operation of cables. The character of the received signals and interference on long cables is such as to require the use of entirely different methods of reception in order to utilize the line time most efficiently, and the higher transmission speeds expected on cables necessitated departure from standard land line practices in the matter of apparatus design and number of channels employed. The system as finally developed embodies the following important improvements over previous methods.

1. An entirely automatic means for quickly reversing the direction of transmission on a simplex circuit at short intervals which can be altered as required to accommodate varying traffic loads in the two directions.

2. A synchronous vibrating relay which corrects for the residual distortion in the signals delivered by the amplifier, and practically doubles the speed of transmission.

3. A high degree of precision and refinement in the design and construction of apparatus which is justified by the great cost of the cable relative to that of the terminal apparatus.

The inclusion of these improvements in a modified multiplex system involved, of course, the solution of a number of important incidental problems such as the provision of Morse "talking circuits" which could be made instantly operative, and the development of suitable arrangements for linking two simplex cable sections together through repeaters.

Two Directional Working

The use of duplex methods in the operation of non-loaded cables enables communication to be carried on simultaneously in both directions and usually effects an increase of from 60 to 90 per cent in the total traffic capacity of the cable. As only a moderate capital expenditure is required to equip a non-loaded cable for duplex operation practically all cables of this type are now equipped in this way as a matter of economy. The characteristics of the loaded cable, however, are such as to require the use of highly complicated and extremely expensive artificial lines and balancing equipment for duplex operation and it is quite doubtful whether the total duplex traffic capacity thus secured would equal that obtainable by the use of simplex methods. Duplexing the loaded cable therefore appeared to afford no certain economic gain over simplex operation and the extremely high cost of duplexing could hardly have been justified merely for the sake of securing simultaneous transmission in both directions.

The apparatus and methods formerly employed for reversing the

direction of transmission on manually operated simplex cables were so time-consuming that it was impracticable to reverse direction oftener than once every quarter or half hour. The delay in transmission which would result from the adoption of the older methods could not be permitted on the loaded cable and it therefore became necessary to develop special apparatus for automatically reversing the direction of transmission at comparatively short intervals in order to approximate simultaneous transmission in both directions and reduce traffic delays to an absolute minimum.

The design of suitable switching arrangements which would permit stopping transmission on a long cable operated with multichannel printing equipment and almost immediately starting transmission in the opposite direction presented several difficult problems. On account of the lack of uniformity in the lengths of the messages to be transmitted and the number of channels employed, it rarely happens that the transmitters on all channels complete the transmission of their respective messages at exactly the same instant, therefore it was necessary to arrange for making the change in direction of transmission at more regular and frequent intervals even though the transmitters on all channels had only partly completed the transmission of their respective messages at the time the change was made. To accomplish this without introducing any errors or other evidence of the interruption into the final printed message necessitates first stopping the transmitters on all channels at precisely the right instant, then allowing an interval equal to the time of signal propagation over the cable to elapse before cutting off the printers at the distant end, and finally upon resumption of transmission in that direction starting all of the transmitters and printers at the proper time and in the correct sequence to avoid the loss, repetition, or mutilation of any character.

The last signals transmitted into the cable before changing to the receiving position result in leaving the cable charged to a potential which would paralyze or "block" the amplifier were it to be immediately connected. Part of this charge must be dissipated and the current due to the residual charge and the presence of any interference or earth currents must be allowed to attain its steady value in the shaping network and input transformer elements of the amplifier before connecting any of the actual amplifying elements to the cable. The switching operations involved in applying the amplifier to the cable must be effected in the proper sequence and at precisely timed intervals in order to leave the amplifier in the proper condition to avoid mutilation of the first signals received from the distant end.

The required degree of accuracy in timing the various switching operations involved in reversing the direction of transmission was secured by utilizing the rotating shafts of the distributors at both stations to control a timing mechanism which determined the lengths of the transmission intervals in the two directions.[4] This timing mechanism is essentially an electrical revolution counter which can be set to count any desired number of revolutions of the distributor shaft and close within a fraction of a revolution of that number the circuit which controls the operation of the various contacts which do the actual switching. As the distributor shafts at the two ends of the cable are maintained in exact synchronism in a manner previously described, the timing mechanisms will therefore also operate in synchronism, and if at the time of setting up the circuit they are started in the proper phase relation the correct phase relationship will be maintained as long as the operation of the circuit continues without interruption. The timing mechanisms are driven from the distributor shafts through the medium of an electrically operated clutch which when disengaged permits the timing mechanisms at all stations to be manually set in their proper positions and started together in this relationship by means of a starting impulse sent over the line which causes the clutches to engage.

In order to provide for transmission intervals of various lengths in the two directions, the timing mechanism includes a number of timing elements each representing a different division of the line time, any one of which can be quickly selected at will by the movement of an indicating lever to control the length of the transmitting and receiving periods.

Upon the completion of the predetermined number of revolutions of the distributor the timing mechanism operates a direction control relay, see Fig. 7, the contacts of which are arranged to operate and cut off the transmitters, discharge the cable, and connect the amplifier and the printers in properly timed sequence. The actual time consumed in making all of the circuit changes necessary to reverse the direction of transmission, measured from the time of transmission of the last signal combination to the time of printing the first character on the printer at the same station, is of the order of five seconds but will vary somewhat on different cables according to the length of the cable and the magnitude and character of the interference and earth currents encountered.

During the interval in which the actual switching operations are taking place no signals are being transmitted in either direction so

[4] A. A. Clokey, U. S. Patent No. 1,601,941.

that neither of the distributors will receive any correction impulses and as a result the sending and receiving brushes may depart considerably from their normal phase position. This would cause errors to occur in the first signals received upon resumption of transmission if means were not taken to bring the brushes back into proper phase relationship before the transmission of actual signals was begun. This is provided for by arranging to have the distributors transmit, at the close of each switching period, a number of "spacing" signals which do not affect the receiving printers since they are not connected in circuit until a sufficient number of reversals have occurred in the line current to correct the receiving brush into the proper position. The transmission of signals which must be recorded by the printer is then started.

As the length of the interval allowed for these switching operations is determined by a definite number of revolutions of the distributor shaft, which may be set to rotate at various speeds, the gearing between the distributor shaft and the timing mechanism is designed to allow for a five-second switching period when the distributor is rotating at a speed which corresponds to the maximum transmission speed of the circuit.

Although this system lacks the advantage of absolutely continuous communication in both directions, it possesses another feature which goes far toward offsetting, if it does not entirely outweigh, the advantages afforded by the duplex method. Almost all of the long cables of the world run in an east and west direction and the difference in time between the terminal stations of those cables results in an unequal distribution of traffic in the two directions except perhaps during a comparatively short time each day. The provision of the selective timing mechanism permits the total traffic capacity of the cable to be divided between eastward and westward transmission in about the same proportion as the eastward traffic load bears to the westward load and thus permits efficient utilization of the entire traffic carrying capacity of the cable.

The Synchronous Vibrating Relay

The vibrating relay principle was first suggested by Gulstad [5] who applied it to short cables for overcoming the effects of distortion. As originally used, it consisted of a sensitive polarized relay provided with a line winding, upon which the received signals were impressed, and two auxiliary windings included in a local vibrating circuit adjusted to cause the relay armature to vibrate continuously when

[5] K. Gulstad, "Vibrating Cable Relay," *Elec. Rev.*, London, Vol. 42, 1898; Vol. 51, 1902.

the line winding was de-energized. The rate of vibration was adjusted to be approximately the same as the frequency of the transmitted signals and the amplitude of the vibrating current was adjusted to be approximately equal to the received signalling current so that the latter, if of one polarity, would neutralize the effect of the vibrating impulse and prevent the movement of the relay armature and if of the same polarity would aid the vibrating impulse. The effect of this combined action of the vibrating and received signalling impulses is to reproduce, in the local circuit, signals of approximately the same shape and duration as the original transmitted signals.

The frequency attenuation characteristic of a system comprising a long telegraph cable and its signal shaping amplifier and networks when the latter are adjusted for the maximum transmission speed is such as to cause the impulses of unit length, which represent half cycles of the fundamental signalling frequency, to be received in considerably smaller amplitude than the impulses of two units (or more) length which represent half cycles of one half (or less) the fundamental signalling frequency.[6] The highest signalling speed obtainable on a given cable is therefore determined by the length of the shortest impulses which must be received in sufficient amplitude to exercise control over the receiving apparatus and at that speed the two-unit and longer impulses will be received in much greater amplitude than is necessary for operation of the receiving apparatus. Gulstad pointed out [7] that as the received impulses of unit length always occur in the proper direction to aid the vibrating impulses they may therefore be greatly reduced in amplitude without impairing the accuracy of reception. On account of this fact the speed of signalling may be increased to a point where only the two-unit and longer impulses are received in sufficient amplitude to overcome the effect of the locally generated vibrating impulses and control the movement of the relay armature. At this increased speed the impulses of unit length will be either greatly diminished in amplitude or entirely removed by the attenuating effect of the cable and at such times the armature of the vibrating relay will be operated by the locally produced impulses.

As the rate of vibration of the Gulstad relay was determined entirely by the values of the resistances and capacities in the local vibrating circuit, the vibrations of the relay armature did not exactly coincide either in frequency or phase relation with the signals sent by the distant transmitter so that complete restoration of the incoming signals to their original form was impossible and full advantage of the speed

[6] The fundamental signalling frequency is defined as the fundamental frequency of a train of alternate positive and negative impulses of unit length.

[7] K. Gulstad, loc. cit.

possibilities of the device could not be realized. For these reasons its use was limited almost entirely to comparatively short cables where the strength and shape of the received signals were sufficiently good to control the relay directly with only a small improvement in shape and with no amplification. The original arrangement was later modified [8] to adapt it to the operation of longer unloaded cables.

One of the principal features of the cable multiplex herein described is the synchronous vibrating relay [9] which was developed particularly for high speed operation on long cables, and is a great improvement over the Gulstad device. The vibrating impulses, instead of being derived from an adjustable vibrating circuit, are generated by a segmented commutator located on the receiving head of the distributor. As the brushes on the receiving head of the distributor rotate in nearly exact synchronism with the transmitting brushes, it is evident that the rate of vibration of the relay will coincide exactly with the frequency of the transmitted signals and by properly adjusting the angular position of the vibrating segments, the time of closure of the relay contacts with respect to the incoming signalling impulses can be accurately fixed. The accuracy with which the missing impulses of unit length in the received signals are reinserted by this means makes it possible to realize the full speed possibilities of the vibrating relay principle and obtain faithful reproduction of signals on a given cable at almost double the speed obtainable through the use of ordinary non-vibrating relays.

Another important advantage gained through the use of the synchronous vibrating relay is greater freedom from the effects of extraneous interference. The amplitude of the received signals is sufficiently great to permit of its being reduced by the effects of interference to approximately half of the normal value before the distortion becomes sufficiently great to cause errors in printing. Likewise interference occurring during the zero intervals in the received signals must attain a value of approximately half of the normal received signal amplitude before causing errors. Interference occurring during the intervals between vibrating impulses will, of course, produce no effect upon the relay unless the amplitude of the interference attains a sufficiently large value to operate the relay directly. This ratio of interference to received signal amplitude represents the absolute limit of operation and some margin must obviously be allowed. It has been found that continuous satisfactory operation can be maintained so long as the interference does not

[8] W. Judd, British Patent No. 9,768, April 25, 1913; G. R. Benjamin and Herbert Angell, U. S. Patent No. 1,579,999, April 6, 1926.

[9] A. A. Clokey, U. S. Patents Nos. 1,521,870 and 1,522,865.

exceed one third of the normal signal amplitude. The presence of a proportionate amount of interference in the received signals in ordinary cable code operation would cause the recorder record to be so mutilated as to render it entirely illegible.

A detailed description of the operation of the synchronous vibrating relay is given in the appendix.

Refinement

The extreme speed at which the apparatus must operate to utilize the entire capacity of a loaded cable precludes making any adjustments while in operation and the importance of maintaining uninterrupted service for long periods demands that the apparatus shall be absolutely reliable in its operation and as free as possible from any variation in adjustments which would require occasional correction. This degree of reliability is secured through a refinement in mechanical and electrical design and a precision in construction which might be considered uneconomical for ordinary land line operation. The extra expense incurred in the design and construction of such highly refined apparatus is well justified by the resultant large increase in traffic capacity of the cable and the small cost of even the most refined apparatus relative to that of the cable on which it is used.

A general idea of the type of apparatus and construction used can be gained from one of the terminal distributors which is illustrated in Fig. 4. The greatest permissible variation in the phase relation

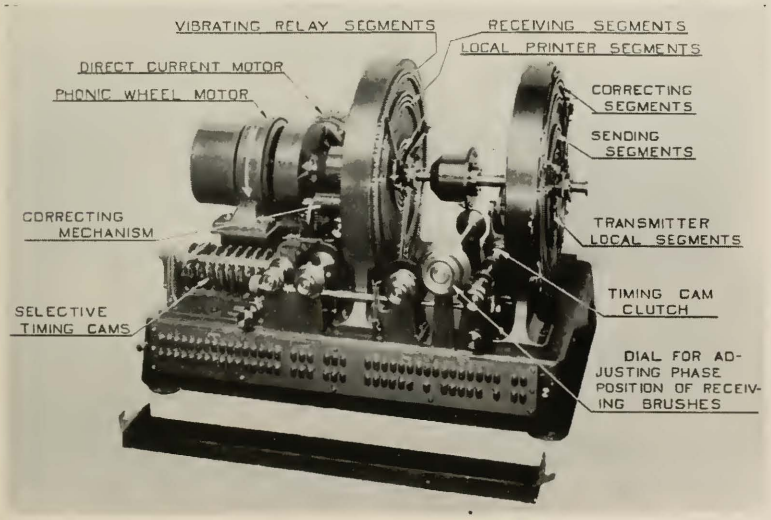


Fig. 4

between the brushes on the distributors at the two ends of the cable is only about one and one half degrees of revolution and in order to hold within this limit it was necessary to design a driving unit in which the phase shift, resulting from variations in the line voltage, was reduced to a minimum, and to arrange the gearing and coupling between the driving motor shafts and the various rotating brush arms so as to reduce to a minimum any lost motion or back lash. The driving unit consists of two motors: the one which supplies the power for driving the brushes is a dynamotor in which the DC side is used as a motor to supply the power and the AC side is included in a circuit with an electrically driven tuning fork which controls the motor speed within very close limits; and the other is a phonic wheel or La Cour motor driven from the same driving fork. This motor normally supplies little if any power for driving the distributor but by increasing or decreasing the load it prevents the occurrence of any appreciable phase shift in the DC motor due to variations in the driving voltage. In order to prevent slight shifting in the phase of the brushes due to vibration and axial twisting in the shafts and gears, it was necessary to employ much heavier construction in the rotating parts than is actually required to transmit the small amount of power used. The cutting of the gears, the distributor segments, and timing cams was done with the utmost precision to eliminate mechanical errors. The distributor segments included in the vibrating relay circuit are heavily faced with coin silver to reduce variation in the resistance of the contact between them and the rotating brushes.

The satisfactory operation of the system depends upon the accuracy with which the various relays in the system follow and repeat the signals. None of the available types of relays were found to be sufficiently reliable to permit of use in the system and it became necessary to develop for the purpose a new type of high speed relay shown in Fig. 5. The size and inertia of the parts comprising the moving system of this relay were reduced as much as possible in order to secure quick response and freedom from contact chatter at the highest operating speeds. A magnetic circuit was designed in which the effects of magnetic hysteresis are practically negligible, which results in the relay always operating upon the same value of current irrespective of its previous magnetic history. Permanency of adjustment, which is essential in relays used in this class of service, was obtained by adhering to standards of accuracy and precision of manufacture heretofore considered unnecessary in relay construction. The accuracy with which relays of the new type will operate at high speeds and the entire freedom from contact chatter is illustrated in

the oscillogram reproduced in Fig. 6, which shows 200- and 600-cycle sine waves applied to the relay windings and the character of the reversals repeated by the contacts. At these and lower frequencies

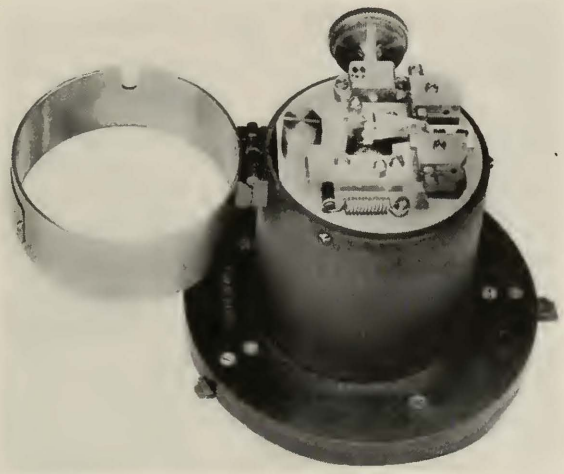


Fig. 5

the adjustment of the relay is sufficiently stable to permit of its being operated continuously for long periods without requiring readjustment or other attention.

Apparatus of this nature is frequently installed in isolated stations where materials or parts needed for making repairs cannot be obtained promptly, and the climatic conditions at some of these stations often

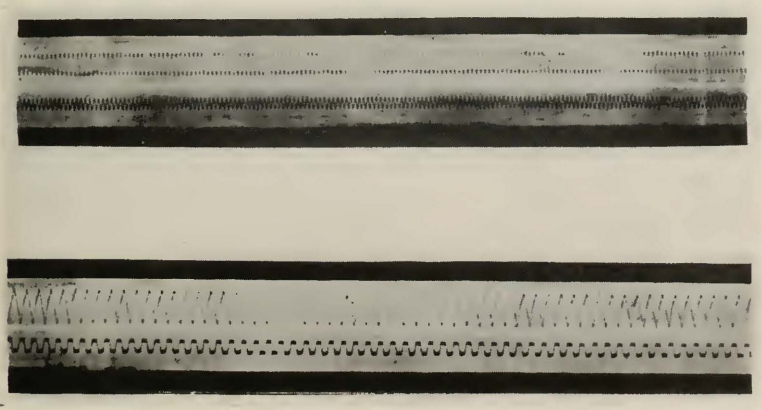


Fig. 6

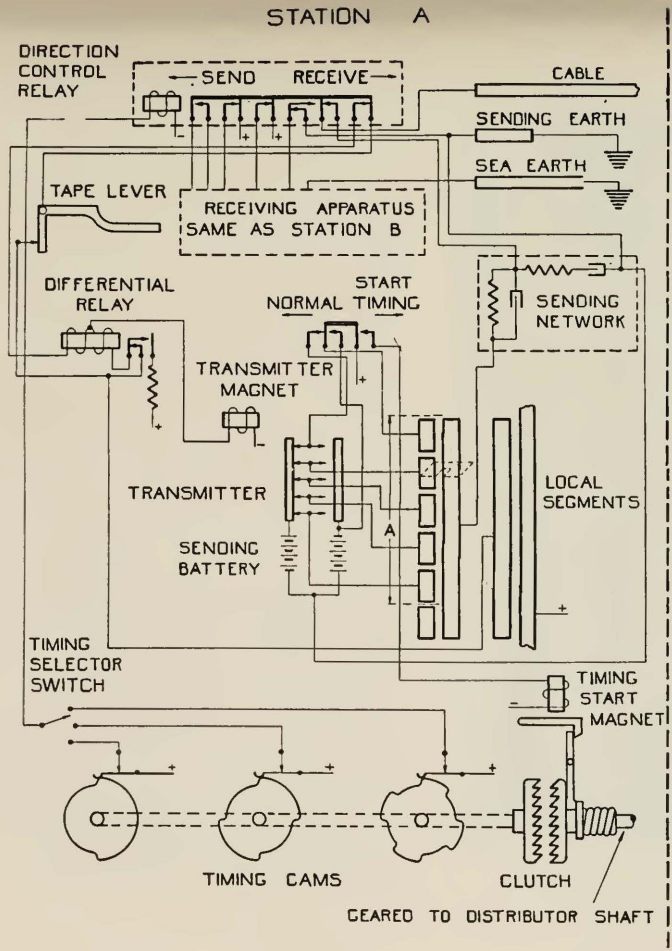
impose quite severe requirements upon the mechanical as well as the electrical portions of the apparatus. In designing the apparatus, the greatest care was therefore exercised in selecting, for the construction of even the smallest details, materials which would withstand the most severe usage and be unaffected by the most severe climatic conditions.

Summary of System

The inclusion of these newly developed features in the multiplex system and the application of the modified system to a long loaded cable presents a number of interesting aspects and new possibilities. The entire system is shown schematically in Fig. 7.

The use of an amplifier containing no mechanical moving parts, in which all adjustments are made by alteration of the constants of electrical circuits, makes it possible to determine at the time of installation the proper amplifier adjustments to give satisfactory signal shape at a number of different transmission speeds and thereafter the amplifier may be quickly set for any speed by duplicating the adjustments that were previously found suitable for that speed. As the operation of the correcting relays and circuits and the vibrating relay depends to some extent upon the shape of the signals delivered by the amplifier, the ability to reproduce accurately a signal shape which has been previously found satisfactory is of considerable importance in the operation of the system.

Although the amplifier and shaping networks are considered a part of the cable system rather than an element in the transmitting and receiving system, their operation must be controlled by the direction control switching mechanisms. The relays included in the direction control system which switch the amplifier circuits are built in the amplifier to simplify wiring and maintenance. The speed of the distributors is controlled as in the multiplex system by vibrating tuning forks, but in order to secure under certain conditions greater stability and freedom from speed variations due to alteration in the fork contact adjustment and changes in room temperature and voltage of the power supply there was developed a constant temperature vacuum tube driven fork. The distributor, with its driving fork, the relays included in the direction control and vibrating relay circuits, and the apparatus usually provided in land line multiplex equipments for phasing and lining up the circuit, including the Morse talking circuit, are mounted in accessible positions on a table which is separated from the operating tables on which the printers and transmitters are located.


STATION A
DIRECTION CONTROL RELAY
SEND
RECEIVE
CABLE
SENDING EARTH
SEA EARTH
TAPE LEVER
RECEIVING APPARATUS SAME AS STATION B
DIFFERENTIAL RELAY
START NORMAL TIMING
SENDING NETWORK
TRANSMITTER MAGNET
TRANSMITTER
LOCAL SEGMENTS
SENDING BATTERY
TIMING SELECTOR SWITCH
TIMING START MAGNET
TIMING CAMS
CLUTCH
GEARED TO DISTRIBUTOR SHAFT

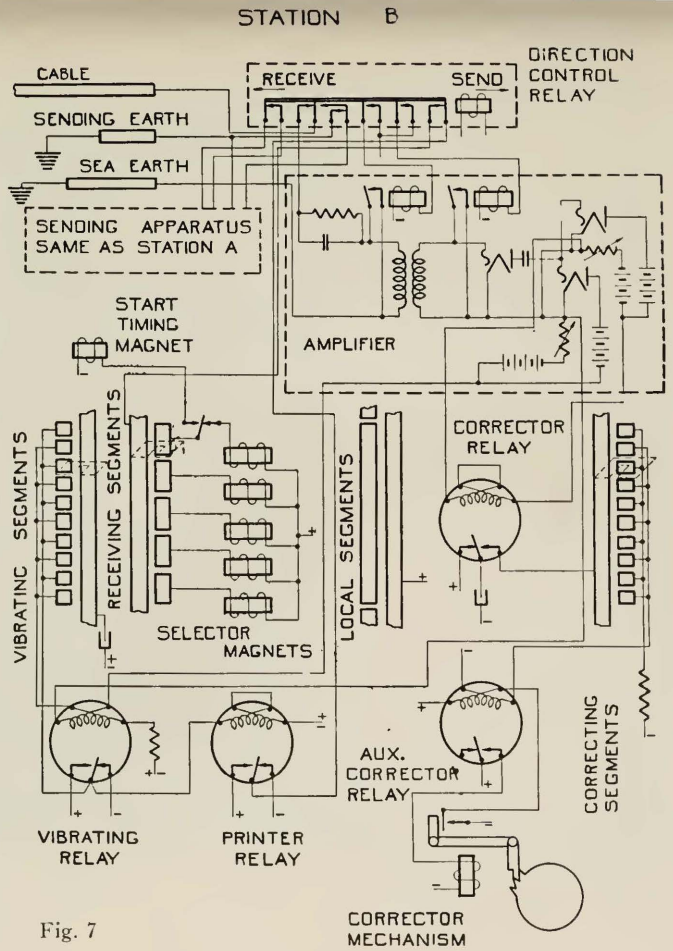

STATION B
CABLE
SENDING EARTH
SEA EARTH
RECEIVE
SEND
DIRECTION CONTROL RELAY
SENDING APPARATUS SAME AS STATION A
START TIMING MAGNET
AMPLIFIER
VIBRATING SEGMENTS
RECEIVING SEGMENTS
SELECTOR MAGNETS
LOCAL SEGMENTS
CORRECTOR RELAY
CORRECTING SEGMENTS
AUX. CORRECTOR RELAY
VIBRATING RELAY
PRINTER RELAY
CORRECTOR MECHANISM

Fig. 7

The adaptation of the multiplex to cable operation does not involve any modifications which affect the design of the perforators, transmitters, or printers, so that it is possible to employ in this system the same type of instruments used in land line operation.

In cases where it is desirable to link two cable sections together automatic repetition is provided for by the provision of additional sending and receiving commutators on the repeater distributor for transmitting and receiving on the second section. A photograph of such a repeater distributor is shown in Fig. 8. The incoming signals

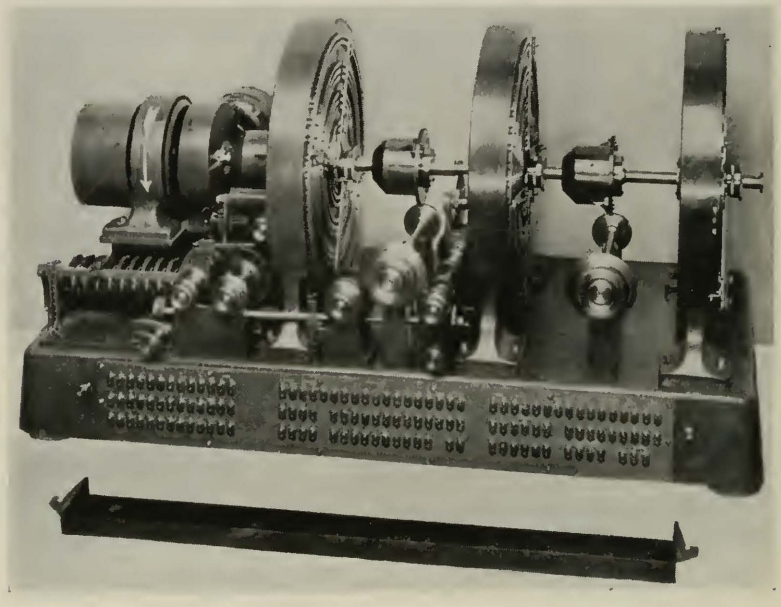


Fig. 8

at the repeater are received in the regular way and operate the vibrating relay which applies the completely corrected impulses to the receiving segments. The receiving segments, instead of being connected to the selector magnets of a printer, are connected to the windings of storing relays which are operated by the incoming signal combinations and set up the identical combinations on the corresponding sending segments associated with the next section of cable. The storing relays thus perform the same function as a tape transmitter except that they are controlled directly by the received signals instead of a perforated tape. With this method [10] of repetition it is possible to replace the storing relays on any channel with a printer and trans-

[10] E. P. Bancroft, et al., U. S. Patent No. 1,541,316.

mitter on each section so that one or more channels on both cable sections may be terminated at the repeater station without interfering with automatic repetition of traffic on the remaining channels.

Not only is it possible to link two or more simplex cable sections together through automatic repeaters, but it is also possible to link such a system through repeaters with a duplexed land line multiplex without introducing serious complications. The printer on the receiving side of each channel of both the land and cable circuits at the repeater station may be replaced with an automatic reperforator which will prepare from the incoming signals a perforated tape for retransmission. As this tape leaves the reperforator it is automatically drawn through a standard transmitter which will transmit the signals into the corresponding channel of the next section of line or cable. In moving between the reperforator and transmitter the tape passes under a contact closing lever arranged to stop the operation of the transmitter when the slack in the retransmitting tape drops below a predetermined minimum as the result of a difference in transmission speed on the two sections or the stoppage of the reperforator on the simplex section during the transmitting periods. This avoids the possibility of mutilation of the transmitted signals or tearing the tape.

The provision of a comparatively large number of traffic channels and automatic repeaters by means of which traffic on any or all channels may be automatically repeated into the other cable sections or land telegraph lines affords a high degree of flexibility in handling and routing traffic and permits the several channels to be terminated at the two ends in widely separated points.

Conclusion

Although the general principles of the system and the general design of the apparatus described herein are applicable to all loaded cables irrespective of length or construction, it is quite obvious that the detailed design of the various pieces of apparatus required will be determined to a great extent by the electrical characteristics of the particular cable to which they are to be applied and by the operating and traffic requirements which that system must fulfill. Equipment of this type can not therefore be standardized to the degree possible in the case of similar equipment for land line service, and the provision of apparatus for each cable becomes a special engineering problem which must be worked out with the cooperation of the engineers of the operating company in order to make the apparatus capable of satisfactorily meeting all of the conditions which will obtain in subsequent commercial use.

A complete operating equipment embodying the general principles described has been designed with the cooperation of the engineers of the Western Union Telegraph Company for the New York-Azores permalloy loaded cable and has been in actual commercial operation for many months. Provision has been made in the design of this apparatus for the extension of the circuit to Emden, Germany, over the Azores-Emden cable of the Deutsch-Atlantische Telegraphengesellschaft, and automatic repeaters for the Azores station and terminal equipment for the Emden station have been installed and are now undergoing tests preliminary to the establishment of through-operation between New York and Emden.

Appendix I

Synchronous Vibrating Relay

There are several methods which may be employed to obtain the synchronous vibrating feature, one of which is shown schematically in Fig. 9. The relay is of the polarized type having two separate

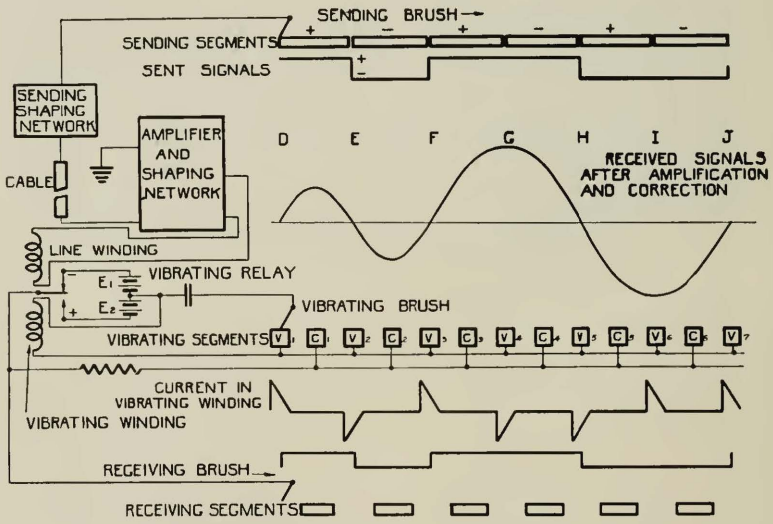


FIG. 9

Fig. 9

windings, one of which, termed the line winding, is connected directly in the output circuit of the amplifier, and the other, or vibrating winding, is included in a circuit comprising the vibrating condenser,

the vibrating segments V and the vibrating brush of the distributor. The distributor segments are shown developed for the sake of clearness. Disregarding the line winding for the moment, the passage of the vibrating brush over segment C_1, when the relay armature is resting against the negative contact, causes the vibrating condenser to be negatively charged by the battery E_1, and as the brush continues its rotation and passes upon segment V_1, the charged condenser is disconnected from the charging circuit and is connected to the vibrating winding through which it immediately discharges in the proper direction to cause the relay armature to be moved against its opposite or positive contact. This change in the position of the relay armature connects all of the "C" segments to positive battery, so that the passage of the brush over segments C_2 and V_2 in succession causes the condenser first to be positively charged, then to discharge through the vibrating winding which restores the relay armature to its former position against its negative contact. This cycle of operations will be repeated as long as the brush rotates and the rate of vibration can be made to coincide exactly with the frequency of the transmitted signals by suitably arranging the vibrating brush so as to be corrected from the incoming signals in a manner similar to that employed in the standard multiplex system.[11] The armature of the vibrating relay, in addition to controlling the polarity of the charge upon the vibrating condenser, controls the polarity of a battery applied to the receiving brush which distributes the received and corrected signalling impulses to the selector magnets of the receiving printer. In practice an intermediate relay, not shown in the figure, is employed between the armature of the vibrating relay and the receiving brush.

The line winding of the vibrating relay is connected in the amplifier circuit in the direction which will cause its armature to move toward its positive contact in response to incoming signalling impulses of positive polarity and vice versa, and as the amplitude of the current in that winding is adjusted to be approximately equal to that of the vibrating impulses, the effect of impressing the amplified and partially corrected positive impulse D upon the line winding at the time the vibrating brush is passing over segment V_1 would be only to aid the condenser discharge current in reversing the position of the relay armature. Every received impulse of unit length and the current during the first interval of unit length in every sustained pulse will produce the same result as shown at D, E, F and H, but the effect of current in the line winding due to the second and all succeeding time units of every sustained pulse will be neutralized, as shown at G, I

[11] J. H. Bell, loc. cit.

and J, by a pulse of approximately equal amplitude and opposite polarity in the vibrating winding and the position of the relay armature will therefore remain unchanged. Thus as the relay is actually operated by the energy supplied by the locally generated vibrating impulses and the received signalling current is employed only to neutralize the effect of the vibrating impulses during the second and succeeding time units of the sustained pulses, a considerable amount of distortion may be present in the signals without causing errors in reception, in fact the impulses of unit length may be entirely missing without interfering in any way with the accuracy of reception.

Ordinary polarized relays not provided with the vibrating feature are operated by the energy supplied by the amplified signals, which must consequently be quite free from distortion in order to insure faithful reproduction of the transmitted signal. The speed of transmission on long submarine cables employing non-vibrating relay reception cannot exceed that frequency at which the received signalling impulses of unit length are reduced by the attenuation of the cable to an amplitude which is only enough greater than the interference currents present to cause positive operation of the relay. The vibrating relay, however, does not require the pulses of unit length to be present at all, so that its limiting speed is that at which the impulses of two units length are received in just sufficiently large amplitude to prevent the relay from vibrating and allow a reasonable margin for overcoming the effects of any interference currents that may be present. This limiting speed is approximately double that obtainable with the use of non-vibrating relays.

The Application of Vacuum Tube Amplifiers to Submarine Telegraph Cables

By AUSTEN M. CURTIS

SYNOPSIS: Vacuum tube amplifiers have been developed for use in submarine telegraph reception and at present are in successful operation on four high speed permalloy cables. There is no limit to the speed at which vacuum tube amplifiers may be operated and in the present stage of development, the rate at which messages may be passed over loaded cables of the length used in the Atlantic Ocean is determined by the cable itself and the mechanical transmitting and receiving apparatus. In regard to maintenance, vacuum tube amplifiers have a great advantage in that they do not require any delicate mechanical adjustments.

THE laying of the new permalloy loaded cable between New York and Fayal (Azores) in September 1924 marked the most radical change in construction and operation of submarine cables that has taken place in many years. During 1926 three additional cables of this type were laid, the four sections being arranged to provide a line of communication between New York and England and another between New York and Germany. The traffic handling capacity of these cables when ultimately developed to its maximum by suitable terminal apparatus, will be nearly equal to that of all the other cables between North America and Europe combined.

The construction of these cables and the principles underlying their operation have already been described by O. E. Buckley[1] before the American Institute of Electrical Engineers in June 1925. The speed of operation of loaded cables of this type is many times that of the older non-loaded cables, and new apparatus has had to be developed to realize the full advantage offered by permalloy loading. One of the most important of these new developments has been the signal shaping vacuum tube amplifier, which is now in use on the four North Atlantic loaded cables.

The purpose of this paper is to point out the requirements which must be met by cable amplifiers, particularly those used on high speed loaded cables, and to describe how these requirements have been fulfilled in the present signal-shaping amplifier of the Western Electric Co. It will not be necessary to consider in any detail the general principles of operation of telegraph cables as these have been discussed with reference to non-loaded cables by Mr. J. W. Milnor[2] and are

[1] *Bell System Technical Journal*, Vol. 4, July 1925; *Journal of the American Institute of Electrical Engineers*, Vol. XLIV, No. 8, August 1925.

[2] "Submarine Cable Telegraphy," *Journal of the American Institute of Electrical Engineers*, Vol. 41, February 1922.

considered with particular reference to loaded cables in the paper by Mr. A. A. Clokey published in this issue of the *Bell System Technical Journal.*

The Necessity of Cable Signal Amplifiers and Their Requirements

Telegraph signals passing over a long submarine cable are distorted so severely that only a small fraction of the ultimate speed would be possible if extraordinary means were not taken to compensate for this distortion. It may be found that a certain cable attenuates very low frequencies to only one half of their original voltage while the higher frequencies may be received at less than one ten-thousandth of their initial strength. The transmission of an ideally perfect signal requires that its components of all frequencies be received at amplitudes proportional to those transmitted, consequently the reshaping of a signal received from a cable involves equalizing the strength of all its component frequencies by reducing the amplitude of the lower frequencies and amplifying the higher frequencies.

As the voltage which may be impressed upon a cable is limited by considerations of the safety of its insulation, the sensitivity of the receiver to currents of the highest frequency necessary in a properly defined signal is one of the limitations of the speed at which a cable may be operated. Unfortunately this is not the only limitation or it would be possible to increase the speed indefinitely by simply increasing the sensitivity of the receiver. All cables are exposed to extraneous interfering currents, some natural in origin and some the result of human activities, and while a great deal may be accomplished by proper design of the cable and the associated apparatus the speed is ultimately limited by interference. A large proportion of this interference is similar in nature to "static" and the bane of radio communication is also, but to a lesser degree, the bane of cable communication.

Experience has shown that continuous communication of the high standard of accuracy required in the transmission of code and cypher messages cannot be maintained unless the voltage received at the nominal signaling [3] frequency is between two and five millivolts. A receiver must therefore be capable of responding to voltages of this order at the signaling frequency in order to utilize the cable efficiently. On an average cable the power available at this voltage is of the order of 2×10^{-8} watts and there is at present no signal recording device

[3] This is defined in the case of the Morse cable code as the fundamental frequency in a series of alternate dots and dashes.

known to the art which will operate on so small a power except at uneconomically low speeds. For this reason it is necessary to insert between cable and recorder an instrument which will amplify the received signal.

A cable signal-shaping amplifier must fulfill many severe requirements. With its associated apparatus it must be capable of correcting the attenuation of the cable by equalizing the strength of all important component frequencies of the signal and it must also be capable of controlling in its output circuits a power many times as large as it receives.[4] It must be as insensitive as possible to interfering currents not included in the band of frequencies necessary to the signal and it is very desirable that overloading, which may be caused occasionally by these currents, should not permanently influence its adjustment or destroy any of its elements. The strength of its output current should be readily adjustable. It should be mechanically rugged, as otherwise its maintenance will require too large a proportion of the time of the staff at the cable station, and delays to traffic will be caused. Finally it should be protected as well as possible against local electrical fields and mechanical vibration and its operation should not be affected by conditions of extreme humidity.

Comparison of Mechanical Amplifiers and Vacuum Tube Amplifiers

In recent years several satisfactory mechanical amplifiers (called magnifiers in cable parlance) have been invented and their use has led to radical improvements in the speed of transmission over non-loaded cables. Most of these magnifiers utilize a sensitive moving-coil galvanometer, which moves some device a small distance in order to control a much greater power than that which caused the original motion of the coil. We may consider as typical of these the selenium magnifier which causes a beam of light to move over one or the other of two groups of selenium cells and thus varies their resistance, the Heurtley hot wire magnifier which changes the resistance of two pairs of almost microscopic heated wires by causing them to move relatively to each other, and the electrolytic magnifier which changes the resistances of a group of immersed electrodes. With all of these devices the controlled power is obtained from a local battery, but it is so small that it can do little more than operate a sensitive siphon recorder or a delicate moving coil relay. The latter may of course control a larger power which may in turn cause the operation of a comparatively rugged electromagnetic relay and thus indirectly a considerable power

[4] In practice the power amplification factor of the various types of amplifiers may range between five thousand and one hundred million.

may be controlled. With any of these magnifiers the suspended coil forms a mechanical oscillating system which is of great assistance in correcting the distortion of the cable, and allows signals to be shaped properly with the aid of a simple network of inductance, capacity and resistance. The inertias of the suspended coil and of the controlled devices make these magnifiers insensitive to high frequencies, and while this has some advantages in discriminating against high frequency disturbances, it also sets a rather definite limit to the speed at which they may be used. In order to utilize them as efficient signal shaping devices the natural frequency must be not far from one and one half times the nominal signaling frequency.[5] On this account and because their sensitivity decreases roughly as the square of the natural frequency to which they are adjusted, the moving coil magnifiers are rarely operated at signaling speeds of more than fifteen cycles per second. As they are easily damaged by relatively small overloads it is not safe to keep them in circuit when the approach of a thunder storm to a cable terminal makes the reception of induced surges in the cable likely. This sometimes results in keeping a cable out of operation for several hours, although the surges would only occasionally cause the loss of a letter if the magnifiers were not subject to damage by overloading.

A vacuum tube amplifier is free from many of the disadvantages of the mechanical amplifiers. It contains no delicate parts which require skilled manipulation, and once adjusted it maintains its adjustment indefinitely. There is no inherent limitation to the speed at which it may be operated; this being determined only by the requirement that the signal be sufficiently stronger than the interference. There is no practical limit to the amount of power which may be controlled and at the same time it is easy to limit this power and insure that momentary overloading shall not damage the amplifier or the associated apparatus. A multi-stage vacuum tube amplifier possesses still another important property in that there is practically no reaction between its various stages at telegraph frequencies. For this reason a number of interstage shaping networks may be used, and it will be found that the adjustment of one network is entirely without influence on the effects of the others.

History of Development of Vacuum Tube Amplifiers in Bell Telephone Laboratories

The signal shaping amplifier now in use is the outgrowth of studies of the applications of vacuum tubes begun in the laboratories of the

[5] Milnor, *A. I. E. E.*, February 1922.

American Telephone and Telegraph Company and the Western Electric Company in 1912. The vacuum tube amplifier appeared to offer important advantages for use on submarine cables because of its lack of distorting effects which are a function of the frequency of the current amplified, and also because of the ease with which signal distortion correcting circuits could be associated with the vacuum tubes. The initial studies on amplifier circuits suited to currents of the low frequencies involved in submarine cable telegraphy were made by Mr. R. V. L. Hartley and Mr. B. W. Kendall.[6] One of the difficulties which loomed quite large at that time was that most cables were operated duplex and the connection of an amplifier to a duplex circuit would involve the insertion of a transformer which promised to introduce distortion[7]. A suitable distortionless amplifier was first tried and subsequently distortion correcting networks were introduced between its stages. It was found that this permitted the use of more correcting elements than had been feasible in previous practice and thus indicated the possibility of attaining higher speeds than were usual at that time. The development of the shaping circuits employed was at first based on the principle of producing the various derivatives of the arriving current wave and adding them in proper phase relation to the arriving wave. This principle and methods of applying it had been developed mathematically by Mr. J. R. Carson of the American Telephone and Telegraph Company.[8] Mr. R. C. Mathes who conducted the experimental investigation beginning in 1916 simplified his work somewhat by recognizing that this principle was equivalent to a statement that the received signal would be satisfactory if the attenuation and phase distortion of the entire system of cable and amplifier for steady state alternating currents were corrected by the shaping networks over a range of frequencies from nearly zero to approximately the nominal signalling frequency. By the middle of August 1918 the employment of improved shaping methods made speeds of 22 cycles possible in simplex working on an artificial cable having a KR. of 2.7. The then standard cable apparatus would have permitted a speed of not more than 9 cycles, on a cable subject to interference of the magnitude usually encountered.

[6] B. W. Kendall, U. S. Patent No. 1,491,349, April 22, 1924.

[7] Expedients for avoiding distortion of this nature were suggested by Dr. H. W. Nichols of these Laboratories and by Mr. Lloyd Espenschied of the A. T. & T. Co. Their plans contemplated the modulation of an alternating current of relatively high frequency by the incoming signal, the amplification of the modulated current by suitable apparatus and its subsequent demodulation for obtaining the amplified low frequency signal (H. W. Nichols U. S. Patent No. 1,257,381, February 16, 1918; Lloyd Espenschied, U. S. Patent No. 1,428,156, September 5, 1922).

[8] See U. S. Patents No. 1,315,539, September 9, 1919, No. 1,450,969, April 10, 1923, No. 1,516,518, November 25, 1924 and No. 1,532,172, April 7, 1925. See also article by Dr. K. W. Wagner, *Electriche Nachrichten-Technik*, October 1924.

This apparatus [9] was then demonstrated to officials of the Western Union Telegraph Co. and with the cooperation of their engineers tests extending over a period of about a year were carried on at Rockaway Beach on several of the cables entering that station. It was shown in these experiments that while the vacuum tube amplifier together with suitable distortion-correcting networks would permit a considerable increase in the simplex speed (limited only by the interference present in the cable), the duplex speed was limited by the imperfect balance between the cable and the artificial line, and the increased sensitivity and signal shaping ability of the amplifier were of little value under the conditions then obtaining. Serious efforts were made to utilize the current limiting properties of vacuum tubes in conjunction with differentiating and integrating networks in reducing the effect of the unbalance on the signal and some successful results were attained.[10]

By 1920 the research leading to the development of the permalloy loaded cable had progressed to a point where it was evident that a new type of high speed cable amplifier would be required and the investigations were continued with this end in view. After the solution of numerous difficulties an amplifier was produced which was capable of correcting almost any variety of signal distortion which might be caused by a loaded cable. An amplifier of this type was tested on a trial length of 120 miles of loaded cable laid in a loop out of Devonshire Bay, Bermuda, and found to be generally satisfactory. The amplifier was then redesigned in a form suitable for commercial use and two amplifiers were built and installed at Rockaway Beach and Fayal in readiness for the New York-Azores loaded cable. They were put into successful operation and the predicted speed of 1,500 letters per minute was demonstrated within an hour after the cable had been released by the electricians of the ship which laid it.

Circuit Arrangements of Signal Shaping Vacuum Tube Amplifier

The electrical requirements of a cable signal shaping amplifier suitable for use on high speed loaded cables may be briefly stated as follows. It must take an input signal having components as low as one half millivolt and as high as possibly ten volts in amplitude, and correct the distortion by amplifying the weaker components much more than the stronger, at the same time making any necessary phase

[9] See U. S. Patents to R. C. Mathes No. 1,311,283, July 29, 1919, No. 1,426,755, August 22, 1922, No. 1,493,216, May 6, 1924 and No. 1,586,821, June 1, 1926; also Canadian Patent No. 207,231, January 4, 1921, granted to B. W. Kendall.

[10] D. K. Gannett and M. Kirkwood, U. S. Patent No. 1,483,172.

corrections. It must also be able to operate satisfactorily with signals in which the weaker components may be as strong as 100 millivolts. An output of about fifteen milliamperes at 15 volts should be available and this output must be adjustable by small steps. It must be capable of handling currents containing frequencies between a small fraction of a cycle and about 180 cycles, the particular part of this band of frequencies which is utilized depending on the nature of the cable and the speed at which it is operated.

These requirements are met in the present cable amplifier,[11] by circuits which are shown in the upper half of Figure 1. The amplifier proper consists of four stages of vacuum tubes, the first three being designed for high voltage-amplification and the last for large current output. An additional output stage is provided for the purpose of increasing the flexibility of the amplifier by permitting two separate classes of apparatus to be operated simultaneously. The coupling between stages is a combination of two types, the coupling for the very low frequencies being through a resistance capacity network while that for the higher frequencies is by means of an auto-transformer of special design or by highly damped resistance, inductance and capacity networks. The amplifier is connected to the cable through an input network and a shielded transformer. The input network assists in shaping the signal and prevents the first stage of the amplifier from being overloaded by the strong low frequency components of the signal arriving in the cable. The transformer permits earthing the filaments of the amplifier tubes and their associated batteries and avoids the short circuiting of the long balanced sea earth. The latter is used to reduce the effect of electrical disturbances on that part of the cable which lies in shallow water near the shore.[12] The requirements of this transformer are quite unusual as it must have a satisfactory voltage regulation from .2 to 200 cycles per second.

The ability to operate on voltages which may vary widely from time to time makes it necessary to provide a suitable range of adjustment of amplification. This is accomplished by providing that the secondary windings of the input transformer may be connected in series or in parallel and the plate coupling resistances of the tubes varied by a factor of four to one. A potentiometer placed between the second and third stages of the amplifier allows a variation of twenty to one in the voltage transmitted to the third stage, and with other adjustments as mentioned above, increases the total range

[11] A. M. Curtis, U. S. Patents Nos. 1,586,970 and 1,586,972, June 1, 1926, and 1,624,395 and 1,624,396, April 12, 1927.

[12] J. J. Gilbert, *Bell System Technical Journal*, July 1926; also British Patent No. 218,261, August 31, 1925, and Canadian Patent No. 265,944, Nov. 16, 1926.

of amplification adjustment to about 150 to 1. In addition a set of constant resistance potentiometers in the relay control panel associated with the amplifier allows the current through any of the relays to be varied in small steps without influencing the current in the other relays or changing the impedance of the amplifier output circuit.

The characteristic of the amplifier system may be measured by applying a certain input voltage at varied frequencies, and noting the corresponding output voltage. If the amplifier has been adjusted to give a satisfactory signal when connected to a cable, measurements will show that its amplification increases rapidly to a maximum which occurs at about 1.5 times the signaling frequency, and then falls to practically zero at about twice the signaling frequency. This elimination of the higher frequencies is effected by proper adjustment of the inter-stage shaping networks, and it results in suppressing that portion of the interfering currents received from the cable which lie above the band of frequencies required to form the signal.

The amplifier as described above is perfectly suitable for recorder operation and will permit communication at speeds up to at least ninety cycles per second which in cable code is equivalent to about 2,800 letters per minute, provided that a suitable recorder is employed. It is, however, not entirely suitable for multiplex printing telegraph operation under all conditions without the addition of apparatus to prevent "zero wander."

System for Correcting "Zero Wander"

In general, printing telegraph systems have been designed on the assumption that they were to work over land telegraph lines and they contain no provision for avoiding the effects of the "wandering zero" which is caused by the inability of a practical cable transmission system to transmit direct current. This inability to transmit direct currents is due to the fact that there is usually present in a submarine cable an earth current which is many times as strong as the signal, and it is necessary to block out this earth current by series condensers (as is usual in ordinary cable practice) or to keep it out of the amplifier by a transformer. The syphon recorder does not require that the zero of the signal be maintained closely but cable signal relays operate on a fixed value of current of either polarity and are incapable of determining whether or not part of this current is due to a displacement of the zero. It is therefore necessary to reconcile in some way the printing telegraph systems, which under some conditions require the reception of a direct current, with the cable system which cannot transmit a direct current. Several methods of doing this have been

used with the mechanical amplifying systems on low speed cables; they usually supply directly to the relay a "zero correcting" current which depends upon the past history of the signal.[13]

When a vacuum tube amplifier is employed it is more convenient to apply the zero correction to the grid of the last stage vacuum tube as this results in the most economical utilization of the correcting battery and its circuits. The zero correcting apparatus is mounted in a cabinet adjacent to the amplifier and differs considerably in principle from that hitherto used with mechanical amplifiers.

The three element moving armature polarized relay, which had been designed for use in loaded cable operation generally, was changed in some details and adapted for use in the zero corrector. It is capable of operating at a high speed and also discriminates very accurately between currents of slightly differing values. When actuated by the normal signal its armature contacts vibrate between the fixed contacts, not touching either unless the zero of the signal deviates more than about three per cent from its proper position. When this deviation does occur the relay contacts close the circuit for an instant at the peak of a signal wave, and permit the battery to which they are connected to charge a condenser through a comparatively low resistance. The charge on this condenser then passes gradually to a second condenser through a high resistance and at the same time commences to be discharged from the second condenser by a shunt resistance. The voltage on this second condenser is applied to the grid of the last stage of the amplifier in such a sense that it produces a deflection of the amplifier zero in the direction opposite to the deflection which caused the relay contacts to come together. This correcting voltage is applied at a rate which is slow enough to prevent it from distorting the signal and the rate at which it is dissipated by the shunt if no further contacts take place is still slower. It should be noted that these rates of charge and discharge, while adjustable, need not bear any accurate relation to the shape of the signal itself. The correction is usually applied rapidly enough so that the zero is brought back to normal within the duration of about five signal pulses and the proper operation of the circuit prevents the zero from passing beyond limits about five per cent of the signal amplitude either side of the normal position. A somewhat simplified circuit of the relay control unit which includes the zero corrector is shown in the lower part of Fig. 1.

[13] A system in common use is due to S. G. Brown (British Patent No. 6,275, February 20, 1913). Other systems have been invented by D. K. Gannett and M. Kirkwood, U. S. Patent No. 1,548,878, Aug. 11, 1925, and R. C. Mathes, U. S. Patent No. 1,295,553, February 25, 1919.

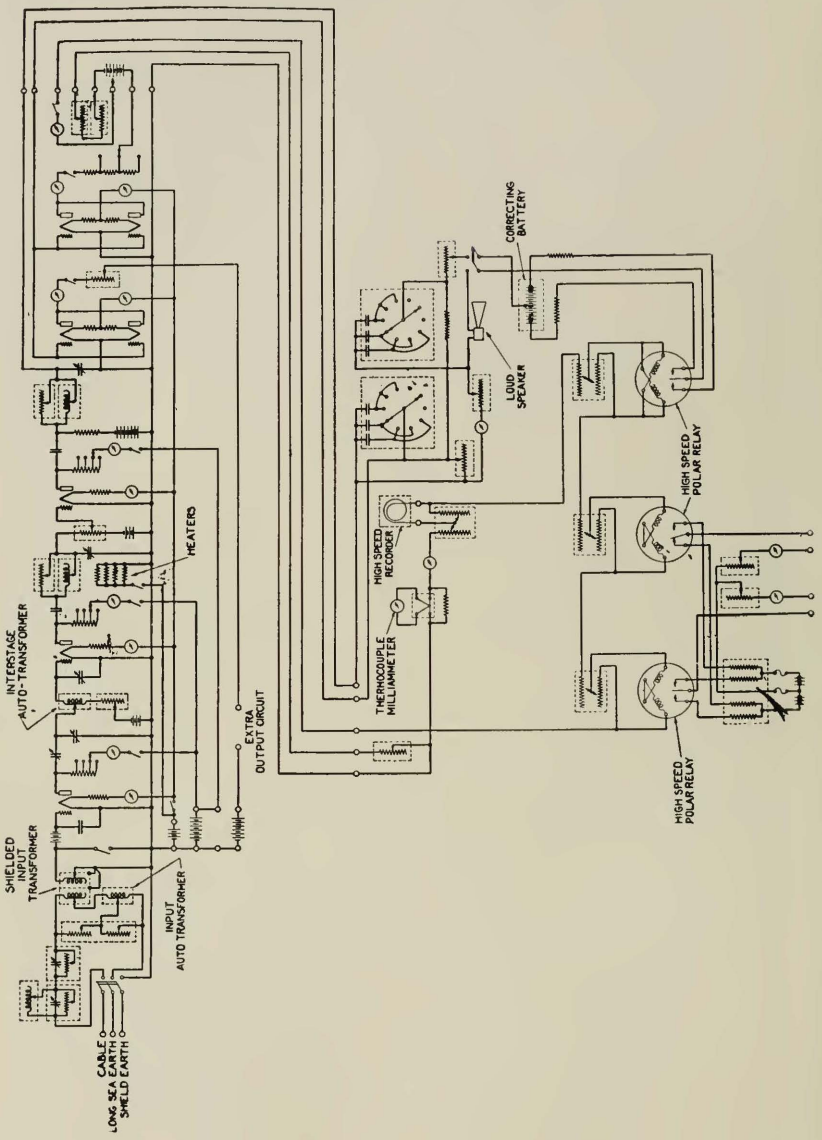

CABLE
LONG SEA EARTH
SHELD EARTH
INPUT AUTO TRANSFORMER
SHIELDED INPUT TRANSFORMER
INTERSTAGE AUTO-TRANSFORMER
HEATERS
EXTRA OUTPUT CIRCUIT
THERMOCOUPLE MILLIAMMETER
HIGH SPEED RECORDER
LOUD SPEAKER
CORRECTING BATTERY
HIGH SPEED POLAR RELAY
HIGH SPEED POLAR RELAY

FIG. 1.

Mechanical Details of Amplifier and Relay Control Desk

The construction of the amplifier and relay control desk is shown in the accompanying figures. Fig. 2 is a front and Fig. 3 a back view of the amplifier in its cabinet, Fig. 4 is a back view of the panel frame removed from the cabinet and Fig. 5 is a front view of the relay control panel associated with the amplifier. Mechanically the amplifier



Fig. 2.

consists of a frame of brass angles supporting on its front face four large hard rubber panels and sixteen small ones. The four large panels are mounted on the upper part of the frame, and hold the switches and the meters which it has been found desirable to provide in the filament and plate circuit of each vacuum tube. The adjustable elements of the amplifier interstage shaping networks are contained in the sixteen smaller panels mounted in the lower part of the framework. Each of these panels is a complete unit, comprising either a variable condenser, a variable resistance, or a switch for the adjustment of an

inductance. As these panels are all of the same size, it is possible if necessary to change radically the arrangement of the interstage networks, or substitute entirely different ones without any difficult mechanical work on the amplifier. Each of the condensers and resistances is contained in an earthed metal box, into which it is sealed by a wax insulating compound. The frame in which the panels are



Fig. 3.

mounted is earthed and metallic fins are provided between the various panels in such a manner that any current which leaks through a film of moisture which might be deposited on the surface of the panels must pass to earth. The plan of mounting each individual piece of apparatus in a metal box, and sealing it in with insulating compound, has been adhered to throughout, the tubes, meters and switches of course being excepted. This is principally for protection from the serious humidity frequently found in cable stations. Additional protection is provided by electric heaters drawing current from the battery which operates the filaments of the vacuum tubes. Frame-

work shelves provide space for mounting the tubes and heavy apparatus such as coils and coupling condensers. The tube of the first stage is held in a spring suspended socket, damped by an oil dashpot.[14] The socket of the second stage tube is sufficiently protected from vibration by a sponge rubber mounting, while the sockets of the third and fourth stages need no special protection. Dry cell grid batteries are mounted on the lower shelf of the framework. The numerous external connections are brought to the terminal strip which may be seen along the lower part of the back of the framework. The panel assembly is mounted in the upper part of a mahogany case, lined with copper. The lower part of the case contains the shaping

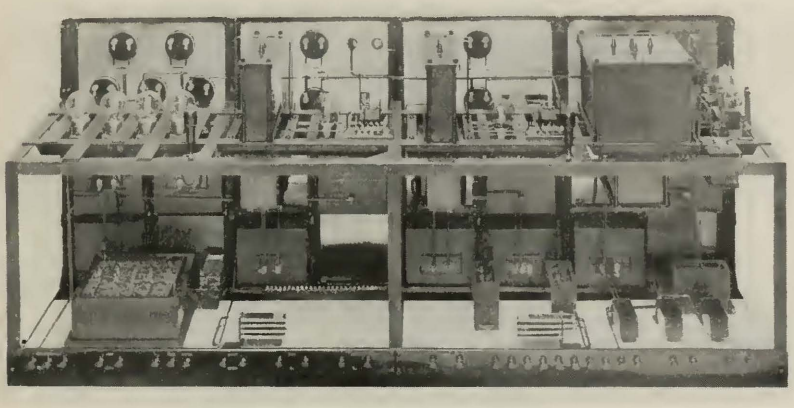


FIG. 4.

networks which are connected between the cable and the amplifier All adjustments are made either from the front of the panels or on apparatus contained in the lower part of the cabinet, and as all the elements of the amplifier are inherently stable, when the adjustments for shaping the signal at any given speed have once been determined they may be quickly duplicated at any time.

The relay control desk combines the apparatus for correcting the signal zero wander with means for adjusting the current through the relays used in the multiplex printing telegraph system and includes several switches used in the control of the latter apparatus.

In designing the amplifier the necessity of maintaining continuous operation and easily and quickly remedying any minor troubles was considered of the utmost importance, and this led to its being made large enough so that work may be done inside of it without having to remove

[14] W. A. Knoop, Patent U. S. No. 1,523,430, January 20, 1925.

it from the circuit and take it apart. All of the circuit elements may be reached from the back of the cabinet without disturbing anything else, and on several occasions this arrangement has proved of great value.

Power Supply for Amplifier

Three sets of storage batteries are required for an amplifier. The filaments are heated by a 6 V. 500 AH storage battery. The plate voltage for the first three stages is supplied by a 250 volt 1 AH storage

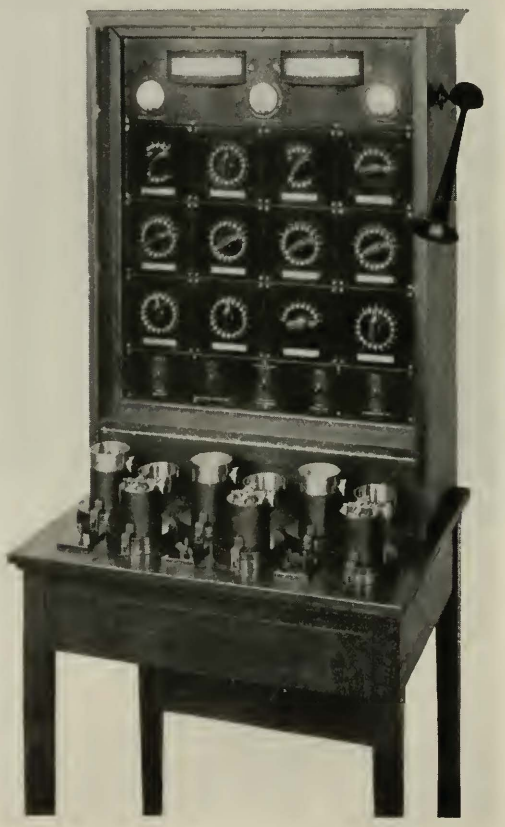


Fig. 5.

battery while the plate voltage for the last stage is supplied by a 275 V. 4 AH storage battery. These batteries are in the general battery room of each cable station and are handled by ordinary methods, the only special precautions necessary being the shielding

of the leads from battery to amplifier and the avoiding of loose switch contacts.

Results Obtained in Service

The first two amplifiers built were put in operation on the New York-Azores cable in September 1924, and after a few weeks' testing a speed of 65 cycles per second or about 2,080 letters per minute in cable code was demonstrated. In the fall of 1926 three additional permalloy loaded cables were completed and equipped with vacuum tube amplifiers. They are laid between New York and Bay Roberts, Newfoundland, between Bay Roberts and Penzance, England, and between Fayal, Azores, and Emden, Germany. A speed of ninety cycles has been demonstrated on the New York-Bay Roberts cable, and the longer section from Bay Roberts to Penzance has worked at eighty cycles. The adjustment of amplification and the flexibility of the shaping networks is such that it has proved possible to remove an amplifier adjusted for operation at about 40 cycles from the long New York-Azores cable and readjust it for use on the short New York-Bay Roberts cable at 20 cycles in about fifteen minutes.

During the two and one half years of operation of amplifiers on the New York-Azores cable the maintenance required has been almost negligible and the rare cases of trouble have usually been found in the external connections. The longest delay to traffic caused by the amplifiers during this period was about two hours, and was due to the disarrangement of some temporary wiring. The reliability of the amplifiers is well attested by the fact that during the first two years there was only one available at each station, and there was no difficulty in keeping cable and amplifiers in continuous operation.

In connection with maintenance the vacuum tube amplifiers have a great advantage in that they do not require any delicate mechanical adjustments, while the electricians responsible for the operation of mechanical amplifiers must frequently spend hours at tasks requiring the skill and patience of a watchmaker.

It has been found possible to handle messages during thunderstorms which prevented operation of the non-loaded cables and their mechanical amplifiers for several hours. As an experiment the loaded cable and amplifier were worked continuously through an unusually severe thunderstorm during which stop watch observations of the intervals between lightning flashes and thunder claps showed that lightning had struck within a thousand feet of the cable terminal on three occasions. Although the induced surges were frequently several times as strong as the signals, the automatically limited output of the

amplifier protected the recorders from damage, and the effect of each lightning discharge was limited to the possible mutilation of one or two letters.

The protection of these amplifiers from mechanical vibration has proved entirely satisfactory. During alterations to the Western Union Cable station building at Rockaway Beach a brick wall six feet from the amplifier was broken down with sledge hammers without interfering with the normal handling of messages.

Other Applications of Vacuum Tube Amplifiers in Cable Telegraphy

While as yet vacuum tube amplifiers have been utilized principally on high speed loaded cables they are not necessarily restricted to such use. It was mentioned in an early part of this paper that, since the non-loaded cables are ordinarily operated duplex at a speed which is set, not by the sensitivity of the receiver, but by the strength of the interference due to imperfect balance between cables and artificial line, no increase in speed might be expected to result from the substitution of vacuum tube amplifiers for the mechanical amplifiers now used. Nevertheless the superior ruggedness of the vacuum tube amplifier, combined with its ability to operate safely through thunderstorms which would ruin the mechanical amplifiers, might reduce appreciably the amount of lost time, particularly during the summer months, and thus improve the traffic capacity of these cables.

In addition to the use of vacuum tube amplifiers for operating terminal apparatus they have another important field as repeaters intermediate between two short sections of a long cable. As the speed at which any cable can be operated is roughly inversely proportional to the square of its length, it is customary to lay cables connecting distant centers of population in two or more sections, interrupted at some conveniently located but often inconveniently isolated island. This involves repeating the signals received from one section of the cable into the next section and for years this was done manually, that is, an operator received and translated the signals and passed them on to another operator for transmission on the other section. Within recent years through relay operation by means of repeaters has become general. At the intermediate station a device receives the signal from an amplifier and retransmits it to the next section usually correcting it completely to its original form. These repeaters, while very successful, are still quite complicated mechanically, and require skillful maintenance, particularly as they utilize delicate mechanical amplifiers and moving coil relays. It is possible to replace

them by vacuum tube amplifiers having no moving parts, and thus requiring a minimum of maintenance. The vacuum tube amplifier is capable of reshaping the signal almost as completely as is done by the mechanical repeaters, and, in case of a cable worked simplex, the direction in which the amplifier at the intermediate station repeats the signal may be automatically controlled from either the sending or the receiving station.

Modulation in Vacuum Tubes Used as Amplifiers

By EUGENE PETERSON and HERBERT P. EVANS

Synopsis: Recent developments in amplifier design tending toward more rigorous quality requirements have shown that the solutions of Van der Bijl and Carson are inadequate for certain purposes since they are based upon a convenient assumption which is not satisfied in fact. In particular, a detailed investigation of carrier current repeaters used for the simultaneous transmission of several channels, and upon which in consequence the modulation or crosstalk requirements are particularly severe, showed the modulation currents measured to be quite different from those specified by the theory, as was the law of variation of these currents with the circuit constants.

The cause of the discrepancy was found to reside in the neglect of the variation of the amplification factor (μ) with both plate and grid potentials. When the actual state of affairs was taken into account in the analysis by the application of a general method involving no assumptions, theory and experiment were found to be in good accord. The new expressions have been developed in terms of the amplification factor (μ), the internal output resistance of the tube (R_0), and their differential parameters, which are involved in the representation of the characteristic tube equation by a double power series. Expressions for the current components are developed in terms of the coefficients of the series, and modifications of Miller's method for greater convenience and precision in determinations of tube characteristics are described from which the series coefficients may be evaluated.

Conclusions are drawn from the solutions as to desirable tube characteristics by which, for example, a single tube may take the place of two tubes in push-pull connection. Finally, certain properties of different types of tubes under conditions of maximum output power are compared on the basis of μ constant and μ variable.

THE amount of modulation produced in vacuum tube amplifiers is in many cases a controlling factor in their application and it becomes of importance to determine how modulation products arise, so that the possibility of reducing them by tube and by circuit design may be studied.

In restricting discussion to amplifiers, and particularly to those used in communications, we are treating cases most amenable to analysis; in which, normally, the applied potential variation maintains the grid always negative so that conductive grid current does not flow, but in which, on the other hand, the greatest negative potential does not exceed the negative end of the plate current grid potential characteristic. In applying these two detailed restrictions we are incidentally insuring against prohibitive quantities of modulation; we know that, for example, the flow of conductive grid current may, under special conditions, produce an exceptionally efficient modulator of great service as such, but highly undesirable as an amplifier.

The necessity for suppressing modulation proceeds from the disturbing effects attendant on it, by which there may result reduction

of quality in speech amplifiers, and crosstalk in the multi-channel amplifiers of carrier telephony, to take but two examples. The modulation level in the last case is restricted to much smaller values than are tolerated in the first; it is commonly required to reduce modulation products to the thousandth part, or even less, of the fundamentals which produce them. This last case is the one in which we are primarily interested; other cases of greater distortion referred to above may be treated by an extension of the methods used below in the case of grid current flow, and by Fourier series or expansions in terms of Bessel functions when the negative end of the tube characteristic is exceeded.

A thoroughgoing study of the amplifier problem would relate the static characteristics of a tube and the parameters of the circuit in which it works to its operating characteristics, and then would relate its static characteristics to the internal structure of the tube; it would in brief enable us to link the details of tube structure to the fundamental and modulation currents produced in the output wave of the amplifier. In the following, however, we shall treat only that part of the general problem which relates the operating characteristics and circuit parameters to the static characteristics.

A consideration of the usual plate current characteristics of a three electrode vacuum tube, as shown in Fig. 1, demonstrates the well-

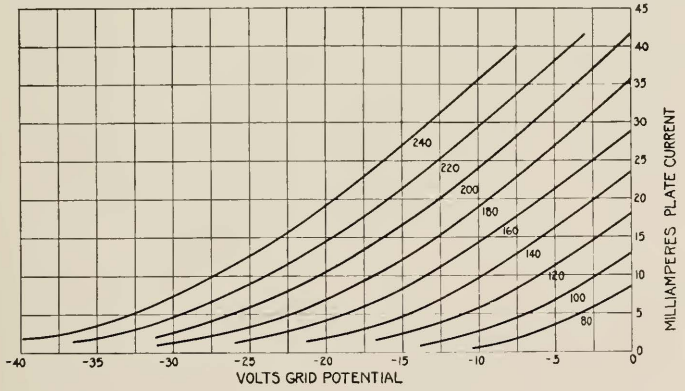


Fig. 1—Plate current as a function of grid potential with plate potential as a parameter. EL tube No. 109,150. $I_f = 1.1$ amperes

known dependence of the plate current upon the two variables, the grid and plate potentials. That is to say, the plate current varies with the grid potential when the plate potential is fixed, and it varies with the plate potential when the grid potential is fixed. It has been found of great convenience in the past to utilize an approximate relation between the grid and plate potentials as expressed in what is

sometimes described as the fundamental theorem of the vacuum tube. The theorem states that a potential change in the grid circuit appears as a voltage generated in the plate circuit, the magnitude of which is equal to the grid potential change multiplied by the amplification factor μ. Solutions have been obtained for the output current components with the aid of this relation through the work of Van der Bijl and of Carson, which have been of great practical importance.

These solutions are approximate because of the simplifying assumption of the constancy of the amplification factor, which is certainly not accurate as the curves of Fig. 2 demonstrate. In this diagram

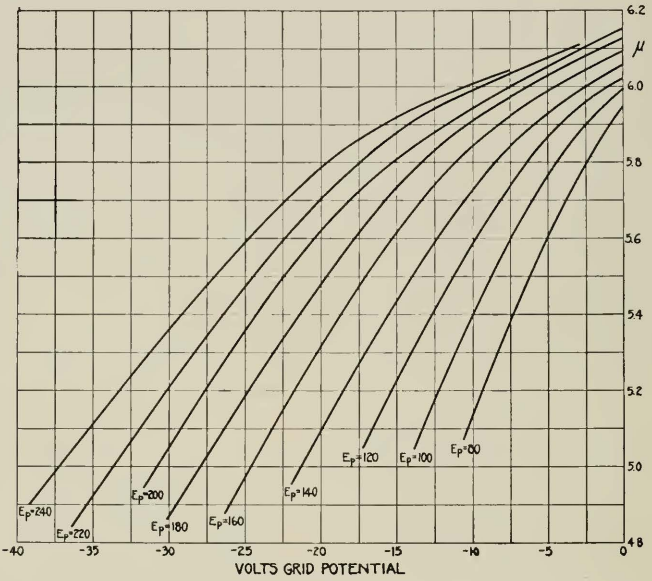


Fig. 2—μ as a function of the grid potential with plate potential as a parameter. EL tube No. 109,150. $I_f = 1.1$ amperes

the amplification factor for small applied potentials is plotted as ordinate with the grid potential as abscissa and the plate potential as parameter. The variation of μ is observed to be of the order of twenty per cent over the operating range. When we are interested in the distortion of the input wave, this variation cannot be ignored since to do so would in some cases yield results of another order of magnitude than those found experimentally. The treatment for our specific needs must therefore be modified to take account of the actual state of affairs.[1]

There are two ways open for a treatment involving the variation

[1] Early calculations to show the effect of a variable μ upon distortion were given by H. Nyquist of the American Telephone and Telegraph Company in an unpublished memorandum of April, 1921.

of μ; one is a modification of Carson's analysis while the other involves a reconsideration of the tube characteristic equation. The modification of Carson's treatment may be carried out with the aid of an expedient as follows: with a single input frequency the wave form of the generator voltage acting in the plate circuit is distorted by the variable amplification factor of the tube, and it is this distorted wave which acts in the plate circuit, instead of the pure sine wave operating with constant μ. The method of procedure is then evident; if we refer the actual distorted generator potential to the grid circuit—that is, if we divide it by the average value of μ—we have an input wave which, when treated by Carson's well-known method [2] in which μ is assumed constant, will yield correct results inasmuch as the μ variation has been taken into account—somewhat indirectly, it is true. When complex waves are applied to the grid circuit the effective grid potential is made up of numerous components and the treatment becomes very cumbersome.

Tube Characteristic Expressed by Double Power Series

A more direct method which has been used with some success consists in expressing the tube current-voltage relation by a double power series, without invoking any special relations regarding the connection between a grid potential change and the equivalent plate potential. If we use the symbols I_b, E_p, E_c to denote the plate current, plate potential, and grid potential, respectively, we may express the plate current as a double power series as follows:

$$\begin{aligned} I_b = f(E_p, E_c) &= \Sigma a_{mn} {E_p}^m {E_c}^n \\ &= a_{00} + a_{10}E_p + a_{01}E_c \\ &\quad + a_{20}{E_p}^2 + a_{11}E_pE_c + a_{02}{E_c}^2 \\ &\quad + a_{30}{E_p}^3 + a_{21}{E_p}^2E_c + a_{12}E_p{E_c}^2 + a_{03}{E_c}^3 \\ &\quad + \cdots \end{aligned} \tag{1}$$

where

$$a_{mn} = \frac{1}{m!\,n!} \frac{\partial^{m+n} f(0, 0)}{\partial {E_p}^m \partial {E_c}^n}, \tag{2}$$

and in which it is understood that the development applies with the operating point on the characteristic. The derivatives, it will be noted, are evaluated at the point at which both E_p and E_c are zero. Some of the coefficients of Eq. (1) may of course be eliminated by reference to the evident properties of the tubes, but this need not concern us here since it is more convenient to formulate the tube equation in another way.

[2] *Proc. I. R. E.*, 1919.

Under normal conditions of amplifier operation E_p and E_c are fixed, and the alternating grid and plate potentials vary about these potentials. It then becomes convenient to express the coefficients as derivatives referred to the specific point E_{p_0}, E_{c_0}. If we indicate the variable components of the grid and plate potentials by e and v, respectively, the tube equation may be put in the form

$$\begin{aligned} I_b &= f(E_{p_0} + v, E_{c_0} + e) = \Sigma b_{mn} v^m e^n \\ &= f(E_{p_0}, E_{c_0}) + b_{10}v + b_{01}e \\ &\qquad + b_{20}v^2 + b_{11}ve + b_{02}e^2 \\ &\qquad + b_{30}v^2 + b_{21}v^2e + b_{12}ve^2 + b_{03}e^3 \cdots, \end{aligned} \tag{3}$$

where

$$b_{mn} = \frac{1}{m!\,n!} \frac{\partial^{m+n} f(E_{p_0}, E_{c_0})}{\partial^m E_p \partial^n E_c}.$$

The b coefficients are functions of the operating point—specified by E_{p_0}, E_{c_0}—and change with the operating point, in general. The a coefficients are definitely referred to the origin, however, so that each b coefficient may be expressed in terms of the a coefficients which correspond to it. It is possible, by determining this relation, to follow the variation of the b's in terms of E_{p_0} and E_{c_0}. To do this we substitute for E_c and E_p in Eq. (1) the expressions $E_{c_0} + e$, $E_{p_0} + e$, respectively, find the constant term, and obtain succeeding coefficients by differentiation. Thus

$$\begin{aligned} b_{00} &= a_{20}E_{p_0}^2 + a_{30}E_{p_0}^3 + a_{40}E_{p_0}^4 \\ &\quad + a_{21}E_{p_0}^2E_{c_0} + a_{31}E_{p_0}^3E_{c_0} + a_{22}E_{p_0}^2E_{c_0}^2, \end{aligned}$$

and further

$$b_{10} = \frac{\partial b_{00}}{\partial E_{p_0}}, \qquad b_{01} = \frac{\partial b_{00}}{\partial E_{c_0}}.$$

This concludes our consideration of the tube characteristics without reference to the circuit to which the tube may be connected. Eq. (3) rather than Eq. (1) will be used in the following.

It should be noted in terminating this part of the discussion that the treatment is capable of easy extension to characteristics depending upon a larger number of variables. Thus a four element (double grid) tube characteristic may be expressed by a triple power series, and so on. When the potential of the second grid is maintained constant it is evident that the tube characteristic is given by a double power series in which the coefficients depend in addition upon the potential of the second grid. To determine the dependence quantitatively, the triple series will serve.

Solutions for the Plate Circuit Components

We now pass on to a consideration of the operation of the tube working into a plate resistance. The more general case of a load impedance which is a function of frequency may be treated by application of the equations derived above, but it will serve our purpose here to deal with the case of a pure resistance load since the experimental work was done for that particular case which is of considerable practical importance.

If J is the alternating component of the plate current, we have from (3)

$$J = I_b - f(E_{p_0}, E_{c_0})$$

and J, it is seen, is a function of the two variables v and e. The quantity v depends on e of course, so that J may evidently be expressed as a function of e alone or

$$J = \sum_{k=1}^{k=\infty} C_k e^k. \tag{4}$$

A solution of the problem therefore consists in determining the C's in terms of the circuit and tube parameters.

The change in plate potential v may further be expressed as

$$v = -RJ = -R \sum_{k=1}^{k=\infty} C_k e^k. \tag{5}$$

The C's are then determined by putting (4) and (5) in (3) and identifying coefficients of similar powers of the variable. We have then

$$\begin{aligned} v &= -RC_1 e - RC_2 e^2 - RC_3 e^3 \quad \cdots, \\ v^2 &= \quad R^2 C_1^2 e^2 + 2R^2 C_1 C_2 e^3 \quad \cdots, \\ v^3 &= \quad - R^3 C_1^3 e^3 \quad \cdots, \end{aligned}$$

in which powers higher than the third are neglected for this, the first approximation. Carrying through the substitutions we obtain the solutions

$$\begin{aligned} C_1 &= b_{01}/(1 + b_{10}R), \\ C_2 &= (b_{02} + b_{20}R^2C_1^2 - b_{11}RC_1)/(1 + b_{10}R), \\ C_3 &= \frac{b_{03} - RC_1b_{12} + R^2C_1^2b_{21} - R^3C_1^3b_{30} - RC_2b_{11} + 2R^2C_1C_2b_{2G}}{1 + b_{10}R}. \end{aligned} \tag{6}$$

The first equation, which leads to the first approximation to the fundamental current, is identical with that obtained on the basis of μ constant, but the higher orders are distinctly changed. When e is a

pure sine wave C_2 contributes to the constant term which represents the change in direct current, C_1 and C_3 contribute to the fundamental component of the plate current, C_2 gives rise to the second harmonic and C_3 gives rise to the third harmonic current. When e is a complex wave the subscripts indicate the order of modulation [3] to which each coefficient applies.

The b coefficients may be readily converted into quantities dependent on μ, R_0, and their derivatives; we have

$$\mu = \frac{\partial I_b/\partial E_{c_0}}{\partial I_b/\partial E_{p_0}} = \frac{b_{01}}{b_{10}}$$

and

$$1/R_0 = \partial I_b/\partial E_{p_0} = b_{10},$$

so that

$$\mu/R_0 = b_{01}.$$

Succeeding coefficients are obtained by differentiating with respect to E_p and to E_c. For example,

$$b_{20} = -\frac{1}{2R_0^2}\frac{\partial R_0}{\partial E_{p_0}},$$

$$b_{11} = \frac{1}{R_0}\frac{\partial \mu}{\partial E_{p_0}} - \frac{\mu}{R_0^2}\frac{\partial R_0}{\partial E_{p_0}} = -\frac{1}{R_0^2}\frac{\partial R_0}{\partial E_{c_0}},$$

$$b_{02} = \frac{1}{2R_0}\frac{\partial \mu}{\partial E_{c_0}} + \frac{\mu}{2R_0}\frac{\partial \mu}{\partial E_{p_0}} - \frac{\mu^2}{2R_0^2}\frac{\partial R_0}{\partial E_{p_0}}.$$

The b coefficients may be obtained directly from the family of characteristic curves either graphically or analytically, when the operating point is specified. If we obtain our coefficients from the μ and R_0 curves, however, derivatives of a lower order are required than we need in dealing directly with the static characteristics. A family of μ-curves is shown in Fig. 2, and a family of R_0 curves is shown in Fig. 2*a*.

Results applying to the four element tube which are obtained by methods analogous to the above may be stated briefly. If we express the change in plate current by

$$J = C_{10}\epsilon + C_{01}e + C_{20}\epsilon^2 + C_{11}\epsilon e + C_{02}e^2 \cdots,$$

where ϵ and e represent the alternating potentials on the two grids,

[3] The new frequencies produced by modulation are given by the expression

$$F = |mf_1 \pm nf_2 \pm \cdots|,$$

where f_1, f_2 are impressed frequencies and m, n are integers or zero; the order is simply the sum of m, n, $\cdots$.

we find

$$C_{10} = b_{010}/(1 + Rb_{100}),$$

$$C_{01} = b_{001}/(1 + Rb_{100}),$$

$$C_{11} = \frac{b_{011} + 2R^2C_{10}C_{01}b_{200} - RC_{01}b_{110} - RC_{10}b_{101}}{1 + Rb_{100}},$$

in which

$$b_{rst} = \frac{1}{r!\,s!\,t!}\frac{\partial^{r+s+t}f(E_{p_0}, E_{n_0}, E_{c_0})}{\partial E_p{}^r \partial E_n{}^s \partial E_c{}^t}$$

and E_{n_0} is the fixed potential of the second grid.

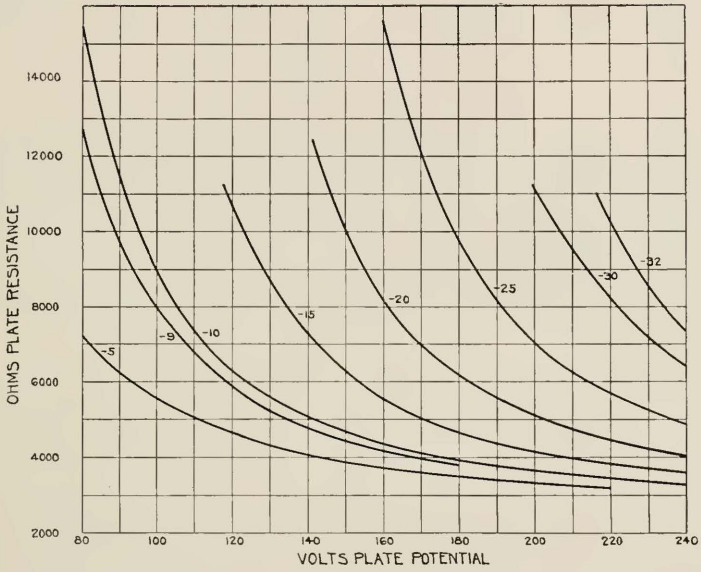


Fig. 2*a*—Plate resistance as a function of plate potential with grid potential as a parameter. EL tube No. 109,150. $I_f = 1.1$ amperes

We proceed to the methods used for the experimental determination of μ and R_0 for the three element tube.

Measurement of Tube Parameters

The amplification constant and plate impedance of a three element tube may be measured with precision by a well-known method due in principle to J. M. Miller [4] which requires no explanation here. The method as originally proposed is somewhat inconvenient in that the space current of the tube under test passes through a resistance common to the alternating measuring current, so that the operating

[4] *Proc. I. R. E.*, Vol. 6.

point of the tube is changed during manipulation for balance. Everitt's modification,[5] which consists in separating the direct and alternating current paths by a retard coil and condenser in the usual manner, is therefore preferable in this respect, but a complicating factor enters in the introduction of a reactive component due to the retard coil which cannot be balanced out by the variable resistances originally provided. This may be taken care of in a more or less obvious way by shunting a reactance around the grid resistance as shown in Fig. 3, the effect of which is to correct for the introduced

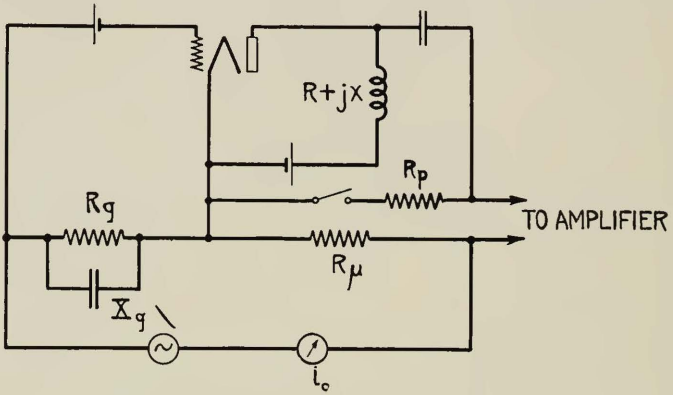


Fig. 3—Modification of Miller's method for determining μ and R_0

phase unbalance due to the retard coil and so lead to a precisely determinable null point instead of to a broad minimum, as is otherwise found.

The effect of the inserted reactance may be calculated by direct methods. Referring to the figure, there is no potential of fundamental frequency across the amplifier used to indicate balance at the null point, and if the grid-filament impedance of the tube is much greater than R_g (50 ohms), the total oscillator current passes through R_g and X_g in parallel, with R_μ in series. The potential impressed on the grid is then

$$e_g = jR_gX_gi_0/(R_g + jX_g),$$

which appears in the plate multiplied by the amplification factor of the tube and reversed in sign. The nomenclature is clearly indicated in the Figure. An alternating current flows in the plate circuit which is just balanced by the drop across R_μ so that we may write

$$\frac{j\mu R_gX_g}{R_g + jX_g} \cdot \frac{R + jX}{R_0 + R + jX} = R_\mu,$$

[5] See p. 201 of van der Bijl's "Thermionic Vacuum Tube."

which yields, after some reduction,

$$\mu = \frac{R_\mu}{R_g}\left(1 + \frac{R_0R + R_0^2}{X^2 + R_0R + R^2}\right).$$

As to the order of magnitude of the various quantities involved, R^2 is usually negligible before X^2, while R_0 and R may be of the same order of magnitude, so that we have

$$\mu = \frac{R_\mu}{R_g}\left(1 + \frac{2R^2}{X^2}\right).$$

In the specific case of a 101-D tube we had $X = 2.1 \times 10^5$, $R = 7 \times 10^3$ and

$$\mu = \frac{R_\mu}{R_g}(1 + 0.002).$$

The correction term, amounting to two parts in a thousand, drops out without the retard coil and we arrive at Miller's formula $\mu = R_\mu/R_g$. In measuring the output impedance of the tube after the settings for μ have been determined, R_g is doubled and R_p is connected in the plate circuit and varied until balance is again attained. It has been shown [6] by extension of the method used above that

$$R_0 = R_p\left[1 + \frac{3}{2}\left(\frac{R_p}{X}\right)^2\right]$$

and the correction term is of the same order of magnitude as that previously found for the amplification factor.

Balances may be obtained precise to one part in a thousand or better, but in much of our own work the observations are not ordinarily corrected for finite reactance. In order for the balancing action to take place the two reactances must be of opposite sign since amplification produces a 180° phase shift. If we balanced by a reactance shunted around R_μ instead of around R_g, the inserted reactance would, of course, be of the same sign as that of the plate retard coil, which was inductive at the frequency of 1,000 cycles at which the balances were made. The alternative scheme of shunting a variable condenser around R_g was adopted purely as a matter of convenience.

Applications of the Analysis

Second Order Modulation in Voltage Amplifiers

A striking illustration of the difference in the results of the two analyses, one based on the assumption of constant amplification

[6] By Mr. V. A. Schlenker.

factor and the other based on actual tube characteristics, is provided by a consideration of the second order modulation in the case of large external plate resistance.

The ratio of second harmonic to the fundamental, when μ is assumed invariable, comes out proportional to

$$(R + R_0)^{-2},$$

which shows that the ratio tends toward zero as R is made indefinitely great, a condition approximated in voltage amplifiers. According to this expression, the distortion would be eliminated by increasing the external plate resistance. That this is not really so is demonstrated by the analysis above which gives for the same ratio

$$\lim_{R \to \infty} \left(\frac{C_2}{C_1} \right) = (b_{02} + \mu^2 b_{20} - \mu b_{11})/b_{01}.$$

The ratio therefore approaches a constant value different from zero as R is indefinitely increased. The second harmonic level referred to the fundamental is about 40 T.U. down with a 101-D tube, which is prohibitively large distortion for certain classes of work such as multichannel amplification used in carrier telephony; for a 104-D tube the level is about 32 T.U. down on the fundamental.

In order to bring out some important points involved in the theory, we shall discuss them in connection with experimental data on a standard type of tube (101-D) which are due to Mr. A. G. Landeen. The method used in measuring the current components is described in his paper on current analysis in the *Bell System Technical Journal* for April 1927.

Output Currents of a Representative Tube

Fig. 4 shows the calculated effect of varying the plate resistance on the fundamental, second, and third harmonic currents produced by a representative 101-D tube, which are indicated by circles, triangles, and crosses, respectively. In this drawing the values of the coefficients as calculated by Mr. J. G. Kreer are plotted as ordinates, and the external plate resistances are plotted as abscissæ. The agreement with the values obtained from experiment, and shown by the full lines, is seen to be rather close and within the limits of accuracy of the measurements except perhaps for the third harmonic at high load resistances. The coefficients used in the calculation of the quantities C_1, C_2, C_3 were obtained by graphical methods, which consisted in determining tangents to curves derived from μ and R_0 measurements. The precision obtained is sufficient for our present

purposes, but for greater precision it may be desirable to use analytical methods for the determination of the b coefficients.

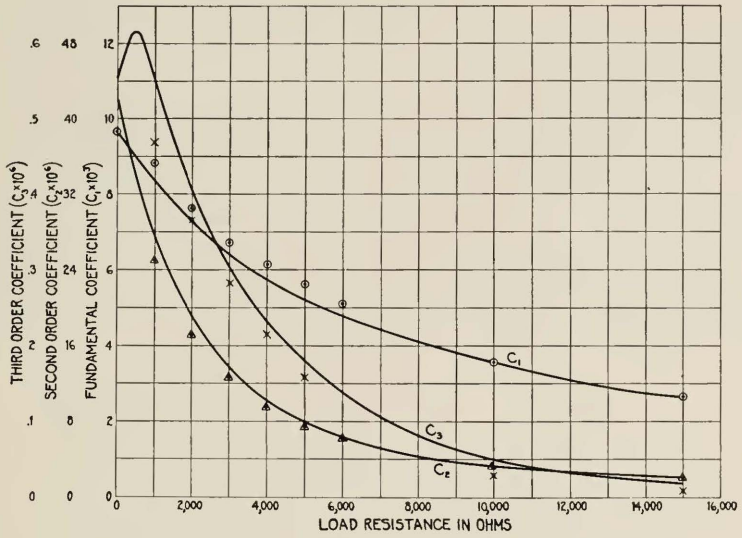


Fig. 4—Modulation coefficients for EL tube No. 109,150. $E_c = -9$, $E_p = 120$

Consideration of the expression for the second order coefficient, C_2, shows that the three terms of the numerator are all important, in general, except that the last term is negligible at very low resistances,

$$C_2 = \left[\frac{\partial \mu}{\partial E_{c_0}} - C_1(R - R_0) \frac{\partial \mu}{\partial E_{p_0}} - R_0 C_1^2 \frac{\partial R_0}{\partial E_{p_0}} \right] \frac{C_1}{2\mu}. \tag{7}$$

Now in amplifiers, the condition for maximum power delivered to the load resistance at maximum gain demands equality of internal and external resistances, and this coincides with the requirement for minimum reflection coefficient [7] when the amplifier is connected to a line of definite characteristic impedance.

Under normal conditions, then, we have for the second order coefficient

$$C_2 = \left[\frac{\partial \mu}{\partial E_{c_0}} - \frac{\mu^2}{4R_0} \frac{\partial R_0}{\partial E_{p_0}} \right] \frac{1}{4R_0}, \tag{8}$$

in which the variation of μ with respect to E_p does not enter, the only determining quantities being the variation of μ with respect to E_{c_0}, and the variation of R_0 with respect to E_{p_0}. The second order modula-

[7] The reflection coefficient is expressed as the quotient of the difference by the sum of the two connected impedances.

tion vanishes when

$$\frac{\partial \mu}{\partial E_{c_0}} \Big/ \frac{\partial R_0}{\partial E_{p_0}} = \mu^2/4R_0, \tag{9}$$

which is a valuable property of amplifier tubes when the equality can be secured.

A more general relation for which the second order vanishes occurs when we set Eq. (7) equal to zero. As a matter of experience these conditions are not satisfied with the usual type of tube; they are found to hold in tubes of rather special construction. The null points are, of course, independent of the character of the applied grid potential, provided that the restrictions on the original development for the tube characteristic are not exceeded and that contributions of higher to lower order terms are negligible.

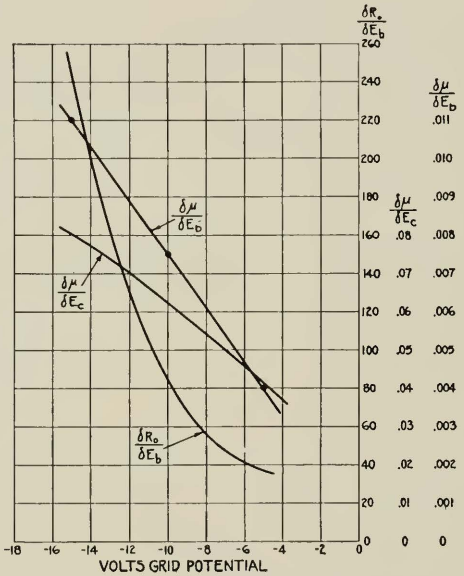


Fig. 5—Variation of tube parameters with grid potential. EL tube No. 109,150. $E_p = 120$

The expression for the third order coefficient contains six terms in the numerator, three of which are of opposite sign. If we consider the contribution of each of these terms as a function of the external plate resistance, we find that at very low resistances the single term b_{03} is predominant, while for resistances comparable to that of the internal plate resistance of the tube itself, no one of the six terms, of which three are negative and three are positive, may be neglected.

As a consequence of the subtraction of quantities of the same order of magnitude, the calculation for the third order coefficient is not capable of any great precision.

Fig. 5 shows how the fundamental coefficients $\partial\mu/\partial E_p$, $\partial\mu/\partial E_c$, $\partial R_0/\partial E_p$, which are involved in the second order term, vary as a function of the grid potential when the plate potential is maintained constant at 120 volts; $\partial R_0/\partial E_p$ is negative in sign.

Variation of C_1 and C_2 with Grid and Plate Potentials

To summarize our analysis up to this point, we have formulated an expression for the characteristic surface of a vacuum tube and have manipulated it to derive expressions for the fundamental and for the second and third order current coefficients. These theoretical

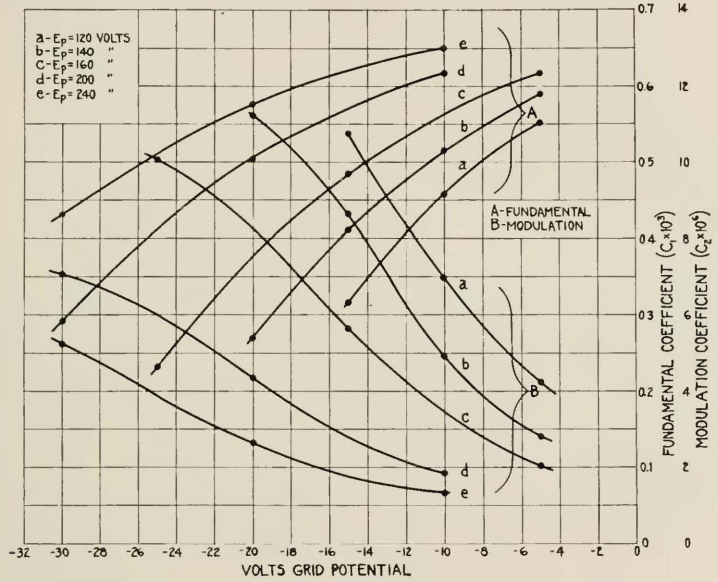


Fig. 6—EL tube No. 109,150. External resistance = 6,000 ohms. I_f = 1.1 amperes

relations have been compared with experimental determinations of the three quantities involved as a function of the external plate resistance, and a sufficiently good agreement has been obtained to indicate that the processes which we have treated are sufficient to account for experimental observations. We now present calculations of the coefficients C_1 and C_2 as a function of plate potentials and grid potentials for several values of the external plate resistance. It is seen from Figs. 6, 7, and 8 that these coefficients vary inversely with the plate and grid potentials.

As the plate resistance is increased all coefficients decrease, but C_2 decreases more rapidly than does C_1. The question then arises as to the conditions which would lead to the smallest amount of distortion while maintaining a definite fundamental power output at a definite

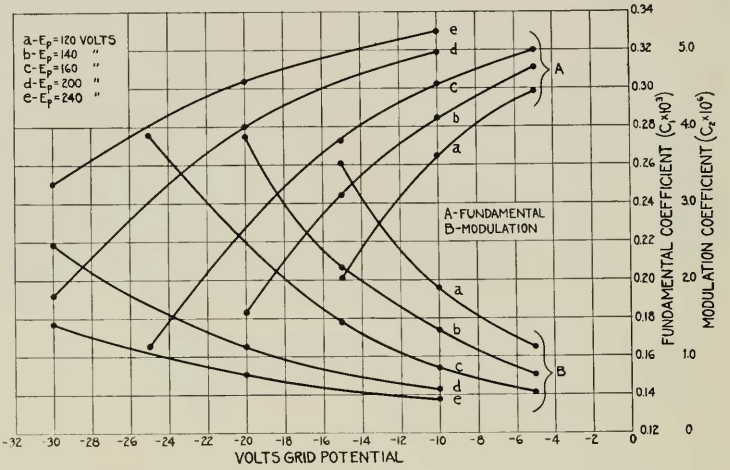


Fig. 7—EL tube No. 109,150. External resistance = 15,000 ohms. I_f = 1.1 amperes

plate potential. This depends evidently upon the desired power. The results in an illustrative case are depicted by Fig. 9, which represents the level of second harmonic current referred to the fundamental current (Δ) plotted as a function of the external plate resistance. At the point of minimum distortion—in which the second harmonic

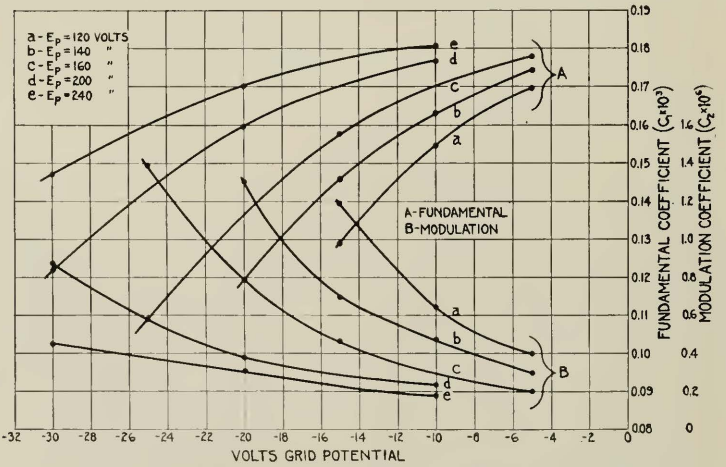


Fig. 8—EL tube No. 109,150. External resistance = 30,000 ohms. I_f = 1.1 amperes

level is 27.5 T.U. down on the fundamental—the external resistance is twice the internal, the grid potential is − 10, and the improvement in the relative reduction of J_2 over customary conditions (for which $R = R_0$ and $E_c = -9$) is about 3.5 T.U. A similar survey made

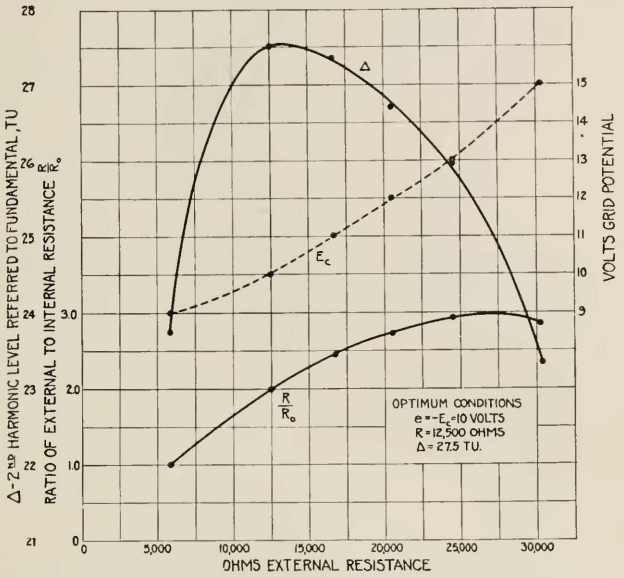


Fig. 9—Ratio of second harmonic to fundamental as a function of load resistance. EL tube No. 109,150. $E_p = 120$, $e = -E_c$ (variable). Power output constant = .056 watt.

with regard to the maximum fundamental power obtainable with constant plate potential, and with plate resistances matched, shows it to be had at − 13.5 volts grid potential (Fig. 10), and to represent a gain of about 35 per cent in output power over that obtainable at the customary operating point. Other considerations such as stability with battery voltage variations operate in repeater practice to fix the grid potential at the customary values.

There has been considerable discussion recently as to the maximum undistorted power obtainable from a tube with fixed plate potential, and variable load resistance and grid potentials. The above analysis shows that in the strict sense of the word, distortion (C_2, C_3) always exists, while with some arbitrary criterion of distortion, results must depend upon the specific criterion adopted. Now as to the maximum fundamental power obtainable, it may be shown that with a parabolic tube characteristic the maximum is obtained for

$$R \doteq \frac{4}{3} R_0,$$

$$E_c \doteq -0.58 E_b/\mu.$$

These values are not very critical, however, since when we put $R = R_0$ and $E_c = -E_b/2\mu$ the power drops by about ten per cent. The last condition is that for maximum power at maximum gain,

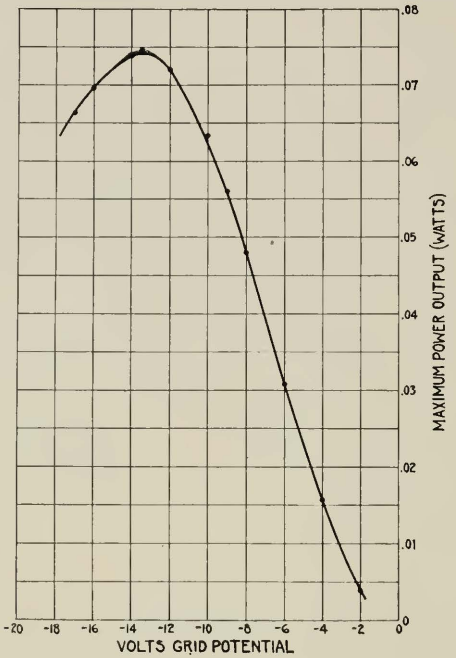


Fig. 10—Maximum power output as a function of the grid potential. EL tube No. 109,150. $E_p = 120$, $E = E_c$, $R = R_0$

in which the output power for a definite alternating grid potential is maximum. In view of the fact that it approximates optimum conditions and is much more convenient as a basis for calculation, we shall use the maximum power—maximum gain criterion of the performance of tubes.

Current and Power Relations.

The preceding analysis has shown that the maximum power obtainable at maximum gain without drawing grid current and without exceeding the negative end of the characteristic—at which grid potential variation produces substantially no plate current variation—is to be found at a grid voltage about midway between the two voltages specified by these conditions. It is again instructive to compare the predictions of the two theories as to the power dissipated in the tube, the a.c. power delivered to the load, the second order modulation, and the relations between these quantities at this

operating point. It is a simple matter to calculate these quantities on the basis of Carson's and of van der Bijl's relations.

Thus the plate current is given by

$$i = \alpha(E_p + \mu E_c)^2,$$

the internal plate resistance to alternating currents is

$$R_0 = 1/2\alpha(E_p + \mu E_c),$$

and the internal plate resistance to direct current is

$$R_{dc} = E_b/\alpha(E_p + \mu E_c)^2.$$

At the operating point for maximum power we have $E_b = -2\mu E_c$, and when the alternating grid potential is equal in amplitude to the grid bias, the above expressions may be manipulated to give

$$\begin{aligned} &1.\ \text{The d.c. power dissipation} \quad P = \alpha E_p^3/4, \\ &2.\ \text{The a.c. power delivered} \quad W = \alpha E_p^3/32, \\ &3.\ \text{The 2d harmonic current} \quad J_2 = \alpha E_p^2/64, \\ &4.\ \text{The fundamental current} \quad J_1 = \alpha E_p^2/4. \end{aligned} \tag{10}$$

From these we have for the ratio of d.c. to a.c. powers

$$P/W = 8,$$

or the efficiency of power conversion at the maximum power condition is $12\frac{1}{2}$ per cent. We find also

$$W/J_2 = 2E_p,$$

or the relation of the fundamental power to the second harmonic current depends upon just one parameter, the plate potential. Other relations of interest are the two following:

$$\begin{aligned} R_{dc}/R_0 &= 4, \\ J_1/J_2 &= 16. \end{aligned}$$

These four relations are independent of tube structure (μ, R_0) and we know that they cannot be accurate in view of the assumptions made in deriving them. In view, however, of the importance of general relations of this type in the design of amplifiers, it is of interest to compare these relations with the ones existing, as calculated by the more accurate theory in which μ variation enters.

Accordingly there are tabulated below the maximum power conditions for a number of tubes of different structure with plate potentials from 120 to 350 volts, operated under the conditions for maximum power output.

CONDITIONS FOR MAXIMUM POWER

Tube	E_b	$-E_c$	μ	R_0	P/W	W/J_2	R/R_0	J_1/J_2
101-D	120	13.5	5.36	8,800	4.8	132	4.55	7.17
101-D	240	28.5	5.43	5,900	4.1	270	4.68	6.95
104-D	120	34.0	2.13	2,930	5.4	132	4.10	7.50
104-D	240	72.0	2.13	2,200	4.3	211	4.54	5.48
205-D	350	32.0	6.95	5,780	3.8	350	5.40	6.53
Special A	250	16.9	9.35	5,800	3.9	267	5.07	6.76
" B	240	9.0	18.1	7,400	3.9	224	4.45	5.50
" B	120	4.4	17.6	9,500	4.5	116	4.14	5.96
" C	130	77.0	1.13	1,430	3.6	131	4.96	6.05

The last four columns represent computations by the double series method which are to be compared with the approximate relations of Eq. (10). Thus J_1/J_2 of the last column is given as 16 when μ-variation is neglected whereas it actually varies between 5.48 and 7.50; R/R_0 on the approximate theory is put at 4, whereas it varies actually between 4.1 and 5.4; W/J_2 is given as $2E_p$ which actually comes out close to half that, and finally the P/W is given as 8, while it really varies between 3.8 and 5.4. On the other hand, the approximate independence of tube structure is shown by these four ratios as given in the last four columns of figures.

Application of the Theory of Probability to Telephone Trunking Problems

By EDWARD C. MOLINA

IF telephone plants were provided in such quantities that when a subscriber makes a call there would be immediately available such switching arrangements and such trunks or paths to the desired point as may be necessary to establish the connection instantly, it would require that paths and switching facilities be provided to meet the maximum demand occurring at any time, with the result that there would be a large amount of plant not in use most of the time.

Obviously, this would result in high costs, particularly in cases of long circuits or where the switching arrangements are complex.

Sound telephone engineering requires, therefore, that we approach this condition only in so far as it is practical and economical to do so, consistent with good telephone service to the subscriber. To take an extreme case—if enough toll lines were provided between New York and San Francisco so that no call would ever be delayed because of busy lines or busy switching arrangements, the rates it would be necessary to charge would be prohibitive, although the speed of service would be very good. Obviously, it is necessary to adopt a compromise between the number of circuits and amount of equipment and the time required to complete a call.

Handling traffic on any other than an instantaneous basis is generally spoken of as handling it on a "delayed basis," even though this delay may be, and generally is, inappreciable to the subscriber. While most of the traffic is handled practically on a no-delay basis, there are certain kinds that are handled on a "delayed basis," such as

1. Calls handled by toll lines, when all toll lines happen to be busy.
2. Calls served by senders or line finders in machine switching systems.
3. Calls handled by operators; the delays implied here being due to the time required by an operator to perform her functions apart from delays due to limitations of equipment.

For traffic handled in this manner it is desirable to have formulas or curves for determining the percentage of calls delayed and the average delay on calls delayed. The product of these two figures will give, of course, the average delay on all calls.

The object of this paper is to present for consideration the results of some theoretical studies made with reference to calls handled on a delay basis.

It is felt that these results may be applied with but slight modifications to many of those traffic problems in which calls are subjected to delay rather than loss when idle mechanisms for advancing them are not immediately available. Such items as service to the subscribers, wear on selecting mechanisms, reliability of circuit operation, etc., are often dependent on the magnitudes of the delays encountered in such cases. Their application to problems in manual traffic is probably less immediate and precise due to the human element which enters into the reckoning. Such factors as an operator's ability to speed up at times of heavy traffic and the facility with which she may reach distant signals appearing before other positions make the problems rather more involved than those dealing with mechanisms whose reactions under various circumstances are more possible to predict. Nevertheless in such cases as these, as well as in problems relating to the delays encountered in clearing trouble conditions, installing telephones, awaiting elevator service, and many other problems of interest to engineers in general, the results and methods discussed here, though probably not directly applicable, may prove highly suggestive in a qualitative way.

Obviously the average number of calls to be handled per unit of time, the average length of holding time per call, and the number of trunks assigned to handle the traffic are factors entering into the mathematical results. A knowledge of these three quantities is not sufficient, however, for the solution of the problem. Quite different results will be secured according to the assumption made as to how the holding times of individual calls vary about their given average, i.e., one set of results follows from the assumption that holding times are all of the same length—other sets of results if holding times vary, the precise set depending on the particular law of variation assumed.

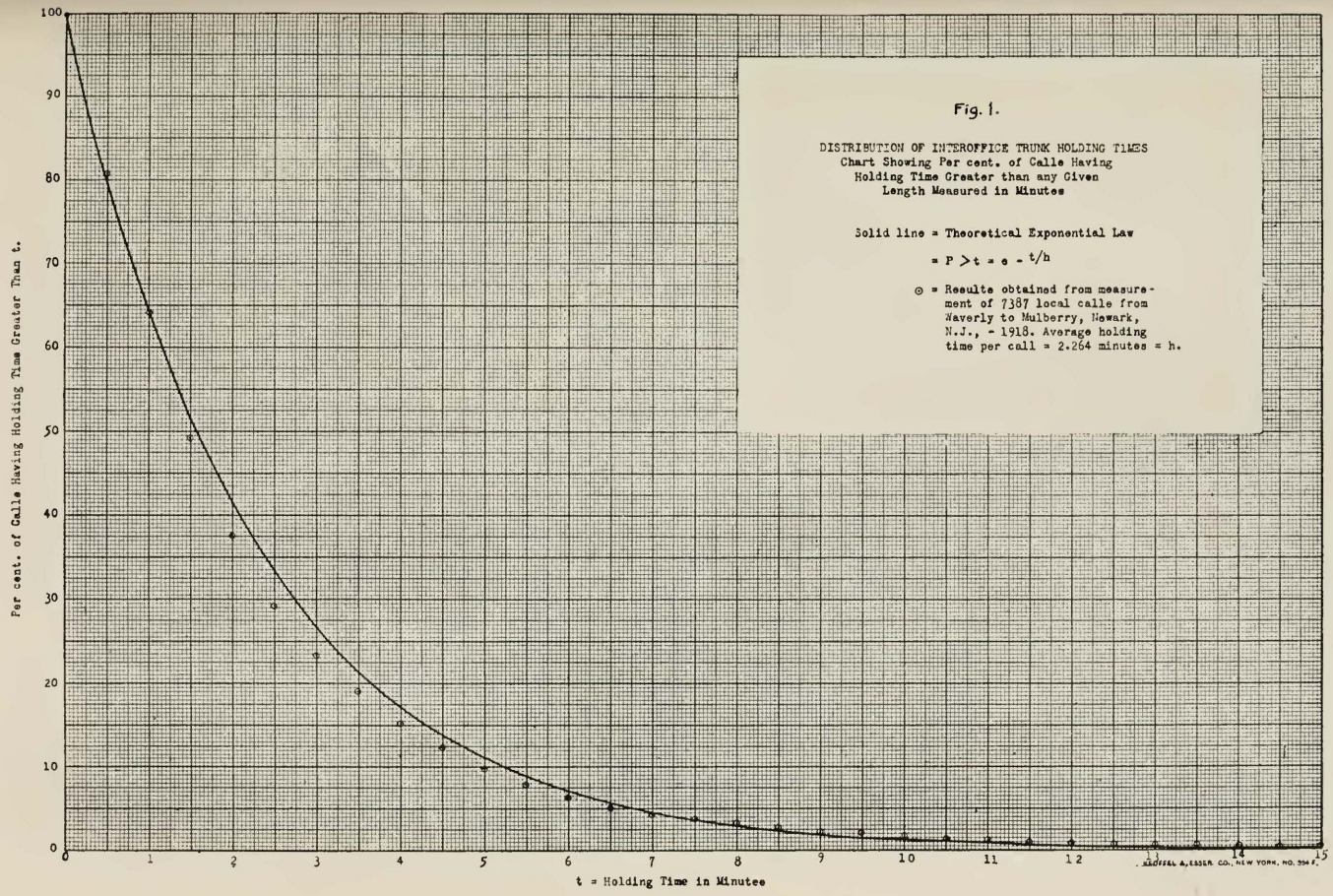
The choice made of laws representing holding time variations must be governed by two considerations:

1. An assumed law must agree at least approximately with the points obtained if we plot the way holding times vary as found from observations.

2. The form of the law must lend itself to the mathematical solution of the delay problem.

Case No. 1

An assumption which permits of an easy and exact mathematical solution of the problem may be stated as follows: If a call is picked


Fig. 1.
DISTRIBUTION OF INTEROFFICE TRUNK HOLDING TIMES
Chart Showing Per cent. of Calls Having
Holding Time Greater than any Given
Length Measured in Minutes
Solid line = Theoretical Exponential Law
= P >t = e - t/h
⊙ = Results obtained from measurement of 7387 local calls from Waverly to Mulberry, Newark, N.J., - 1918. Average holding time per call = 2.264 minutes = h.
Per cent. of Calls Having Holding Time Greater Than t.
100
90
80
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
t = Holding Time in Minutes
KEUFFEL & ESSER CO., NEW YORK, NO. 354 F.

at random, the probability that its holding time is greater than an interval of time of length t is $e^{-t/h}$, where e is the base of natural logarithms and h is the average holding time of all calls. Fortunately there are cases in practice where the variations in length of holding times are closely represented by this exponential law. This is clearly shown by the following Fig. 1 entitled "Distribution of Interoffice-Trunk Holding Times." The points shown on the figure are the plots of actual holding times obtained from pen register records made on a group of trunks running from Waverly to Mulberry in Newark, New Jersey.

Case No. 2

Another assumption covered in this memorandum, because it checks closely with cases arising in practice, is that all calls have exactly the same holding time. The precise solution of the delay service problem becomes extremely difficult on this assumption of a constant holding time. An approximate solution is presented in this paper. Cases in practice where holding times are essentially constant are those of sender holding times of key indicator trunk groups and cordless B boards.

With reference to either Case No. 1 or Case No. 2 consider a group of a certain number of operators handling a certain number of calls having a certain average holding time. If we double the average holding time and halve the number of calls (so that the operators are busy for the same per cent of time on the average), the per cent of calls delayed will not change, but the average delay on calls delayed as measured in seconds will exactly double. Suppose, for example, we wish to obtain the same average speed of answer to line signals with two different groups or teams of operators, the first handling traffic which requires only a short operator holding time or work interval, and the second handling traffic which requires a longer work interval. If the teams are equal in number and ability, we must allow the second team a larger proportion of idle time than the first. If, on the other hand, the proportion of idle time is to be the same for both teams, the second team must be larger or more capable than the first. This general effect is well known, but it is hoped that the results herewith presented will supply more exact knowledge of the subject.

Before proceeding further it will be helpful to give here the notation used on the delay curves following this paper.

h = average holding time per call.

c = number of trunks in a straight multiple.

a = average number of calls originating per interval of time h.

$P\ (>t)$ = probability that a call is delayed for an interval of time greater than t. In other words, 100 times $P\ (>t)$ gives the per cent of calls which will, on the average, be delayed for an interval greater than t.

Charts

The two series of charts following this paper embody, for Case No. 1 and Case No. 2, respectively, curves giving for different values of c and the ratio a/c the probability that a call will be delayed to an extent which will exceed a given multiple of the average holding time. These curves may consequently be read to determine what proportion of the calls will be delayed on the average by an interval as measured in holding times. For example, consider the particular varying holding time chart which corresponds to $c = 10$; we see from the curve marked $a/c = 0.50$ that there is a probability of 0.001 that a call will be delayed for an interval of time which will exceed 0.72 of the average holding time; or, put another way, that 0.1 per cent of the calls will be, on the average, delayed this amount. Again there is a probability of only 0.000019 for a call being delayed at least 1.5 times the average holding time; or 0.0019 per cent of the calls will be delayed by this amount. If, on the same chart, we consider the curve marked $a/c = 0.70$, we find that 22 per cent of the calls will be delayed, 1 per cent will be delayed at least 1.04 times the average holding time, 0.01 per cent will be delayed 2.58 times the average holding time, or more.

The dotted line on each chart gives, at its points of intersection with the curves, the average delay on calls delayed as a multiple of the average holding time interval. For example, on the $c = 10$ varying holding chart we note that for $a/c = 0.70$ the *average delay on calls delayed* is $0.33h$. To obtain the average delay on all calls we multiply by the proportion of calls delayed, $P\ (> 0) = 0.22$, and obtain 0.073 times the average holding time.

A glance at the formulas, given in the Appendix, for the average delay on calls delayed shows that this delay does not reduce to zero when a/c becomes zero. It approaches a lower limit which has the value h/c in Case No. 1 and the value $h/(c + 1)$ in Case No. 2. The latter limit may readily be anticipated from physial considerations as follows. Assume that the group consists of a single trunk; we have to show that the average delay when a call is delayed approaches the limit $h/(1 + 1) = h/2$ as the load approaches zero. Now when the load is very low, those cases where two or more calls have to wait for the trunk to become idle are quite negligible; we only have to

consider the delay incurred by a single call originating while the trunk is busy. But as calls originate at random, the delayed call is just as likely to have fallen near the beginning as toward the end of the constant interval h during which the trunk is busy. In other words, on the average, the delayed call will have originated in the middle of the constant interval h and thus the average delay incurred will be $h/2$. This lower limit for the average delay on calls delayed is indicated in the lower left-hand corner of each sheet of curves by the point where the axis of abscissæ is intersected by a short vertical line.

It will be noted that the constant holding time delay curves change their direction at those points for which the abscissæ are exact multiples of the holding time interval h. No such discontinuities in slope

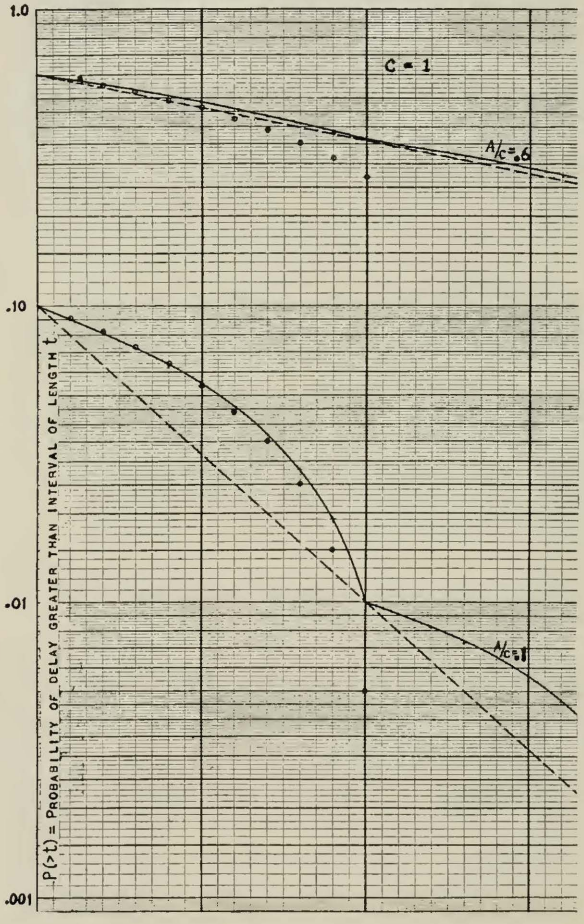


FIG. 4.

appear on the varying holding time curves. This difference in the two classes of curves should occasion no surprise. In the varying holding time case the quantity h has no physical significance; it is merely a numerical value obtained by an algebraic process called averaging. In the other case the quantity h represents a physical characteristic of each and every call.

As stated on page 464 the solution presented in this paper for the case where holding times are all of constant length is not exact. It is desirable therefore to have some idea of the degree of approximation attained.

Figure 4 shows a comparison between our tentative solution and true values which Erlang of Copenhagen, Denmark, succeeded in obtaining by a method which unfortunately becomes impracticable for values of c greater than 3. Our results are shown by the solid curve and Erlang's results by the small circles. We have also indicated on the $c = 1$, 2 and 3 constant holding time charts, Erlang's points for $a/c = 0.50$. For Erlang's work, reference may be had to the Elektrotechnische Zeitschrift, December 19, 1918, page 504.

It may be noted that for the higher values of the ratio a/c the curves are practically straight lines. They depart materially from straight lines for the lower values of the ratio a/c, particularly if c itself is not very large.

Assumptions Made in Mathematical Theory

The mathematical theory back of the curves accompanying this paper is based on the following assumptions:

1. Calls originating independently of each other, and at random with reference to time, have complete access to a single group of trunks.
2. The probability of a call originating during a particular infinitesimal interval, dt, is practically independent of the number of trunks busy or number of waiting calls at the beginning of said interval. This assumption implies that the total interval of time during which the calls fall at random is very large compared with the average holding time per call and that the total number of calls under consideration is very large compared with the number of calls originating per average holding time interval.
3. Calls are served in the order in which they originate. This restriction does not apply to the *average* delays obtained.

4*A*. The average holding time being h, the holding times of individual calls vary around this average in such a way that $e^{-t/h}$ is the probability that for a call taken at random the holding time is greater than t.

4*B*. The holding times of all calls are equal to a constant h.

5*B*. If, at any instant, s of the c trunks are busy, the distribution in time of the instants at which said s busy trunks were seized is identical with the distribution of s points picked individually at random on a straight line of length h.

Assumptions 1 and 2 together imply that the number of *sources* of calls is so large that any blocking of calls due to limitation of sources need not be considered. Assumptions 1, 2, 3 and 4*A* were made in deriving the formulas for Case No. 1. Assumptions 1, 2, 3, 4*B* and 5*B* were made in deriving the formulas for Case No. 2. Assumption 5*B* is not strictly compatible with the physics of the constant holding time case. The distribution in time of the s calls mentioned in 5*B* will, to a certain extent, depend on the history of previous calls. It is because this dependence is ignored that the solution for Case No. 2, presented in the Appendix, is only approximate.

APPENDIX

Mathematical Theory of Delay Formulas

The following mathematical analysis is based on the assumptions given above on pages 467 and 468.

Consider the state of affairs at the instant a particular call "X" originates. Suppose call "X" encounters x other calls; if x is less than c, call "X" will be served immediately but, if not, "X" will have to wait. Our problem is to determine the probability that the delay which "X" may suffer shall have a specified value.

We begin by determining the relative frequency with which the number of calls encountered by "X" has the value x. Let $f(x)$ be the relative frequency with which x calls are encountered by "X." At an instant of time t, x calls will be encountered if at the preceding instant $(t - dt)$ either x, $(x + 1)$ or $(x - 1)$ calls would have been encountered. We ignore as of too rare occurrence the cases where more than $(x + 1)$ or less than $(x - 1)$ calls would have been encountered at time $(t - dt)$ with x at time t. Now in passing from time $(t - dt)$ to time t the probability of an *increase* of one call is proportional to the difference in time, dt, and to the average number of calls, a, falling per holding time interval. Likewise the probability of a *decrease* of one call is proportional to the time difference, dt, and to the number of calls occupying trunks (a decrease must be due to a busy trunk becoming idle). Therefore

$$f(x) = f(x-1)\frac{adt}{h} + f(x+1)\frac{(x+1)dt}{h} + f(x)\left[1 - \frac{adt}{h} - \frac{xdt}{h}\right]$$

when $x < c$, and

$$f(x) = f(x-1)\frac{adt}{h} + f(x+1)\frac{cdt}{h} + f(x)\left[1 - \frac{adt}{h} - \frac{cdt}{h}\right]$$

when $x \not< c$.

For these two equations we may substitute the simpler equations

$$xf(x)\frac{(dt)}{(h)} = af(x-1)\frac{(dt)}{(h)}$$

or

$$cf(x)\frac{(dt)}{(h)} = af(x-1)\frac{(dt)}{(h)}.$$

The solution of these equations gives

$$f(x) = f(0)\left[\frac{a^x}{c^{x-c}\lfloor c}\right], \qquad x \not< c$$

and

$$f(x) = f(0)\frac{a^x}{\lfloor x}, \qquad x \not> c,$$

where $f(0)$ is the arbitrary constant entering in the integration of the finite difference equations. But we must have, evidently,

$$\sum_{x=0}^{x=\infty} f(x) = 1.$$

Substituting in this equation the values for $f(x)$ given above, we obtain

$$1/f(0) = e^a\left[1 - P(c, a) + \frac{a^c e^{-a}}{\lfloor c}\left(\frac{c}{c-a}\right)\right].$$

Since call "X" will be delayed whenever the number x of calls he encounters is equal or greater than c, we have

$$P(>0) = \sum_{x=c}^{x=\infty} f(x) = \frac{\left(\frac{a^c e^{-a}}{\lfloor c}\right)\left(\frac{c}{c-a}\right)}{1 - P(c, a) + \left(\frac{a^c e^{-a}}{\lfloor c}\right)\left(\frac{c}{c-a}\right)}.$$

The next question is to determine the probability, $P(>t)$, of a delay which is greater than an interval of length t.

We will get one answer to this question if we make use of assumption $4A$, and a different answer on the basis of assumptions $4B$ and $5B$. Therefore, from here on, it will be necessary to treat separately the varying and constant holding time cases.

Case No. 1—Holding Times Vary Exponentially

$P\ (> t)$ for this case was obtained by Erlang of Copenhagen. In 1917 he published the formula without its proof. The following deduction of his formula is therefore submitted.

The particular call "X" considered above will be delayed if the number of calls he encountered, x, is equal to or greater than the number, c, of trunks in the group. Suppose that the $x = c + (x - c)$ calls encountered by "X" are handled by the trunks in the manner indicated in the following Fig. 2, where $m_1 + m_2 + \cdots m_c = x - c$.

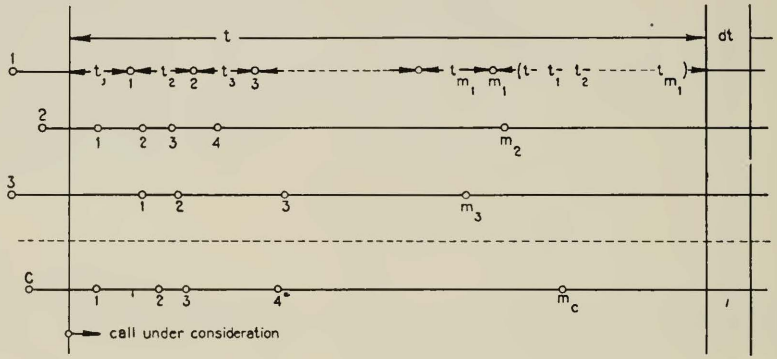


FIG. 2.

By assumption 4A and taking h as the unit of time, the probability that trunk No. 1 will be busy during an interval of time t is

$$(e^{-t_1}dt_1)(e^{-t_2}dt_2)\ \cdots\ (e^{-t_{m_1}}dt_{m_1})e^{-(t-t_1-t_2\ \cdots\ t_{m_1})}.$$

Giving $(t_1, t_2\ \cdots\ t_{m_1})$ all positive values consistent with their sum $\gtrless t$, we obtain (see Todhunter's "Integral Calculus," sixth edition, art. 276)

$$e^{-t}\frac{(t^{m_1})}{(\lfloor m_1)}.$$

Therefore the compound probability that all c trunks are busy during the interval t with the x calls existing at the instant under consideration and that then one of the trunks becomes idle in the interval dt is

$$(e^{-t})^c\left[\sum\left(\frac{t^{m_1}\,t^{m_2}\ \cdots\ t^{m_c}}{\lfloor m_1\ \lfloor m_2\ \cdots\ \lfloor m_c}\right)\right](cdt),$$

where $\sum$ means that we are to give $m_1, m_2\ \cdots\ m_c$ all values such that

$$m_1 + m_2 + \cdots m_c = x - c.$$

By the multinomial theorem

$$\Sigma\left(\frac{\lfloor m_1 + m_2 + \cdots m_c}{\lfloor m_1 \lfloor m_2 \cdots \lfloor m_c}\right)(a^{m_1} b^{m_2} \cdots K^{m_c}) = (ct)^{x-c}$$

for $a = b = \cdots K = t$ and $(m_1 + m_2 + \cdots m_c) = x - c$.

The expression above reduces to

$$(e^{-ct})\left[\frac{(ct)^{x-c}}{\lfloor x - c}\right] cdt.$$

Now all this is on the assumption that "X" encountered x other calls. Therefore we must multiply by $f(x)$ and sum for all values of x from c to ∞. We obtain

$$\sum_{x=c}^{\infty} f(x)e^{-ct}\frac{(ct)^{x-c}}{\lfloor x - c} cdt =$$

$$\sum_{x=c}^{\infty} f(0)\frac{a^x}{c^{x-c}\lfloor c} e^{-ct}\frac{(ct)^{x-c}}{\lfloor x - c} cdt =$$

$$f(0)\left(\frac{a^c}{\lfloor c}\right) ce^{-ct}dt \sum_{x=c}^{\infty}\frac{(at)^{x-c}}{\lfloor x - c} =$$

$$f(0)\left(\frac{a^c}{\lfloor c}\right) cdte^{-ct}e^{at} = f(0)\left(\frac{a^c}{\lfloor c}\right) ce^{-(c-a)t}dt.$$

This is the probability that "X" will be delayed for an interval of length t. To obtain the probability that the delay will be greater than t we must integrate with reference to t from t to ∞. But

$$\int_t^{\infty} e^{-(c-a)t}dt = \frac{e^{-(c-a)t}}{c - a}.$$

Thus, finally,

$$P\,(> t) = f(0)\frac{a^c}{\lfloor c}\left(\frac{c}{c - a}\right) e^{-(c-a)t}$$

$$= P\,(> 0)e^{-(c-a)t}.$$

This formula for $P\,(> t)$ has been deduced by taking h as the unit of time. Evidently we would have obtained

$$P\,(> t) = P\,(> 0)e^{-(c-a)t/h}$$

if h had not been taken as unit of time.

For the average delay on all calls we have

$$\bar{t} = \int_0^{\infty} t\frac{dP\,(> t)}{dt}dt = P\,(> 0)\left(\frac{h}{c - a}\right),$$

and the average delay on calls delayed is

$$\left(\frac{h}{c-a}\right).$$

Case II—Holding Times Constant

Write x, the number of calls encountered by "X," in the form

$$x = nc + m - 1,$$

where n and m are positive integers such that $m \not> c$. In the Fig. 3 below these $nc + m - 1$ calls are shown in groups arranged according

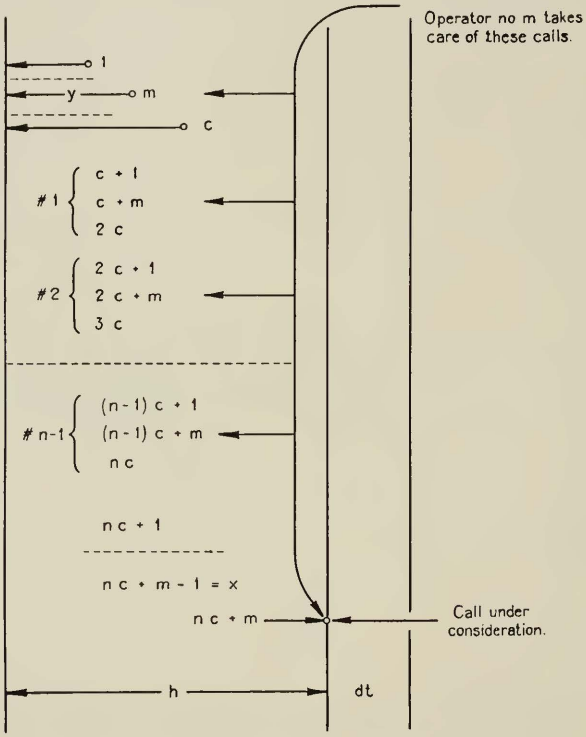


FIG. .3

to the order in which they originated. The first c calls are occupying the trunks. Then $n - 1$ groups, each consisting of c waiting calls, are shown and finally a remainder of $m - 1$ waiting calls; our particular call "X" is the $(nc + m)$th in order of time.

Now evidently, as indicated in the figure, trunk number m will

serve call "X" after said trunk has served the call occupying it and also the mth call in each of the $n - 1$ waiting groups. Therefore our call will suffer a delay of length

$$(n - 1)h + y,$$

where y is the time which elapsed between the beginning of the interval h and the instant at which the mth trunk was seized by the call occupying it. The probability of this delay is a compound one made up of two factors.

1st—The probability that x calls are encountered. This probability is, as derived above,

$$f(x) = f(0)\frac{a^x}{c^{x-c}\underline{c}} = \frac{f(0)c^c(a/c)^{nc}}{\underline{c}}\left(\frac{a}{c}\right)^{m-1},$$

since $x = nc + m - 1$.

2d—The probability that the distance from the beginning of the interval h to the instant at which the mth trunk was seized is y or, in more precise terms, lies between y and $y + dy$. This probability is, on the basis of assumption $5B$,

$$c\left(\frac{dy}{h}\right)\left[\binom{c-1}{m-1}\left(\frac{y}{h}\right)^{m-1}\left(1 - \frac{y}{h}\right)^{(c-1)-(m-1)}\right].$$

The product of these two probabilities gives, writing $y = uh$, $(a/c) = R$,

$$\frac{f(0)c^{c+1}(R)^{nc}}{\underline{c}}\left[\binom{c-1}{m-1}\left(\frac{uR}{1-u}\right)^{m-1}(1-u)^{c-1}\right]du.$$

But the subscribers' interest in a delay of magnitude $(n - 1)h + y$ is totally independent of what value m might have. Therefore the last probability expression must be summed for all permissible values of m, that is from $m = 1$ to $m = c$. We then obtain for the total probability of a delay of extent between $(n - 1)h + y$ and $(n - 1)h + y + dy$:

$$\frac{f(0)c^{c+1}R^{nc}(1-u)^{c-1}}{\underline{c}}\left[\sum_{m=1}^{c}\binom{c-1}{m-1}\left(\frac{uR}{1-u}\right)^{m-1}\right]du =$$

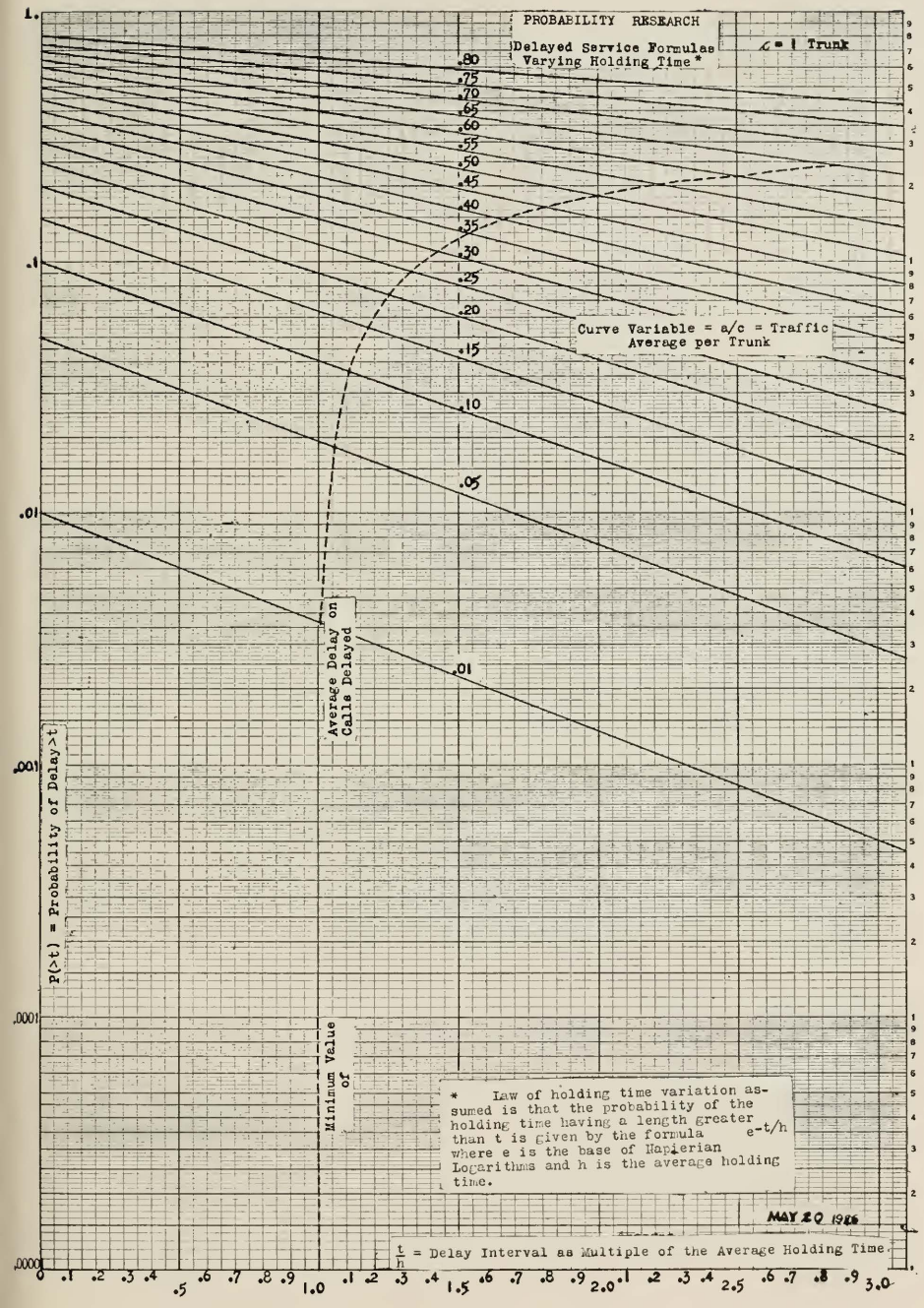
$$\frac{f(0)c^{c+1}R^{nc}(1-u)^{c-1}}{\lfloor c}\left[1+\frac{uR}{1-u}\right]^{c-1}du =$$

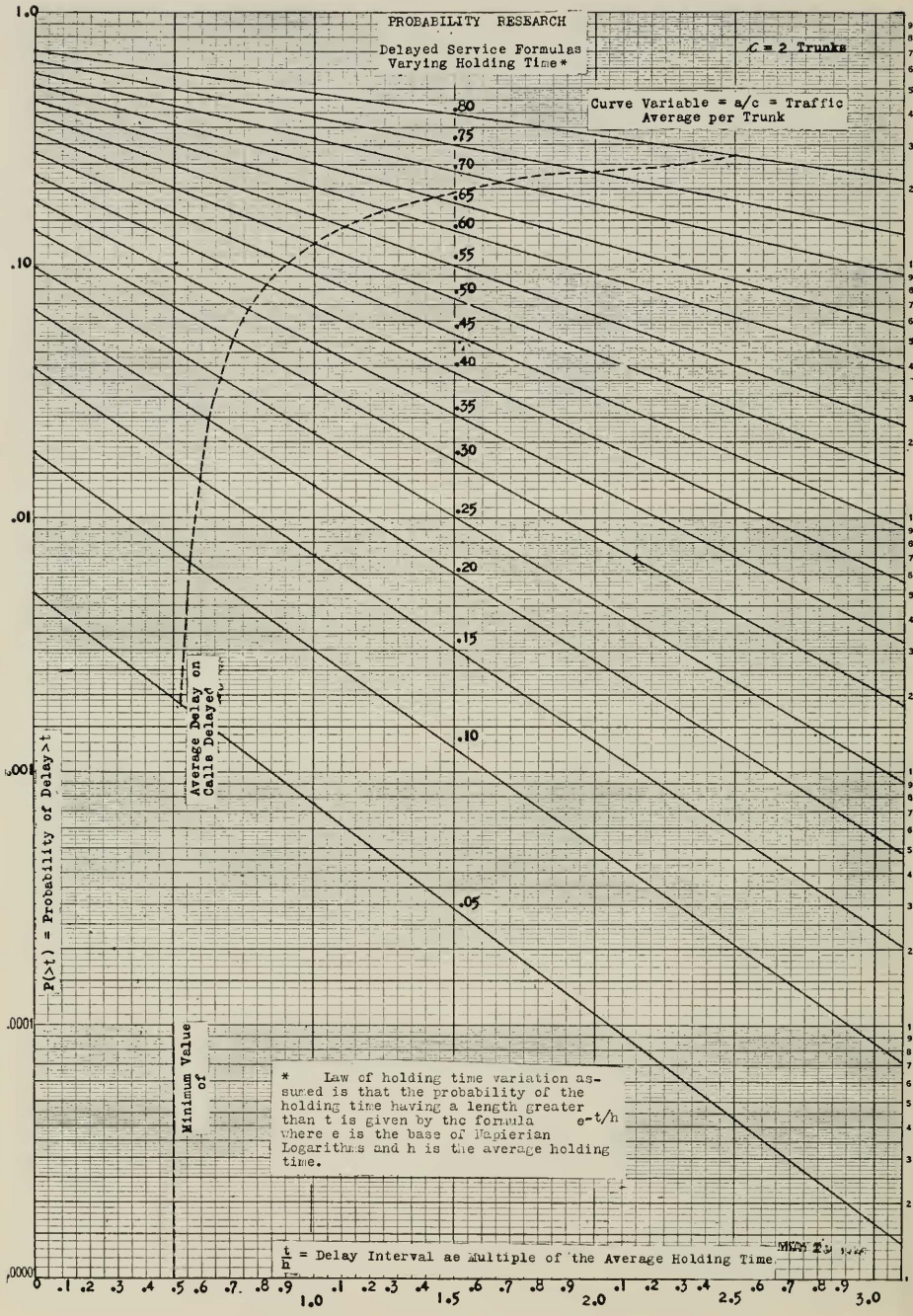
$$\frac{f(0)c^{c+1}R^{nc}}{\lfloor c}[1-(1-R)u]^{c-1}du =$$

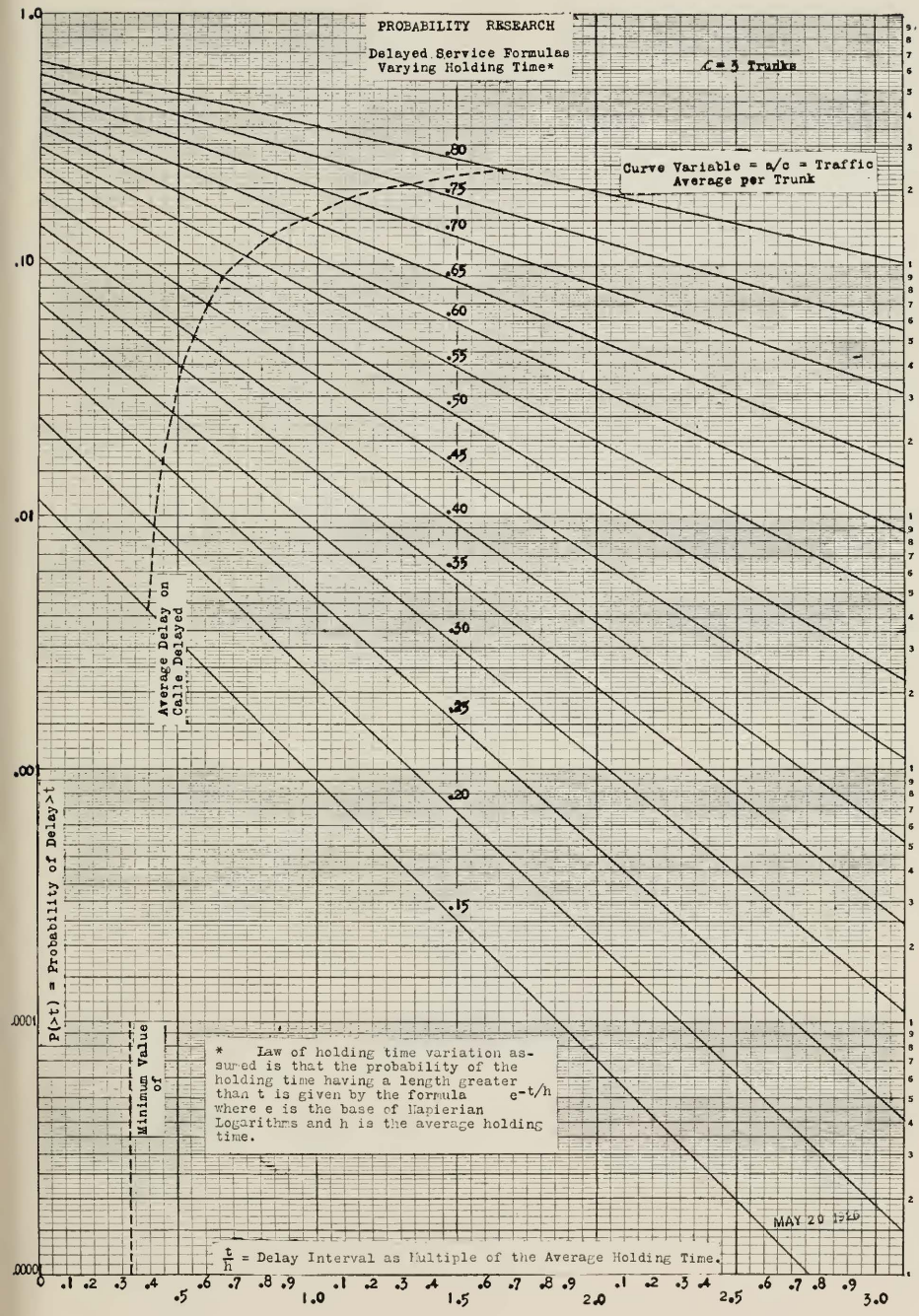
$$P\,(>0)(c-a)R^{(n-1)c}[1-(1-R)u]^{c-1}du.$$

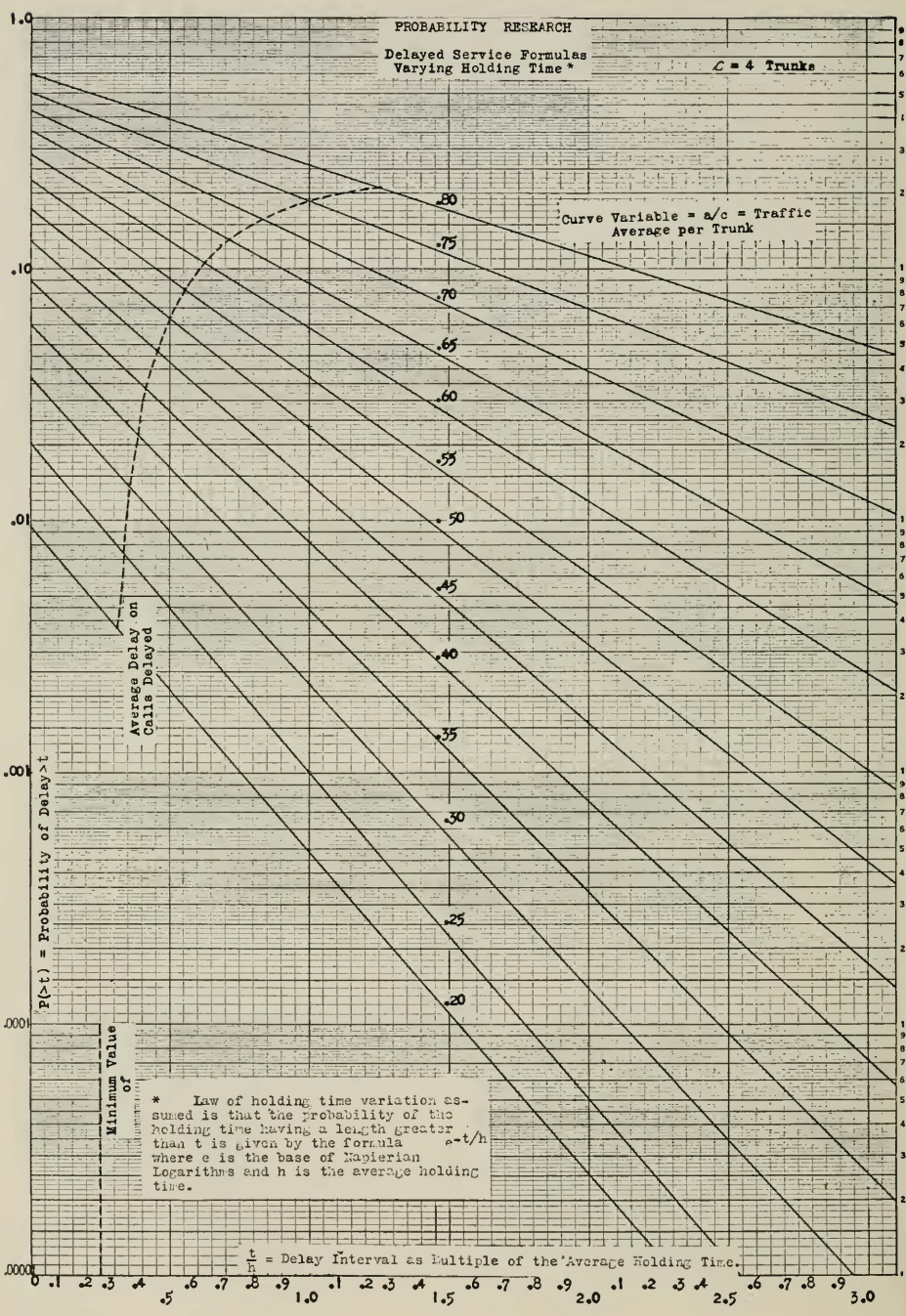
We now obtain by integration with reference to u, summation with reference to n, or both, the following results.

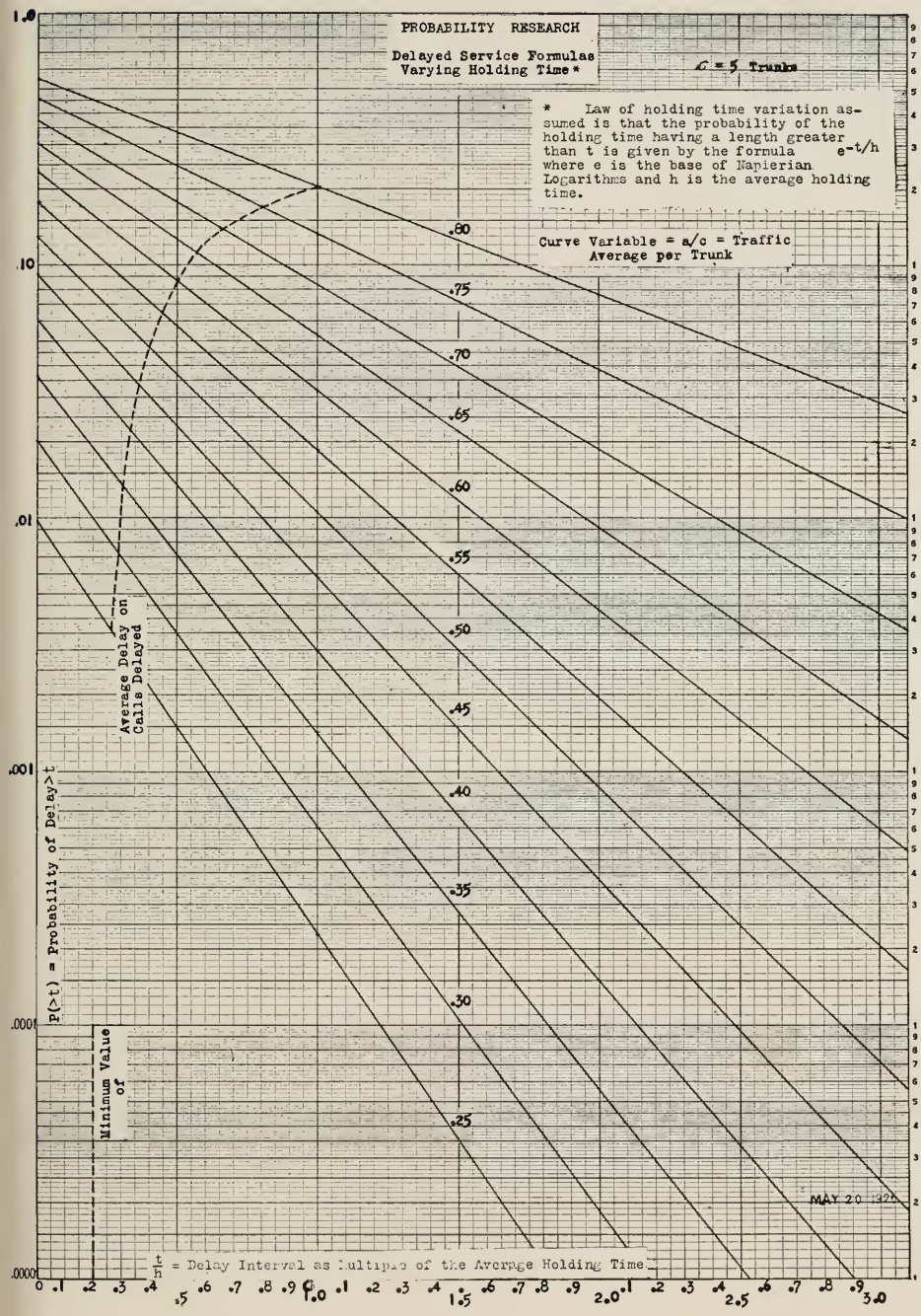
Character of Delay	Probability of Delay
From $(n-1+u)h$ to nh	$P\,(>0)R^{(n-1)c}([1-(1-R)u]^c - R^c)$
From $(n-1)h$ to nh	$P\,(>0)R^{(n-1)c}(1-R^c)$
Greater than $(n-1)h$	$P\,(>0)R^{(n-1)c}$
Greater than $(n-1+u)h$	$P\,(>0)R^{(n-1)c}[1-(1-R)u]^c$
Average delay on all calls	$P\,(>0)\left(\frac{h}{c-a}\right)\left(\frac{c}{c+1}\right)\left(\frac{1-R^{c+1}}{1-R^c}\right)$

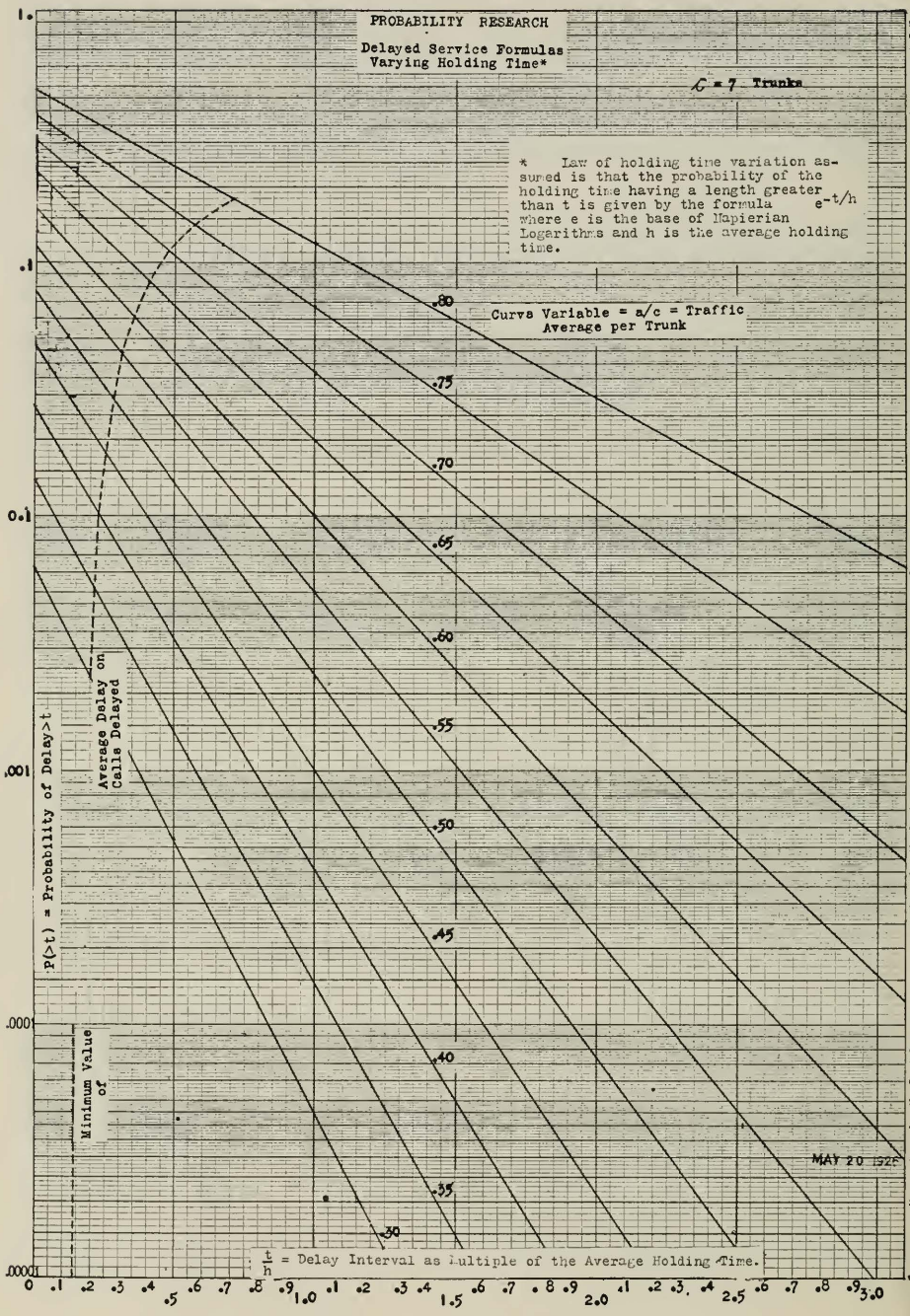

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time*
c = 1 Trunk
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
.15
.10
.05
.01
Average Delay on Calls Delayed
P(>t) = Probability of Delay>t
1.
.1
.01
.001
.0001
.00001
Minimum Value of
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Naperian Logarithms and h is the average holding time.
MAY 20 1926
t/h = Delay Interval as Multiple of the Average Holding Time.
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 .1 .2 .3 .4 1.5 .6 .7 .8 .9 2.0 .1 .2 .3 .4 2.5 .6 .7 .8 .9 3.0

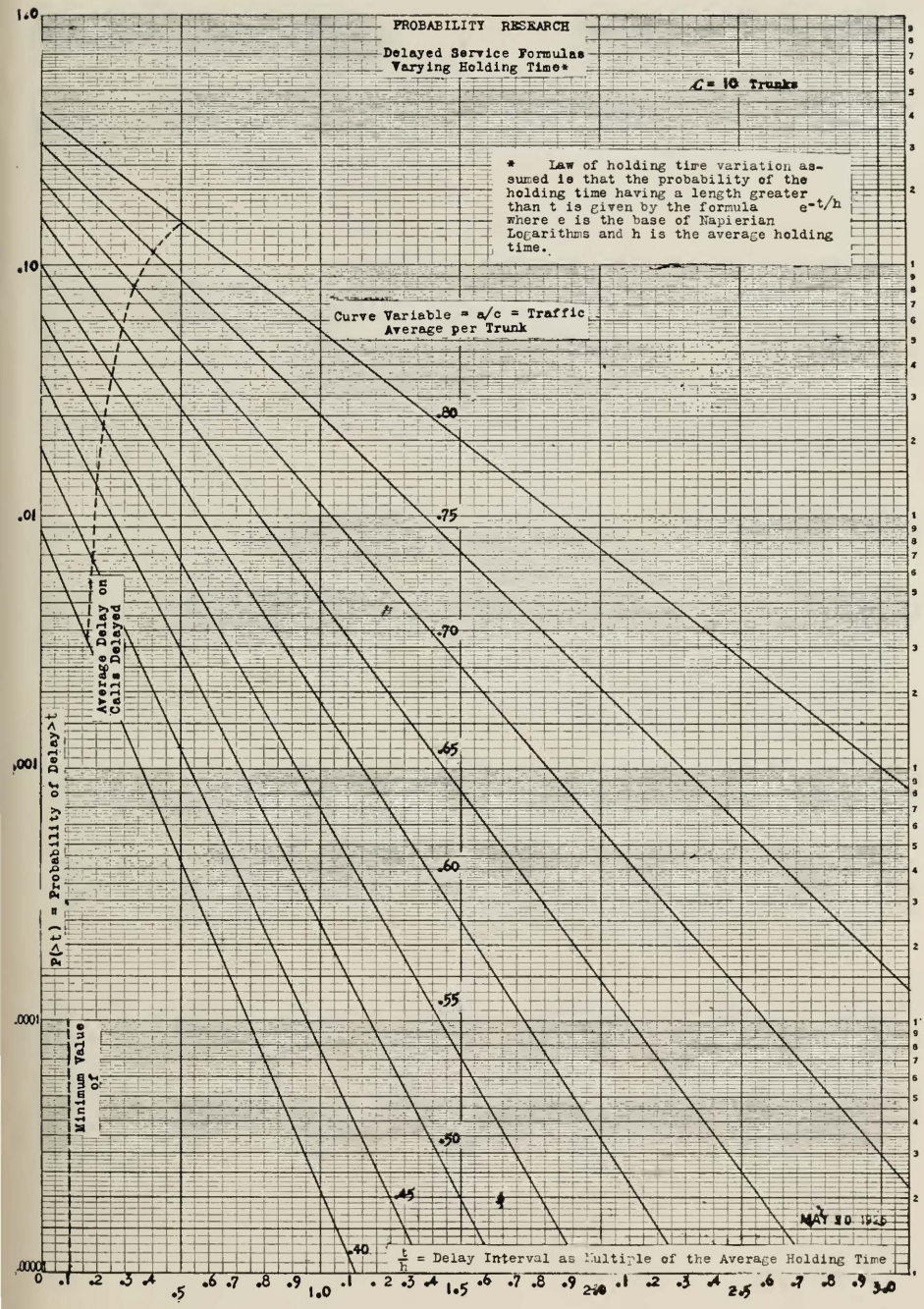

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time*
c = 2 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
.15
.10
.05
1.0
.10
.01
.001
.0001
.00001
P(>t) = Probability of Delay >t
Average Delay on Calls Delayed
Minimum Value of
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Napierian Logarithms and h is the average holding time.
t/h = Delay Interval as Multiple of the Average Holding Time
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 .1 .2 .3 .4 .5 .6 .7 .8 .9 2.0 .1 .2 .3 .4 2.5 .6 .7 .8 .9 3.0

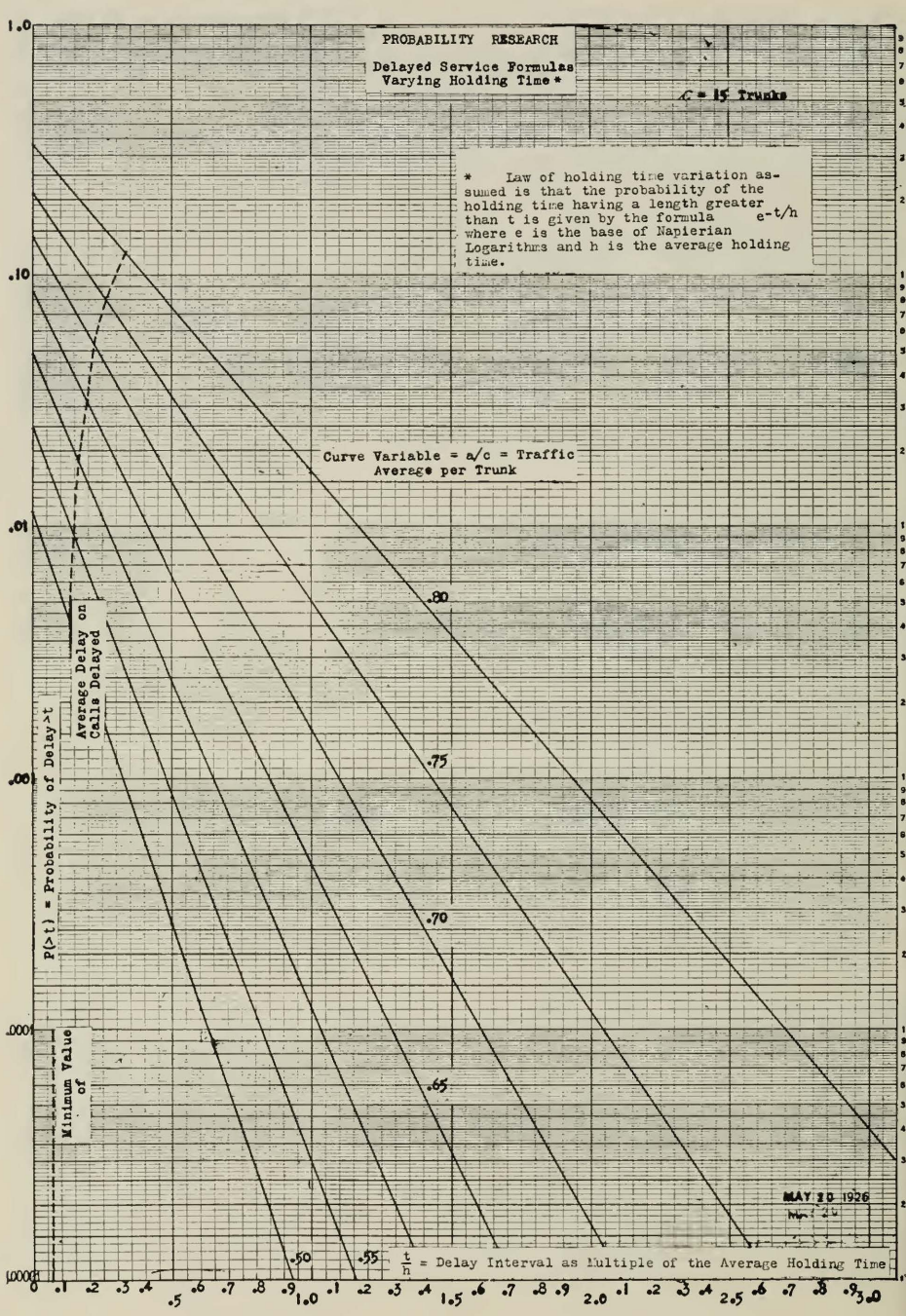

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time*
C = 5 Trunks
Curve Variable = a/c = Traffic Average per Trunk
Average Delay on Calls Delayed
P(>t) = Probability of Delay>t
Minimum Value of
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Napierian Logarithms and h is the average holding time.
MAY 20 1920
t/h = Delay Interval as Multiple of the Average Holding Time.

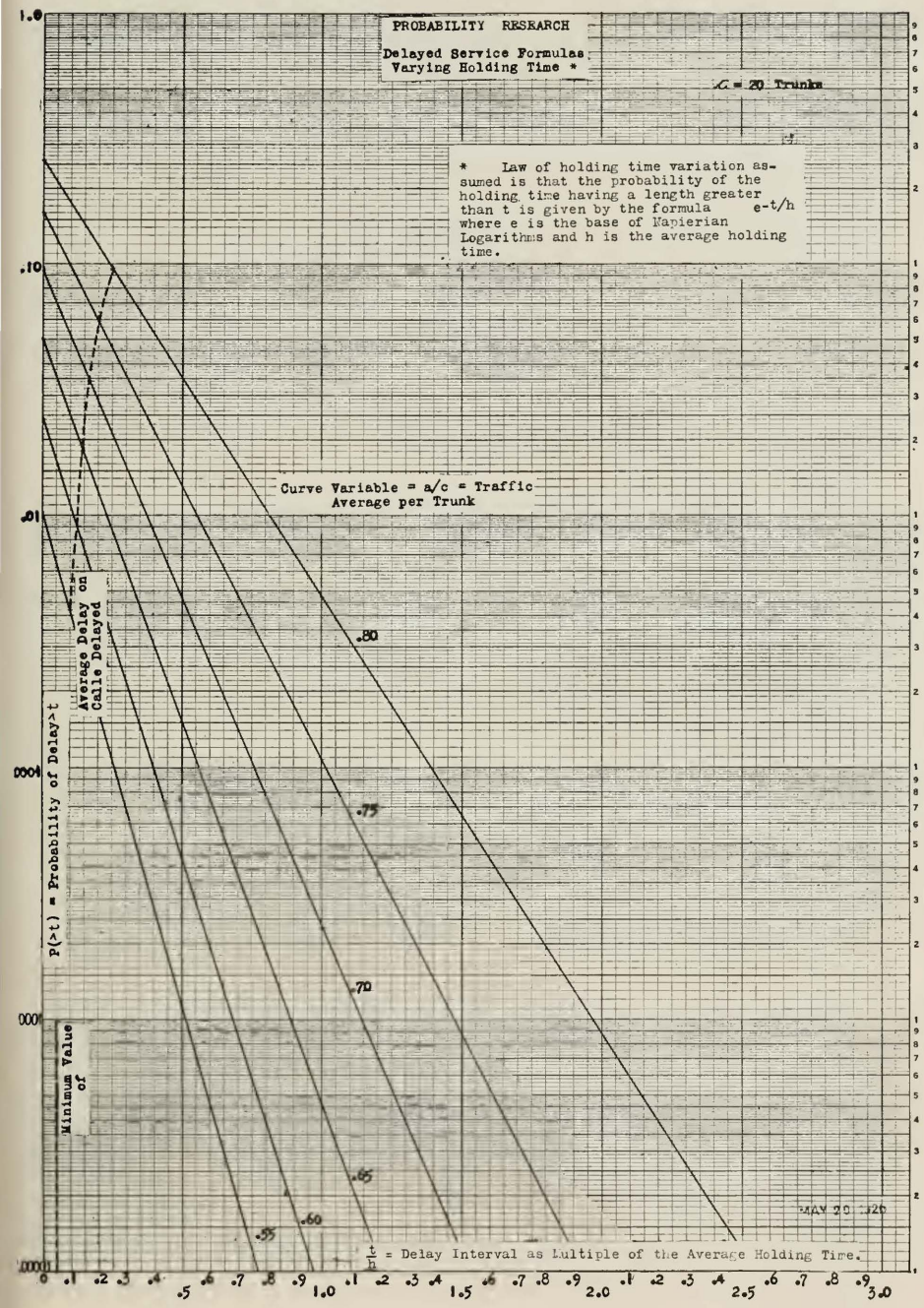

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time *
C = 4 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
Average Delay on Calls Delayed
Minimum Value of
P(>t) = Probability of Delay >t
1.0
.10
.01
.001
.0001
.00001
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Napierian Logarithms and h is the average holding time.
t/h = Delay Interval as Multiple of the Average Holding Time.
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 .1 .2 .3 .4 1.5 .6 .7 .8 .9 2.0 .1 .2 .3 .4 2.5 .6 .7 .8 .9 3.0

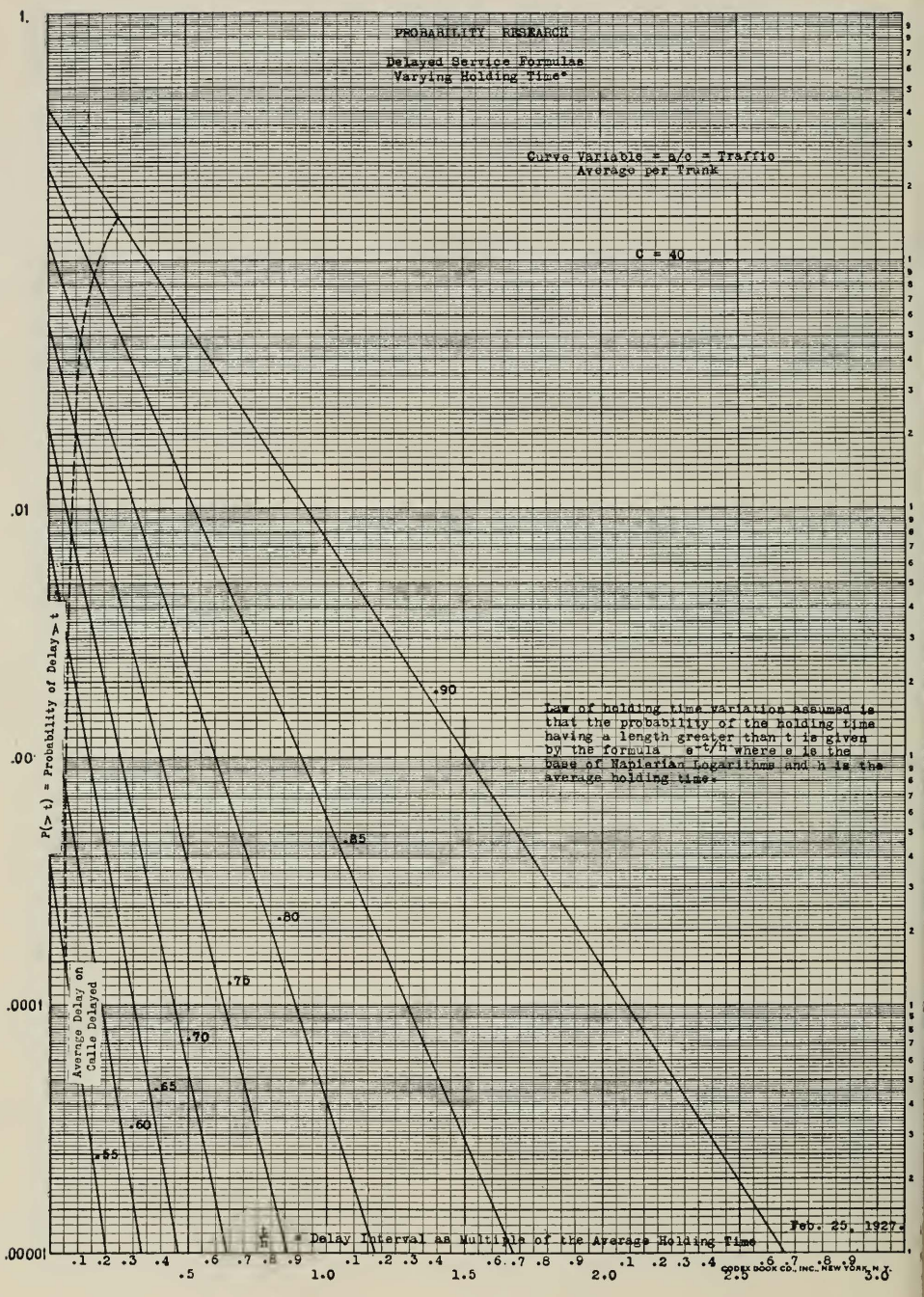

PROBABILITY RESEARCH
Delayed Service Formulae
Varying Holding Time*
C = 5 Trunks
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Napierian Logarithms and h is the average holding time.
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
1.0
.10
.01
.001
.0001
.00001
P(>t) = Probability of Delay>t
Average Delay on Calls Delayed
Minimum Value of
MAY 20 1925
t/h = Delay Interval as Multiple of the Average Holding Time
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 .1 .2 .3 .4 1.5 .6 .7 .8 .9 2.0 .1 .2 .3 .4 2.5 .6 .7 .8 .9 3.0

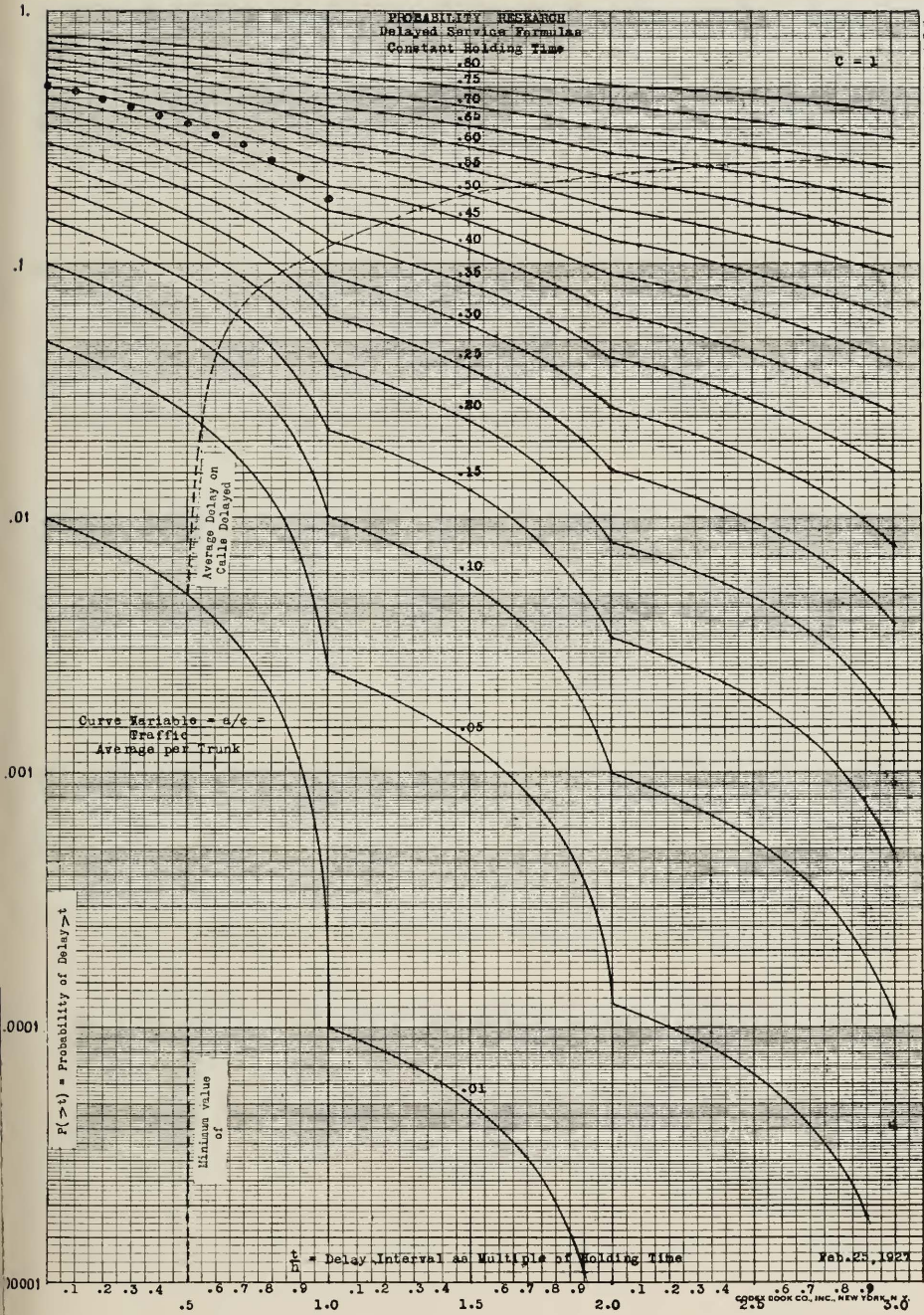

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time*
C = 7 Trunks
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e^-t/h where e is the base of Napierian Logarithms and h is the average holding time.
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
Average Delay on Calls Delayed
Minimum Value of
P(>t) = Probability of Delay>t
1.
.1
0.1
.001
.0001
.00001
t/h = Delay Interval as Multiple of the Average Holding Time.
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 .1 .2 .3 .4 1.5 .6 .7 .8 .9 2.0 .1 .2 .3 .4 2.5 .6 .7 .8 .9 3.0
MAY 20 1925

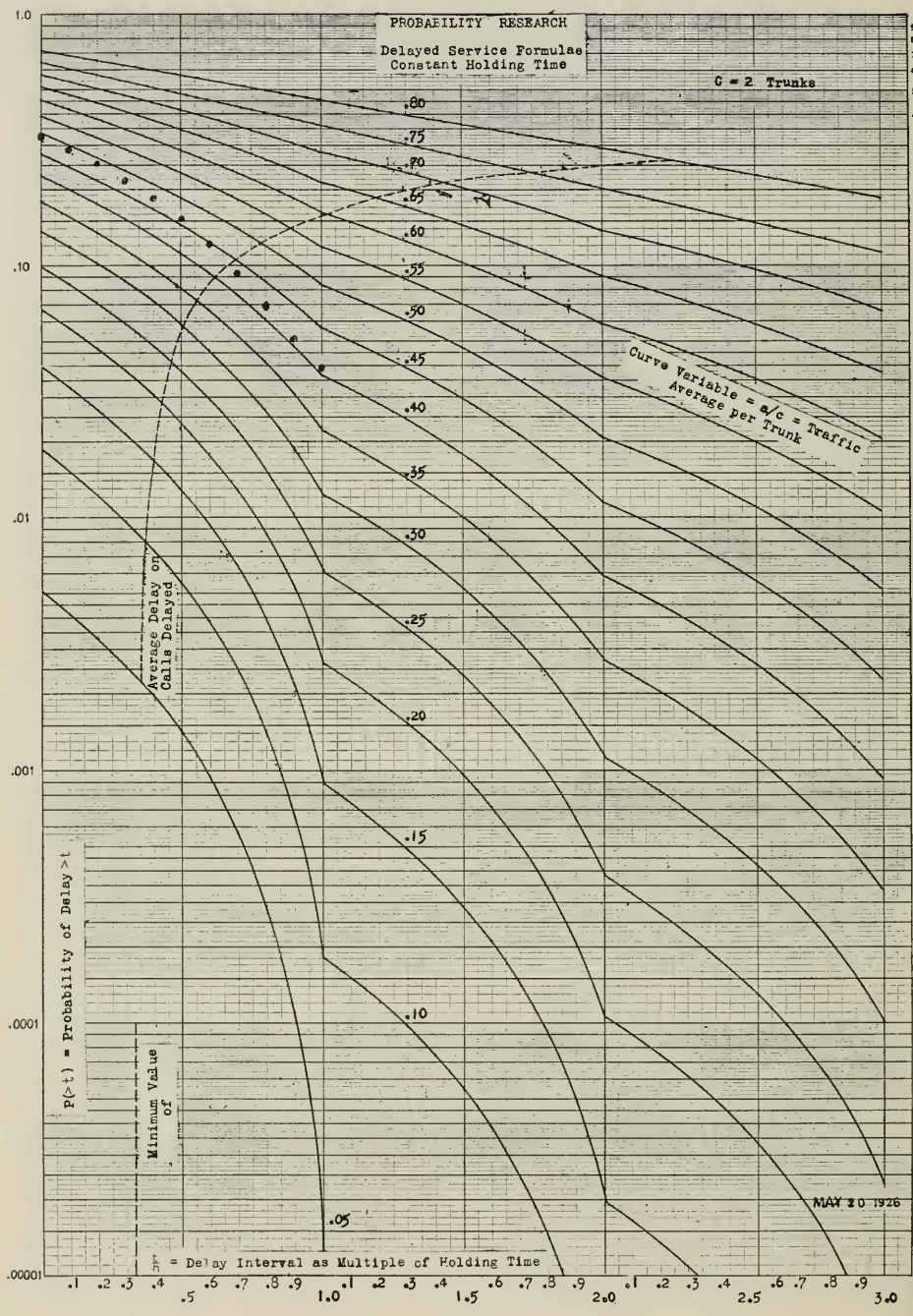

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time*
C = 10 Trunks
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Napierian Logarithms and h is the average holding time.
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
Average Delay on Calls Delayed
Minimum Value of
P(>t) = Probability of Delay>t
1.0
.10
.01
.001
.0001
.00001
t/h = Delay Interval as Multiple of the Average Holding Time
MAY 20 1925

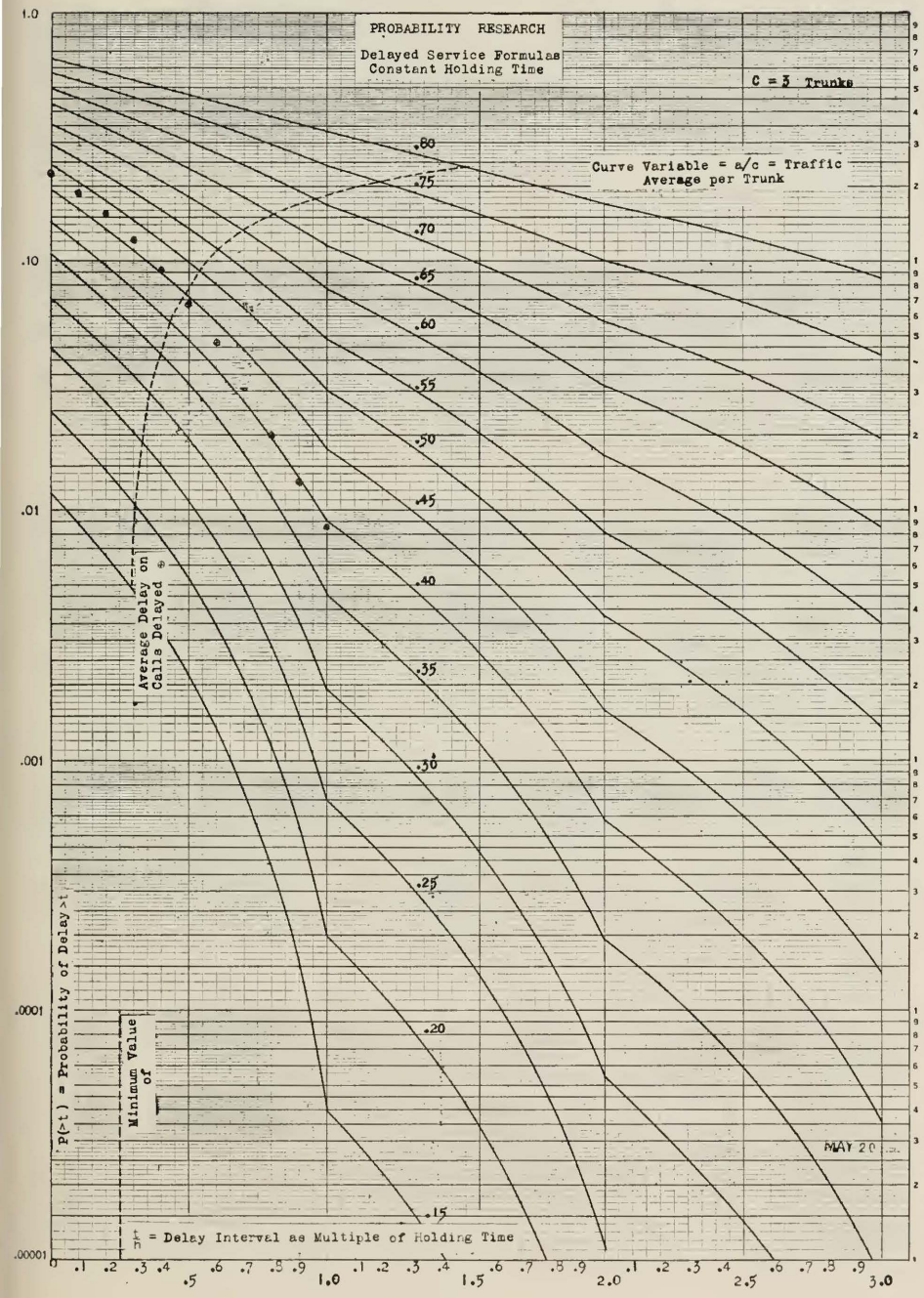

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time *
c = 15 Trunks
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Napierian Logarithms and h is the average holding time.
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.55
.50
P(>t) = Probability of Delay >t
Average Delay on Calls Delayed
Minimum Value of
1.0
.10
.01
.001
.0001
.00001
t/h = Delay Interval as Multiple of the Average Holding Time
MAY 20 1926

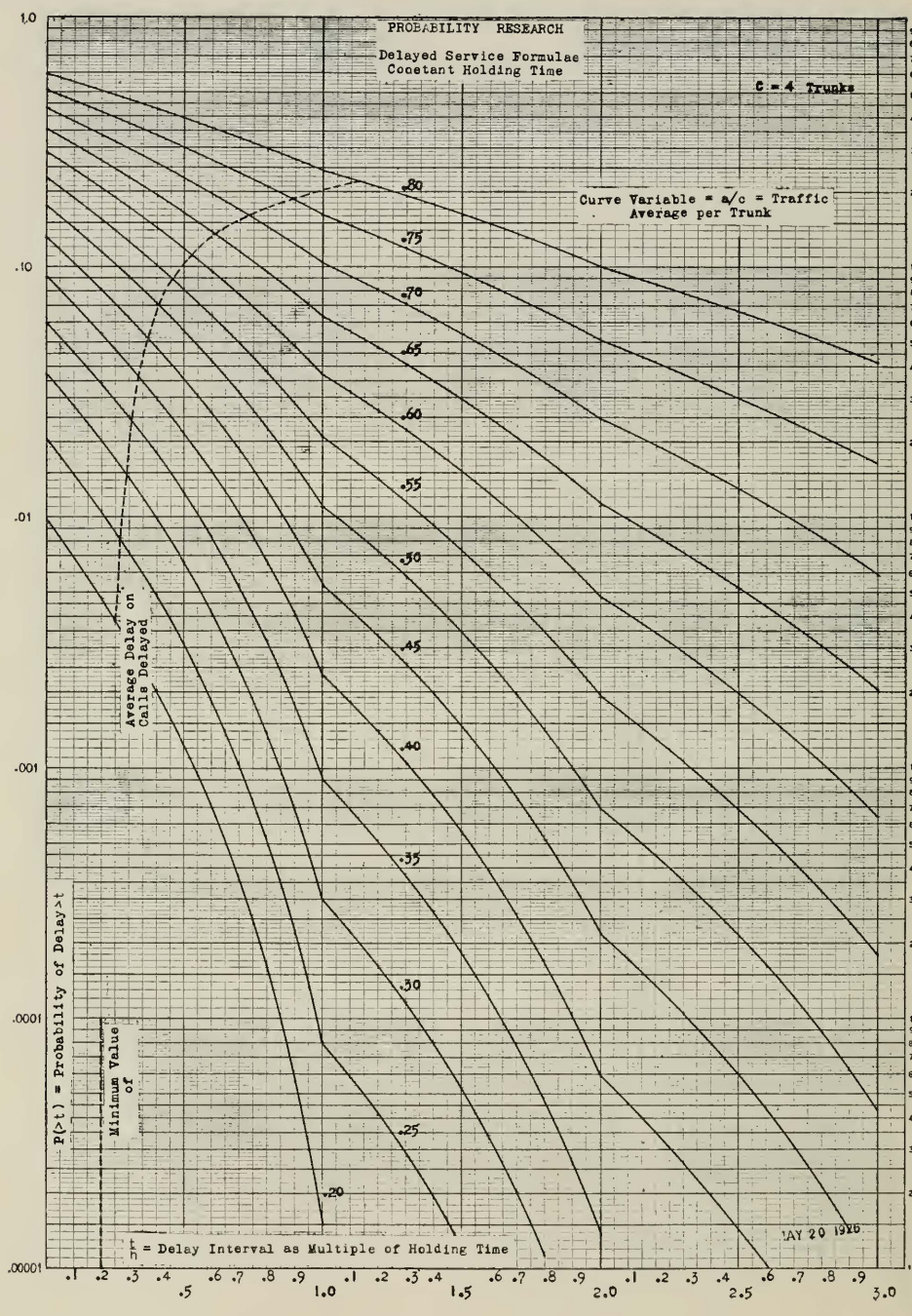

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time *
C = 20 Trunks
* Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e-t/h where e is the base of Napierian Logarithms and h is the average holding time.
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
P(>t) = Probability of Delay>t
Average Delay on Calls Delayed
Minimum Value of
t/h = Delay Interval as Multiple of the Average Holding Time.

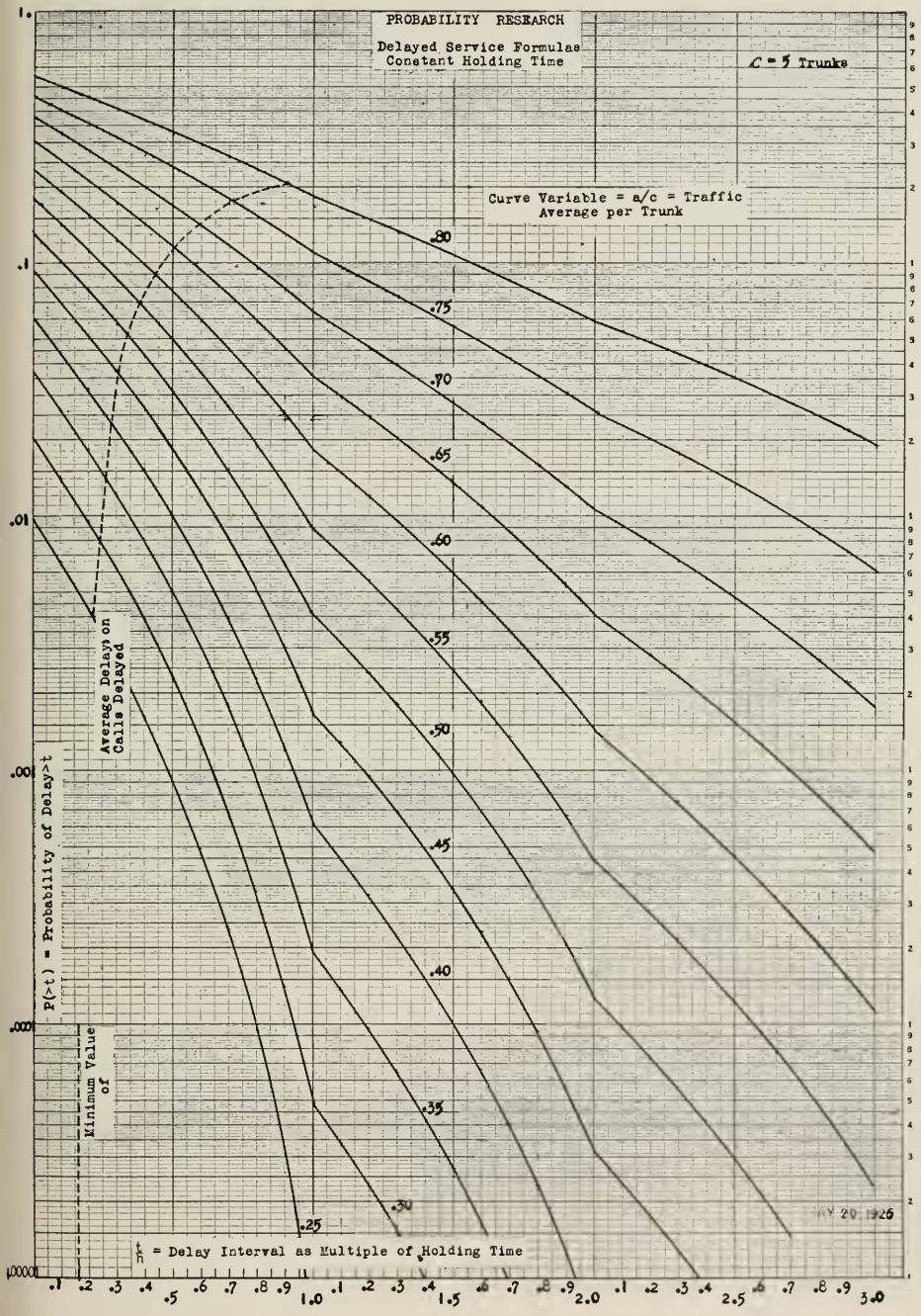

PROBABILITY RESEARCH
Delayed Service Formulas
Varying Holding Time*
Curve Variable = a/c = Traffic Average per Trunk
C = 40
Law of holding time variation assumed is that the probability of the holding time having a length greater than t is given by the formula e^-t/h where e is the base of Napierian Logarithms and h is the average holding time.
P(> t) = Probability of Delay > t
Average Delay on Calls Delayed
.90
.85
.80
.75
.70
.65
.60
.55
1.
.01
.00
.0001
.00001
t/h
* Delay Interval as Multiple of the Average Holding Time
Feb. 25, 1927.
CODEX BOOK CO., INC., NEW YORK, N.Y.

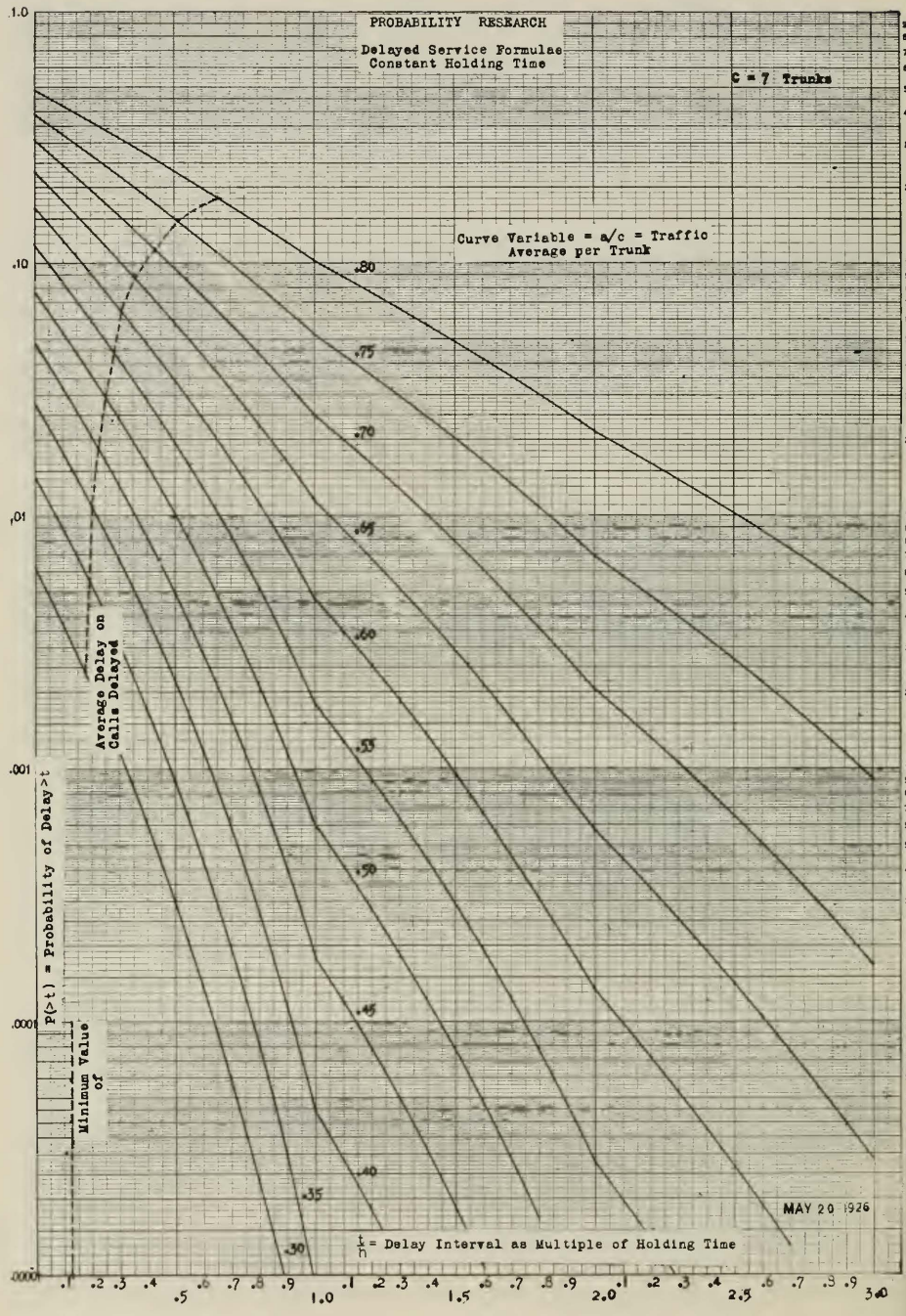

PROBABILITY RESEARCH
Delayed Service Formulas
Constant Holding Time
C = 1
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
.15
.10
.05
.01
Average Delay on Calls Delayed
Curve Variable = a/c = Traffic Average per Trunk
P(>t) = Probability of Delay>t
Minimum value of
t/ht = Delay Interval as Multiple of Holding Time
Feb.25,1927
1.
.1
.01
.001
.0001
.00001
.1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.5 2.0 2.5 3.0

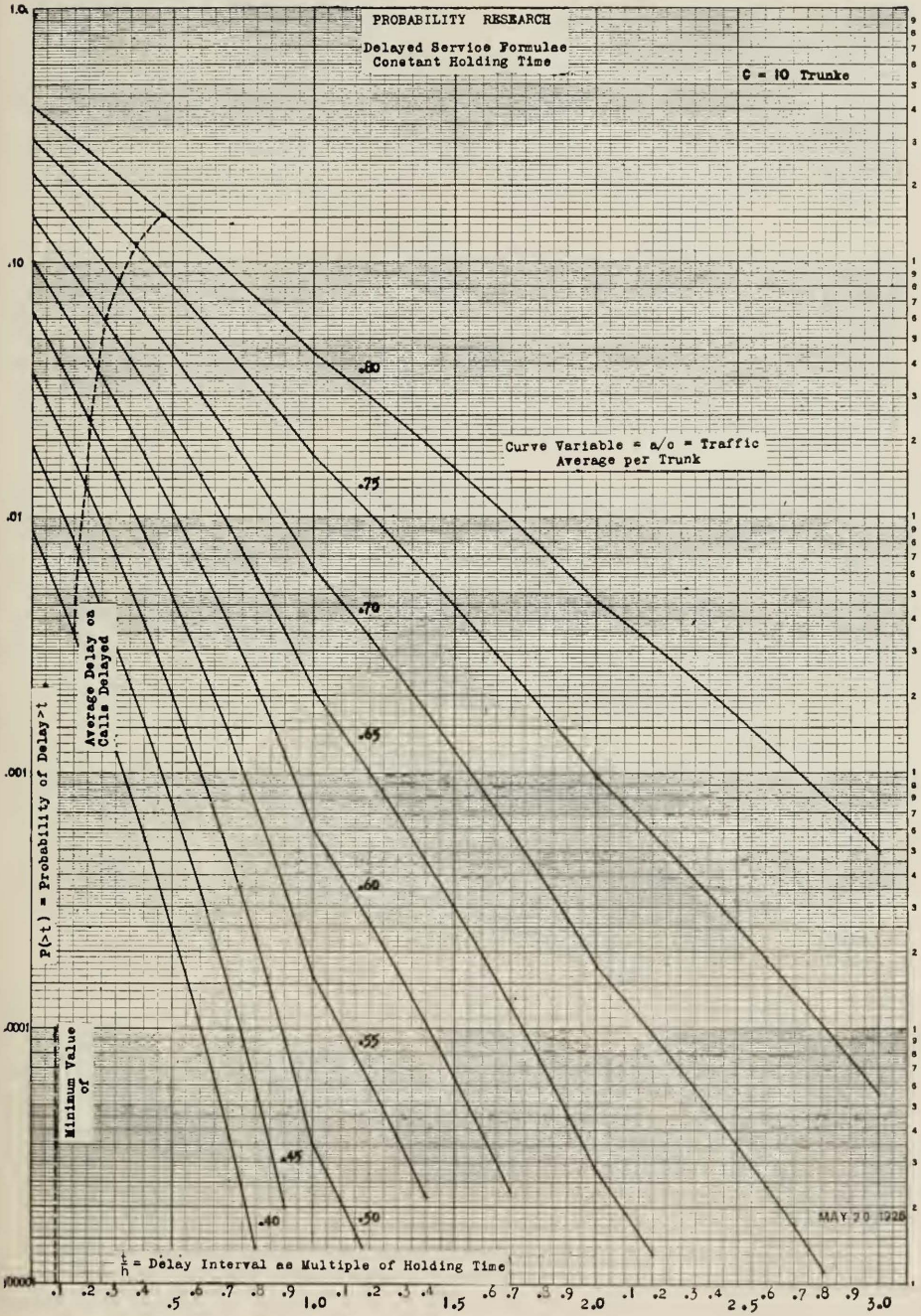

PROBABILITY RESEARCH
Delayed Service Formulae
Constant Holding Time
C = 2 Trunks
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
.15
.10
.05
Curve Variable = a/c = Traffic Average per Trunk
Average Delay on Calls Delayed
P(>t) = Probability of Delay >t
Minimum Value of
= Delay Interval as Multiple of Holding Time
MAY 20 1926
1.0
.10
.01
.001
.0001
.00001
.1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.5 2.0 2.5 3.0

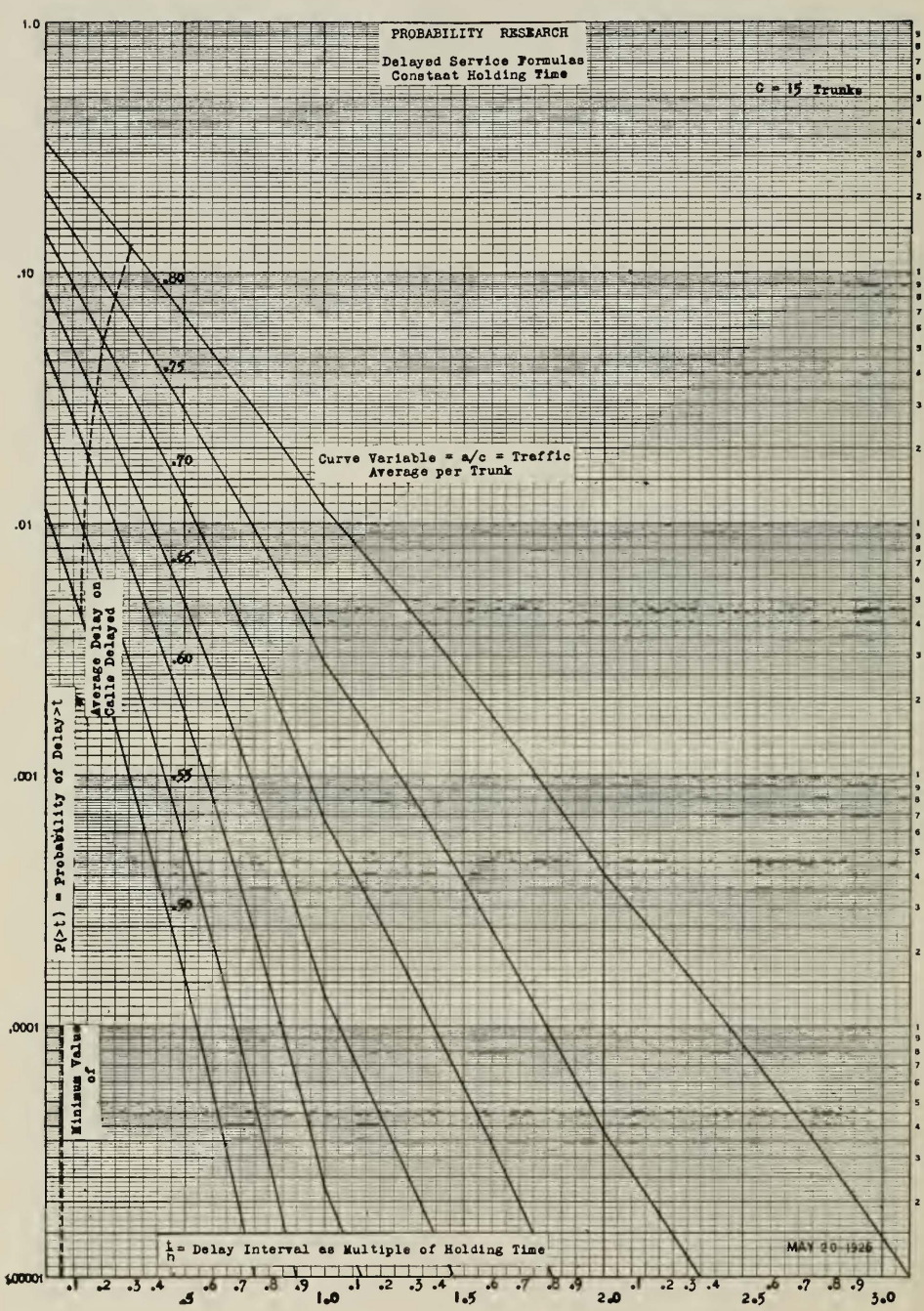

PROBABILITY RESEARCH
Delayed Service Formulas
Constant Holding Time
C = 5 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
.15
Average Delay on Calls Delayed
Minimum Value of
P(>t) = Probability of Delay >t
t/h = Delay Interval as Multiple of Holding Time
1.0
.10
.01
.001
.0001
.00001
.5
1.0
1.5
2.0
2.5
3.0
MAY 20

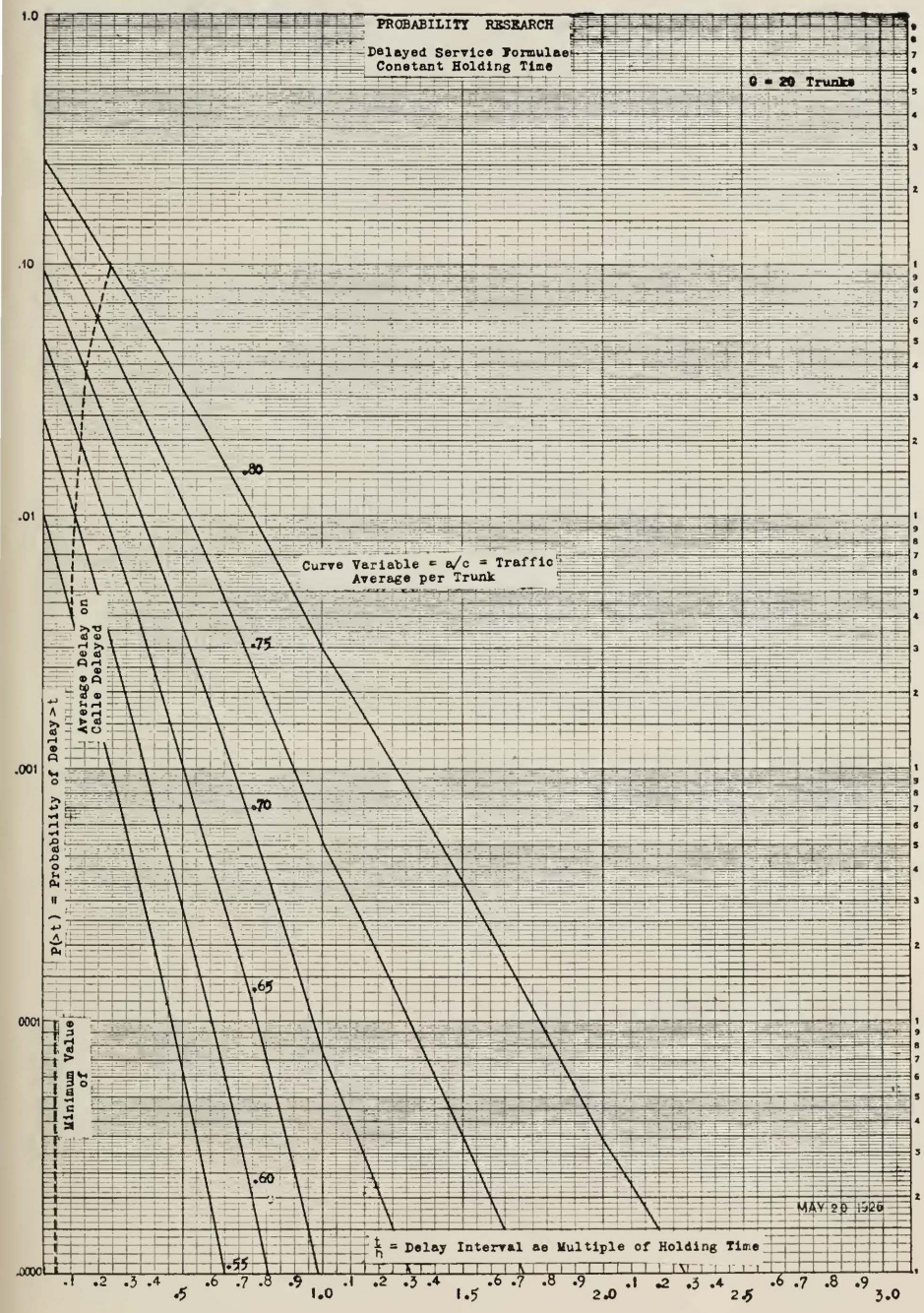

PROBABILITY RESEARCH
Delayed Service Formulae
Constant Holding Time
C = 4 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
.20
Average Delay on Calls Delayed
Minimum Value of
P(>t) = Probability of Delay >t
t/h = Delay Interval as Multiple of Holding Time
1.0
.10
.01
.001
.0001
.00001
.5
1.0
1.5
2.0
2.5
3.0
MAY 20 1926

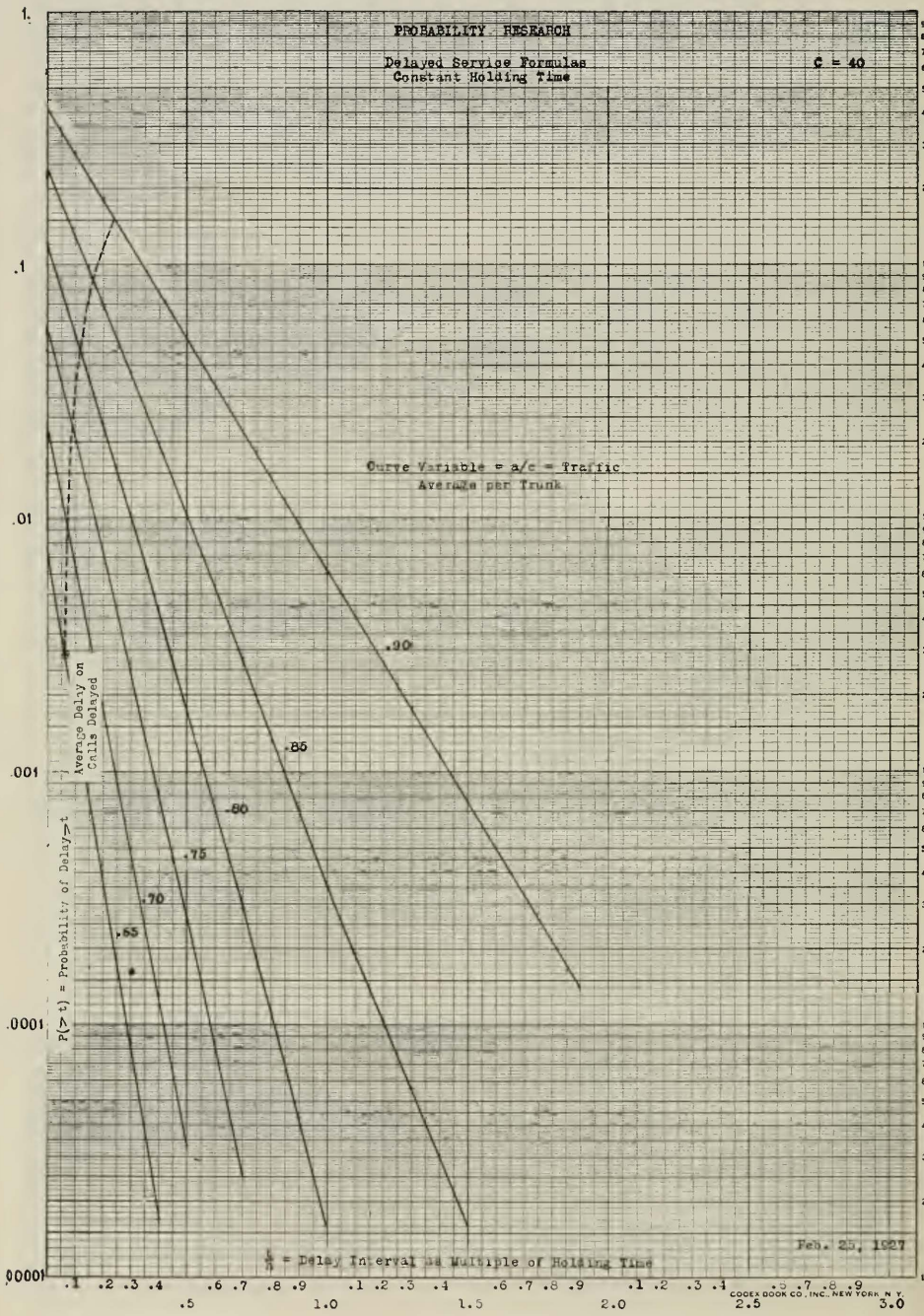

PROBABILITY RESEARCH
Delayed Service Formulae
Constant Holding Time
C = 5 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
.25
Average Delay on Calls Delayed
P(>t) = Probability of Delay>t
Minimum Value of
t/h = Delay Interval as Multiple of Holding Time
1.0
.1
.01
.001
.0001
.1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 .1 .2 .3 .4 1.5 .6 .7 .8 .9 2.0 .1 .2 .3 .4 2.5 .6 .7 .8 .9 3.0


PROBABILITY RESEARCH
Delayed Service Formulae
Constant Holding Time
C = 7 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
.35
.30
Average Delay on Calls Delayed
Minimum Value of
P(>t) = Probability of Delay >t
1.0
.10
.01
.001
.0001
.00001
MAY 20 1926
t/h = Delay Interval as Multiple of Holding Time
.1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.5 2.0 2.5 3.0


PROBABILITY RESEARCH
Delayed Service Formulae
Constant Holding Time
C = 10 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
.45
.40
Average Delay on Calls Delayed
Minimum Value of
P(>t) = Probability of Delay>t
t/h = Delay Interval as Multiple of Holding Time
MAY 20 1926


PROBABILITY RESEARCH
Delayed Service Formulas
Constant Holding Time
C = 15 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
.50
P(>t) = Probability of Delay>t
Average Delay on Calls Delayed
Minimum Value of
t/h = Delay Interval as Multiple of Holding Time
MAY 20 1926
1.0
.10
.01
.001
.0001
.00001
.1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.5 2.0 2.5 3.0


PROBABILITY RESEARCH
Delayed Service Formulae
Conetant Holding Time
C = 20 Trunks
Curve Variable = a/c = Traffic Average per Trunk
.80
.75
.70
.65
.60
.55
P(>t) = Probability of Delay > t
Minimum Value of Average Delay on Calls Delayed
t/h = Delay Interval as Multiple of Holding Time
MAY 20 1926
1.0
.10
.01
.001
.0001
.00001
.1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.5 2.0 2.5 3.0


PROBABILITY RESEARCH
Delayed Service Formulas
Constant Holding Time
C = 40
Curve Variable = a/c = Traffic Average per Trunk
.90
.85
.80
.75
.70
.65
Average Delay on Calls Delayed
P(> t) = Probability of Delay > t
1.
.1
.01
.001
.0001
.00001
t/h = Delay Interval as Multiple of Holding Time
.1 .2 .3 .4 .5 .6 .7 .8 .9 1.0 1.5 2.0 2.5 3.0
Feb. 25, 1927
CODEX BOOK CO. INC. NEW YORK N.Y.

Propagation of Periodic Currents over a System of Parallel Wires

By JOHN R. CARSON and RAY S. HOYT

SYNOPSIS: The first section of this paper is devoted to the formal mathematical theory of the propagation of periodic currents over a system of parallel wires energized at its physical terminals only. The theory developed is essentially a generalization of the classical theory of transmission over a single wire (with ground return) or over a balanced metallic circuit. The solution here given furnishes the fundamental formulas and a good deal of information regarding what takes place in a system of parallel wires; for actual calculations, however, the method of treatment is not so well adapted as that developed in the remaining sections of the paper.

The second section deals analytically with the problem of propagation over a line or a circuit exposed throughout its length to an arbitrary impressed field of force. The resulting solution is immediately applicable to problems of crosstalk and interference, and to the theory of the wave antenna.

The last two sections are devoted to the development and application of a more physical or synthetic method of treatment, based on the substitution of 'equivalent electromotive forces' for the arbitrary impressed field. This synthetic treatment, which permits of an intuitive or physical grasp of the various problems, has been found quite useful in dealing with crosstalk and interference, and also with the wave antenna. The method is illustrated (in the last section) by application to two representative problems of a diverse nature.

IN the modern telephone system, transmission takes place over a circuit which is usually in close juxtaposition to a number of parallel circuits, and which may be, and frequently is, exposed to interference from power circuits or other disturbing sources. The mathematical theory of wave propagation over such a circuit involves two problems: (1) propagation over a system of parallel wires, and (2) propagation over a wire or metallic circuit in an arbitrary impressed field of force.

The first Section of this paper is devoted to the formal mathematical theory of the propagation of periodic currents over a system of parallel wires, energized at its physical terminals only.[1] This problem is essentially a generalization of the problem of transmission over a line of uniformly distributed resistance, inductance, capacity and leakage; and involves the formulation and solution of a differential equation which may be termed the *generalized telegraph equation* in contradistinction to the well-known *telegraph equation* which characterizes transmission over a single wire (with ground return) or a balanced

[1] This is the assumption underlying ordinary transmission theory.

metallic circuit. The analysis of this problem, while furnishing the fundamental formulas and a good deal of information regarding what takes place in a system of parallel wires, is not well adapted for actual calculations, except for relatively simple systems; in particular it is not adapted to deal with the important problems of crosstalk and interference.

In Section II the problem of propagation over a circuit or line exposed throughout its length to an arbitrary impressed field of force is taken up. The resulting solution is immediately applicable to interference problems, where the field of the disturbing source is supposed known, and to the theory of the wave antenna. Moreover, as shown in Section II*a*, it is particularly well adapted to the problems of 'crosstalk,' or interference between circuits of the parallel system.

Sections I, II and II*a* furnish the formal analysis and the fundamental formulas. Sections III and IV, constituting the remainder of the paper, are devoted to the development and the application to representative problems of a more physical or synthetic treatment, in which the general theory and formulas are interpreted in terms of 'equivalent electromotive forces'; this concept permits of an intuitive or physical grasp of the various problems, and has been found quite useful in dealing with crosstalk and interference, and also with the wave antenna.

I

Propagation of Periodic Currents over a System of Parallel Wires, with Impressed Field Concentrated at Terminals

The physical system under consideration is supposed to consist of n parallel wires, numbered from 1 to n, which may either be a system of overhead wires parallel to the surface of the earth, or a multi-wire cable enclosed in a sheath. The formal analysis applies equally well to both cases; but the calculation of the circuit parameters is a matter of considerably greater difficulty in the case of the cable, due to the close juxtaposition of the wires. Even in this case, however, the circuit constants are rather easily calculable to a first approximation from the dimensions of the system; and they are, in any case, experimentally determinable.

Let $I_1, I_2 \cdots I_n$ be the currents in the n wires, which are taken as parallel to the x-axis, which is itself parallel to the surface of the earth or to the sheath (in the cable case). A steady state is assumed; that is to say, the currents are sinusoidal and involve the time t only through the common factor $\exp(i\omega t)$, where $\omega/2\pi$ is the frequency and i denotes $\sqrt{-1}$; consequently the differential operator d/dt is replaceable by $i\omega$ in accordance with the usual methods of alternating current theory.

The first equations of the problem are derived by applying the law

$$\text{curl } E = -\mu(dH/dt)$$

to a contour bounded by a length dx in the surface of the jth wire, a corresponding length dx in the surface of the earth, and two lines normal to the axis of the wire and joining the corresponding ends of the two line elements dx. This gives

$$z_{jj}I_j - E_{gj} = -\frac{dV_j}{dx} - \frac{d\phi_j}{dt}, \qquad (j = 1, 2 \cdots n). \tag{1}$$

In this set of equations, z_{jj} denotes the 'internal' impedance per unit length of the jth wire, that is, the ratio of the axial electric force at the surface of the wire to the current I_j. E_{gj} is the electric force, parallel to the axis of the wire, in the earth's surface. V_j is the line integral of the electric force from the wire to the surface of the earth, that is, the 'potential,' or 'voltage,' of the wire. Finally, ϕ_j is the magnetic flux,[2] per unit length, threading the contour.

Now, both ϕ_j and E_{gj} are linear functions of the n currents $I_1, \cdots I_n$; consequently (1) is reducible to the form [3]

$$z_{jj}I_j = -\frac{dV_j}{dx} - \sum_{k=1}^{n} Z_{jk}I_k, \qquad (j = 1, 2 \cdots n). \tag{2}$$

The calculations of the impedance functions Z_{jk} and, in particular, the effect of the finite conductivity of the earth are dealt with in detail in an earlier paper.[4] The internal impedance, z_{jj}, is of the general form $r_{jj} + i\omega l_{jj}$, where r_{jj} is the resistance of the jth wire and l_{jj} its 'internal inductance.' In the ideal non-dissipative system the mutual impedance Z_{jk} is a pure imaginary of the form $i\omega L_{jk}$, where L_{jk} is the mutual inductance between the jth and kth wires; actually, however, due to the finite conductivity of the earth and to 'proximity effect' between the wires, it is always complex and of the form $R_{jk} + i\omega L_{jk}$. A similar statement holds for the self impedance Z_{jj}. The 'proximity effect' [5] is, of course, the increased internal impedance of the wire due to the currents in the neighboring wires. It may be taken as negligible in open wire lines but is quite appreciable, at telephonic frequencies, in cable circuits.

[2] Expressed in 10^{-8} maxwells if the remaining quantities are in 'practical units.'

[3] It is to be noted that Z_{jj} does *not* include the internal impedance z_{jj} of wire j.

[4] 'Wave Propagation in Overhead Wires with Ground Return,' John R. Carson, *B. S. T. J.*, October, 1926.

[5] See 'Wave Propagation over Parallel Wires: The Proximity Effect,' John R. Carson, *Phil. Mag.*, April, 1921. Rigorously the term $z_{jj}I_j$ of equations (1) and (2) should be replaced by $\Sigma z_{jk}I_k$, the additional terms formulating the proximity effect. This effect will not be explicitly included in the following analysis and the term $z_{jk}I_k$ may be regarded as incorporated with $Z_{jk}I_k$.

Let $Q_1, \cdots Q_n$ denote the charges per unit length on the n wires; the potentials and charges are then related by the set of linear equations

$$Q_j = \sum_{k=1}^{n} q_{jk} V_k, \qquad (j = 1, 2 \cdots n), \tag{3}$$

$$V_j = \sum_{k=1}^{n} p_{jk} Q_k, \qquad (j = 1, 2 \cdots n), \tag{4}$$

in which the q and p coefficients are Maxwell's capacity and potential coefficients. They are calculable by the usual methods of electrostatics, on the assumption that all the conductors, including the earth, are of perfect conductivity.

To complete the specification of the system we have the further set of relations

$$i\omega Q_j + I_j' = -\frac{dI_j}{dx}, \qquad (j = 1, 2 \cdots n). \tag{5}$$

Here I_j' is the 'leakage' current from the jth wire; it is, in general, a linear function of the n potentials, that is,

$$I_j' = \sum_{k=1}^{n} g_{jk} V_k, \qquad (j = 1, 2 \cdots n), \tag{6}$$

where the coefficients g_{jk} depend on the geometry of the system and the conductivity of the dielectric medium. From (3), (5) and (6) we have

$$-\frac{dI_j}{dx} = \sum_{k=1}^{n} (i\omega q_{jk} + g_{jk}) V_k, \qquad (j = 1, 2 \cdots n). \tag{7}$$

This system of linear equations, when solved for the potentials, gives

$$V_j = -\frac{d}{dx} \sum_{k=1}^{n} w_{jk} I_k, \qquad (j = 1, 2 \cdots n). \tag{8}$$

For the special case where the dielectric medium surrounding the conductors is homogeneous and isotropic, the coefficient w_{jk}, which in general is obtained by solving (7), is given by

$$w_{jk} = p_{jk}/(i\omega + \delta), \tag{9}$$

where $\delta = 4\pi\sigma/\epsilon\xi$, σ and ϵ being the conductivity and specific inductive capacity of the dielectric, and ξ a constant whose value depends only on the units.[6] In many cases w_{jk} is calculable with sufficient accuracy from equation (9), so that the solution of (7) is then unnecessary.

[6] A derivation of the formula for δ is outlined shortly after equation (18) of Appendix I.

We are now prepared to write down the *generalized telegraph equation*, which is obtained by eliminating $V_1, \cdots V_n$ from equations (2) by aid of (8); it is:

$$z_{jj}I_j = \sum_{k=1}^{n} \left(w_{jk}\frac{d^2}{dx^2} - Z_{jk} \right) I_k, \qquad (j = 1, 2 \cdots n). \tag{10}$$

This system of n equations, which constitutes the *generalized telegraph equation*, will be written as:

$$\begin{aligned} m_{11}I_1 + m_{12}I_2 + \cdots + m_{1n}I_n &= 0, \\ m_{21}I_1 + m_{22}I_2 + \cdots + m_{2n}I_n &= 0, \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot & \\ m_{n1}I_1 + m_{n2}I_2 + \cdots + m_{nn}I_n &= 0, \end{aligned} \tag{11}$$

where

$$m_{jk} = Z_{jk} - w_{jk}\frac{d^2}{dx^2}, \qquad (j \neq k), \tag{11.1}$$

$$m_{jj} = z_{jj} + Z_{jj} - w_{jj}\frac{d^2}{dx^2}. \tag{11.2}$$

Equations (11) are a system of n homogeneous equations; a finite solution for the currents $I_1, \cdots I_n$ therefore necessitates the vanishing of the determinant of the system, that is,

$$\begin{vmatrix} m_{11} & m_{12} & m_{13} & \cdots & m_{1n} \\ m_{21} & m_{22} & m_{23} & \cdots & m_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nn} \end{vmatrix} = 0. \tag{12}$$

In order to solve this equation the operator d^2/dx^2 is to be replaced by γ^2, which is equivalent to the assumption that the n currents $I_1, \cdots I_n$ involve the variable x only through the common factor $\exp(\gamma x)$. With this substitution, equation (12) is of the nth order in γ^2 and its solution gives, in general, $2n$ values of γ, namely, $\gamma_1, \gamma_2, \cdots \gamma_n$ and $-\gamma_1, -\gamma_2, \cdots -\gamma_n$. The general solution of equations (11) is accordingly of the form

$$I_j = \sum_{k=1}^{n} (A_{jk}e^{-\gamma_k x} - B_{jk}e^{\gamma_k x}), \qquad (j = 1, 2 \cdots n). \tag{13}$$

The potentials are then determined from (8) and (13) in terms of the parameters of the system, the propagation constants, and the arbitrary constants of integration A_{jk}, B_{jk}.

By means of the relations (11) obtaining among the currents, it is

easy to show that the number of independent arbitrary constants of integration is $2n$. These are determined by the $2n$ boundary conditions to be satisfied at the physical terminals of the system. In general these boundary conditions specify $2n$ relations among the impressed voltages, the terminal impedances, and the line currents and voltages. While the evaluation of the constants of integration from these $2n$ relations is formally straightforward, it is actually a matter of very considerable complexity if the system is composed of a large number of wires; furthermore, the evaluation of the propagation constants $\gamma_1, \cdots \gamma_n$ presents great difficulties in such cases.

The results of the foregoing formal analysis may be summarized as follows: In a system of n parallel wires there are in general n modes of propagation, corresponding to the n roots $\gamma_1, \cdots \gamma_n$ of the generalized telegraph equation; these may be termed the normal modes of propagation. Except when special boundary conditions obtain, the current in each and every wire is made up of component waves of all n modes of propagation, and the distribution of energy among the n modes is determined by the boundary conditions at the terminals of the wires. A characteristic and fundamental property of the normal modes of propagation is that a normal mode of propagation is the type which can exist alone. That is to say, if the boundary conditions have a particular set of values, the currents in all the wires may be made up of one mode only; unless, however, the particular conditions obtain, the currents involve components of all modes.

The existence of n modes of propagation in a system of n parallel wires follows from the fact that the determinant is of the nth order in γ^2 and therefore has n roots. In certain cases of practical importance, however, we may have multiple roots, so that the number of distinct modes of propagation is reduced. For example, in the ideal case of perfect conductors and perfect ground conductivity, $L_{ij}/p_{ij} = L_{jk}/p_{jk} = 1/c^2$ and therefore $\gamma = i\omega/c$, where c is the velocity of propagation in the medium. In this case, only one mode of propagation exists, namely, unattenuated transmission with the velocity of propagation of light in the dielectric medium; thus, for the direct wave,

$$I_j = A_j e^{-i\omega/c} \tag{14}$$

and the n constants $A_1, \cdots A_n$ are independent.

Another case of some interest is that in which the wires are all alike, so that $m_{11} = m_{22} = \cdots m_{nn} = m$ and furthermore $m_{jk} = m'$ (a condition which is partially realizable by a properly designed system of transpositions). In this case equation (12) becomes

$$\begin{vmatrix} m & m' & m' & \cdots & m' \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ m' & m' & m' & \cdots & m \end{vmatrix} = 0, \tag{15}$$

and there are only two modes, γ_1 and γ_2, corresponding to $m - m' = 0$ and $m + (n - 1)m' = 0$ respectively. The first mode obviously corresponds to metallic transmission, the second to ground return transmission. It is easily shown that the direct current waves are expressed by

$$I_j = A_j e^{-\gamma_1 x} + Be^{-\gamma_2 x}, \tag{16}$$

$$\sum A_j = 0, \tag{17}$$

with corresponding expressions for the reflected waves. The corresponding potentials are

$$V_j = \tfrac{1}{2} K_1 A_j e^{-\gamma_1 x} + nK_2 Be^{-\gamma_2 x}. \tag{18}$$

Here K_1 is the characteristic impedance of a metallic circuit composed of two wires, and K_2 is the characteristic impedance of the n wires in multiple, with the ground for return.

A case of greater practical importance is that of n balanced pairs, which is the ideal telephone transmission system. To consider this case let

$$\begin{aligned} &m_{11} = m_{22} = \cdots = m_{nn} = m_{2n,\,2n} = m, \\ &m_{jk} = m' \text{ between wires of the same pair,} \\ &m_{jk} = m'' \text{ between wires of different pairs.} \end{aligned} \tag{19}$$

In this case the determinant becomes

$$\begin{vmatrix} m & m' & m'' & m'' & \cdots & \cdots & m'' \\ m' & m & m'' & m'' & \cdots & \cdots & m'' \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ m'' & m'' & m'' & \cdots & \cdots & m & m' \\ m'' & m'' & m'' & \cdots & \cdots & m' & m \end{vmatrix} = 0. \tag{20}$$

There are therefore three modes of propagation γ_1, γ_2, γ_3 corresponding respectively to:

$$\begin{aligned} &m - m' = 0, \\ &m + m' - 2m'' = 0, \\ &m + m' + 2(n - 1)m'' = 0. \end{aligned} \tag{21}$$

The first mode of propagation corresponds to metallic transmission over a pair, the second to transmission over a four wire phantom

circuit, and the third to ground return transmission. It may be easily shown that the solution for the direct waves is (writing I_j and I_j' for the currents in the two wires, j and j', of the jth pair, and V_j, V_j' for the corresponding potentials):

$$\begin{aligned} I_j &= A_j e^{-\gamma_1 x} + B_j e^{-\gamma_2 x} + C e^{-\gamma_3 x}, \\ I_j' &= -A_j e^{-\gamma_1 x} + B_j e^{-\gamma_2 x} + C e^{-\gamma_3 x}, \end{aligned} \tag{22}$$

with the further condition $\sum B_j = 0$. The corresponding potentials are:

$$\begin{aligned} V_j &= \tfrac{1}{2} K_1 A_j e^{-\gamma_1 x} + K_2 B_j e^{-\gamma_2 x} + 2nK_3 C e^{-\gamma_3 x}, \\ V_j' &= -\tfrac{1}{2} K_1 A_j e^{-\gamma_1 x} + K_2 B_j e^{-\gamma_2 x} + 2nK_3 C e^{-\gamma_3 x}. \end{aligned} \tag{23}$$

The characteristic impedances K_1, K_2, K_3 correspond to the three modes of propagation; from the foregoing and equation (8) they are found to have the following values:

$$\begin{aligned} K_1 &= 2(w - w')\gamma_1, \\ K_2 &= (w + w' - 2w'')\gamma_2, \\ K_3 &= \frac{1}{2n}(w + w' + 2[n - 1]w'')\gamma_3. \end{aligned} \tag{24}$$

The importance of the case just considered lies in the fact that the conditions of symmetry which obtain are those of the ideal multi-circuit telephone system. Two of the normal modes of propagation, physical and phantom circuit transmission, are those actually employed in telephone transmission and these modes can exist in any physical or phantom circuit without crosstalk or the induction of current in any other circuit. In fact the problem of crosstalk is the designing of the system, by means of transpositions, to approximate the ideal case, and the calculation of the effect of small departures from the ideal conditions of symmetry.

In investigating problems of crosstalk and of induction from foreign disturbing sources, the general formulas developed in the preceding pages do not lend themselves readily to the necessary calculations and interpretations. In the first place, calculation by means of the general formulas involves the location of the n roots of an nth order equation and thus presents the same difficulties as those encountered in the calculation of the transient oscillations of a network of n degrees of freedom; in fact the problems are mathematically the same, the space variable x of the present problem corresponding to the time variable t of the transient problem. In the second place the formulas, as they stand, are inapplicable to the important case where the circuit

or line is exposed throughout its length to an arbitrary impressed disturbance. Furthermore, in the problem of crosstalk the departures from the conditions of balanced symmetry are necessarily very small, whereas the formulas are so general as to make it very difficult to introduce the essential simplifications which follow from the condition of small departures. For example, if the foregoing formulas are applied, as they stand, to transposed lines, it is necessary to set up new boundary conditions and evaluate a new set of integration constants at every transposition point, since a transposition point is a discontinuity. The difficulty of such a procedure is very great, aside from the fact that it requires as a preliminary the calculation of the n modes of propagation of the system in each transposition interval. In view of these difficulties a more powerful method of attack is required in the analytical investigation of the problems of crosstalk and of interference in general. Fortunately this is furnished by the solution of the problem dealt with in the next section: the propagation of periodic currents over wires in an arbitrary impressed field of force.

II

Propagation of Periodic Currents Over Wires in an Arbitrary Impressed Field of Force

We shall consider first the simplest case, namely, a single wire with ground return. The impressed or disturbing field is assumed to be periodic, of frequency $\omega/2\pi$, so that the problem is a steady state one and the time is involved only through the factor $\exp(i\omega t)$. The resultant field is made up of two parts: first that due to the primary impressed field; and secondly that due to the current in and the charge on the wire, and the corresponding induced currents and charges in the ground. Let $f(x) = f$ denote the component of the electric force of the primary or impressed field parallel to the axis of the wire at its surface,[7] and $F(x) = F$ the line integral of the impressed or primary field from the surface of the wire to ground. It is then proved in Appendix I that the differential equation of the problem is

$$\frac{K}{\gamma}\left(\gamma^2 - \frac{d^2}{dx^2}\right) I = f, \tag{25}$$

the solution of which is [8]

$$\begin{aligned} I = {} & e^{-\gamma x}\left[A + \frac{1}{2K}\int_v^x dy \cdot f(y) e^{\gamma y}\right] \\ & - e^{\gamma x}\left[A' + \frac{1}{2K}\int_v^x dy \cdot f(y) e^{-\gamma y}\right]. \end{aligned} \tag{26}$$

[7] $f(x)$ is assumed to be sensibly constant over the cross-section of the wire.

[8] The lower limit, v, of integration is at our disposal. In case the line begins at $x = 0$, it may be convenient to take $v = 0$.

The corresponding potential of the wire is

$$V = Ke^{-\gamma x}\left[A + \frac{1}{2K}\int_v^x dy \cdot f(y)e^{\gamma y} \right] + Ke^{\gamma x}\left[A' + \frac{1}{2K}\int_v^x dy \cdot f(y)e^{-\gamma y} \right] + F(x). \tag{27}$$

In these equations, γ and K are the *propagation constant* and the *characteristic impedance* of the circuit [9] consisting of the wire with ground return, while A and A' are arbitrary or integration constants which are determined from the boundary conditions. It will be observed that if the arbitrary impressed field is removed ($f = F = 0$), the solution reduces to the usual form. If the terminal impedances are specified, it follows from (26) and (27) that the problem is completely solvable provided that f is specified along the wire and F at its physical terminals.

Two more general cases of practical importance will next be formulated:

(1) Balanced Pair of Wires

Let K_1, γ_1 be the characteristic impedance and propagation constant of transmission over the metallic circuit; and K_3, γ_3 the corresponding quantities for the case of the two wires in multiple, with ground return. Let f_1 and f_2 be the electric force of the primary or impressed field along the surfaces of the wires No. 1 and No. 2 respectively, and I_1 and I_2 the currents in the wires. The solution may then be written as

$$I_1 = a + c, \qquad I_2 = -a + c,$$

where

$$\begin{aligned} a &= e^{-\gamma_1 x}\left[A + \frac{1}{2K_1}\int_v^x \{f_1(y) - f_2(y)\}e^{\gamma_1 y}dy \right] \\ &\quad - e^{\gamma_1 x}\left[A' + \frac{1}{2K_1}\int_v^x \{f_1(y) - f_2(y)\}e^{-\gamma_1 y}dy \right], \\ c &= e^{-\gamma_3 x}\left[C + \frac{1}{4K_3}\int_v^x \tfrac{1}{2}\{f_1(y) + f_2(y)\}e^{\gamma_3 y}dy \right] \\ &\quad - e^{\gamma_3 x}\left[C' + \frac{1}{4K_3}\int_v^x \tfrac{1}{2}\{f_1(y) + f_2(y)\}e^{-\gamma_3 y}dy \right]. \end{aligned} \tag{28}$$

The component a corresponds to transmission over the metallic or physical circuit; while the component c corresponds to transmission over the two wires in multiple, with ground return. A and C are the

[9] It will be observed that in these equations the characteristics of the ground do not appear explicitly. They are, however, implicitly involved in K and γ of the ground return circuit.

integration constants of the direct wave, while A' and C' are those of the reflected wave. The first component, as regards the impressed force f along the wires, depends on the difference $f_1 - f_2$ at the surface of the two wires, while the second depends on the mean value $(f_1 + f_2)/2$. In the case of interference from external sources the latter is usually much the larger and consequently the induction mainly corresponds to the ground return mode of propagation, γ_3.

(2) System of n Balanced Pairs

We shall now write down the expressions for the currents in a system of n balanced pairs ($2n$ wires) when exposed to an arbitrary impressed field. The properties of this system were discussed briefly in the preceding section and formulated in equations (19), $\cdots$ (24). Let I_j and I_j' be the currents in the two wires j and j' respectively of the jth pair, and let f_j and f_j' be the corresponding impressed forces along the surfaces of the two wires, while F_j and F_j' are the corresponding line integrals of the impressed force to ground. By an extension of the previous formulas it is easy to show that the currents are made up of three components:

$$\begin{aligned} I_j &= a_j + b_j + c_j, \\ I_j' &= -a_j + b_j + c_j. \end{aligned} \tag{29}$$

If we write $\bar{f}_j = (f_j + f_j')/2$, the components a_j, b_j, c_j are given by:

$$\begin{aligned} a_j = {} & e^{-\gamma_1 x}\left[A_j + \frac{1}{2K_1}\int_v^x \{f_j(y) - f_j'(y)\} e^{\gamma_1 y} dy \right] \\ & - e^{\gamma_1 x}\left[A_j' + \frac{1}{2K_1}\int_v^x \{f_j(y) - f_j'(y)\} e^{-\gamma_1 y} dy \right], \end{aligned} \tag{30}$$

$$\begin{aligned} b_j = {} & e^{-\gamma_2 x}\left[B_j + \frac{1}{2K_2}\int_v^x \{\bar{f}_j(y) - \frac{1}{n}\sum \bar{f}_k(y)\} e^{\gamma_2 y} dy \right] \\ & - e^{\gamma_2 x}\left[B_j' + \frac{1}{2K_2}\int_v^x \{\bar{f}_j(y) - \frac{1}{n}\sum \bar{f}_k(y)\} e^{-\gamma_2 y} dy \right], \end{aligned} \tag{31}$$

$$\sum B_j = \sum B_j' = 0, \tag{32}$$

$$\begin{aligned} c_j = {} & e^{-\gamma_3 x}\left[C + \frac{1}{4nK_3}\int_v^x \frac{1}{n}\sum \bar{f}_k(y) e^{\gamma_3 y} dy \right] \\ & - e^{\gamma_3 x}\left[C' + \frac{1}{4nK_3}\int_v^x \frac{1}{n}\sum \bar{f}_k(y) e^{-\gamma_3 y} dy \right]. \end{aligned} \tag{33}$$

Here the a component corresponds to transmission over a pair,

the b component to transmission over a phantom circuit, and the c component to ground return transmission. It will be observed that, as regards the impressed field, the a component depends on the difference $f - f'$ of the impressed force at the two sides of the circuit, while the c component depends on the mean impressed electric force averaged over the $2n$ conductors of the system; the b, or phantom component, involves the impressed field in a slightly more complicated way, depending on both the mean impressed force averaged for the two conductors of a pair and also averaged over all the $2n$ conductors of the system.

The extension of the preceding analysis to the general case of n parallel wires, in general dissimilar, is straightforward. The resulting formulas are, however, extremely complicated and for this reason, as well as their small practical utility, they will not be written down.

Formulas (25), $\cdots$ (33) are immediately applicable to the wave antenna and to interference problems in general where the impressed disturbance is supposed to be known. Their application to the problem of crosstalk, which will now be taken up, is not immediate in the same sense because here the primary disturbance which sets up crosstalk is itself a function of the unbalances among the wires composing the system. That is to say, the primary disturbance or impressed field causing the crosstalk is implicitly rather than explicitly given.

II*a*

In discussing the theory of crosstalk a representative problem will be dealt with rather than a formulation of the general problem. The types of problem encountered in practice are extremely varied, depending on whether we have to do with 'side-to-side,' 'side-to-phantom' or 'phantom-to-phantom' crosstalk, etc.; and each problem may call for special treatment. The representative problem, however, besides showing the underlying mathematical theory should serve to indicate the correct procedure in other specific problems.

Let us return to the general system of n parallel wires, dealt with in Section I, and let us suppose that two of them, say wires No. 1 and No. 2, constitute a metallic circuit which, for convenience, we shall suppose would be balanced with respect to ground if the other wires were removed. We now suppose that this metallic circuit is energized by an electromotive force impressed at $x = 0$, which in the absence of the other wires would produce a current I^0 in wire No. 1 and an equal and opposite current $-I^0$ in wire No. 2. Our problem is now to calculate the currents induced in the neighboring wires and the additional

currents induced in wires No. 1 and No. 2 due to the reactions in the system.

It will be observed that, if the system were ideally balanced, no currents would be induced in the neighboring wires and we should simply have the current I^0 in the metallic circuit; in engineering language there would be no crosstalk. This is the ideal to which the correctly designed telephone system approximates by means of 'transpositions.' It is never, of course, completely realized but the approximation, as regards the neighboring metallic circuits, must be extremely close, since the allowable amount of crosstalk is very small.

Let us now return to the original system of equations for n parallel wires discussed in Section I and let us write $I_1 = I^0 + I_1'$, $I_2 = -I^0 + I_2'$ and replace $I_2, I_3 \cdots I_n$ by $I_2', I_3' \cdots I_n'$ respectively, the primes indicating that the currents are 'unbalance' currents. Similarly write $V_1 = V^0 + V_1'$, $V_2 = -V^0 + V_2'$; $Q_1 = Q^0 + Q_1'$, $Q_2 = -Q^0 + Q_2'$; and for the rest of the wires add primes to the symbols for potential and charge. Equations (2) may then be written as

$$\begin{aligned}
(z_{11} + Z_{11})I_1' + \frac{dV_1'}{dx} &= -Z_{12}I_2' - Z_{13}I_3' - Z_{14}I_4' - \cdots, \\
(z_{22} + Z_{22})I_2' + \frac{dV_2'}{dx} &= -Z_{21}I_1' - Z_{23}I_3' - Z_{24}I_4' - \cdots, \\
(z_{jj} + Z_{jj})I_j' + \frac{dV_j'}{dx} &= -(Z_{j1} - Z_{j2})I^0 - Z_{j1}I_1' \\
&\quad - Z_{j2}I_2' - Z_{j3}I_3' - \cdots, \\
&\qquad (j = 3, 4 \cdots n),
\end{aligned} \tag{34}$$

or, denoting the right hand sides of the equations by f_1', f_2', f_j', respectively,

$$\begin{aligned}
(z_{11} + Z_{11})I_1' + \frac{dV_1'}{dx} &= f_1', \\
(z_{22} + Z_{22})I_2' + \frac{dV_2'}{dx} &= f_2', \\
(z_{jj} + Z_{jj})I_j' + \frac{dV_j'}{dx} &= f_j', \\
&\qquad (j = 3, 4 \cdots n).
\end{aligned} \tag{35}$$

This set of equations in the unbalance currents $I_1', \cdots I_n'$ and unbalance potentials $V_1', \cdots V_n'$ admits of immediate interpretation. This is to the effect that the unbalance currents may be regarded as due to an impressed field characterized by an axial electric intensity

$f_j' - dF_j'/dx$ along the jth wire ($j = 1, 2, \cdots n$), and an impressed potential F_j' (line integral of impressed field from jth wire to ground), where

$$\begin{aligned} F_1' &= p_{12}Q_2' + p_{13}Q_3' + p_{14}Q_4' + \cdots, \\ F_2' &= p_{21}Q_1' + p_{23}Q_3' + p_{24}Q_4' + \cdots, \\ F_j' &= (p_{j1} - p_{j2})Q^0 + p_{j1}Q_1' + p_{j2}Q_2' + p_{j3}Q_3' + \cdots, \\ & \qquad (j = 3, 4 \cdots n). \end{aligned} \tag{36}$$

Consequently if f_j' and F_j' were known, equations (25), $\cdots$ (27) would be immediately applicable to the calculation of the unbalance currents. Inspection of equations (34), $\cdots$ (36) shows, however, that while I^0, V^0, Q^0 are supposed known, the expressions for f_j' and F_j' involve the unbalance currents and charges themselves. The solution of the equations calls therefore for a process of successive approximation, now to be discussed. While this method of solution is theoretically sound and applicable in all cases, its success in practical applications depends largely on the fact that the unbalance currents must be extremely small, compared with the primary current I^0, if the crosstalk is to be kept within tolerable limits.

Returning to equations (34), $\cdots$ (36), the first approximate solution is obtained by (1) ignoring the unbalance currents and charges in their effect on the current in the primary wires (No. 1 and No. 2), and (2) replacing $f_3', \cdots f_n'$ and $F_3', \cdots F_n'$ by

$$\begin{aligned} f_j' &= -(Z_{j1} - Z_{j2})I^0, \\ F_j' &= (p_{j1} - p_{j2})Q^0, \\ & \qquad (j = 3, 4 \cdots n). \end{aligned} \tag{37}$$

Consequently in the first approximate solution the primary current, charge and potential are I^0, Q^0, V^0, which are calculable in terms of the impressed e.m.f. and the terminal impedances for the circuit composed of the primary wires (No. 1 and No. 2) by ignoring the reaction of the other wires.[10] The unbalance currents, charges and potentials of the other wires are then calculable on the supposition that those wires are energized by the known impressed field f_j', F_j', as given by (37), which depends only on I^0 and Q^0.

The second approximate solution is obtainable by substituting the first approximate values of I_j' and Q_j' in the right hand side of equations (34) and (36) and then proceeding precisely as in the first approxi-

[10] It should be clearly understood that this particular procedure is not required and is not always followed in practice. For example, it is customary in calculating the crosstalk induced in a metallic or 'side' circuit to take into account, in the first approximate solution, the reaction between the wires making up the disturbed circuit.

mate solution. This process can, theoretically, be repeated indefinitely and successively closer approximations thereby obtained. Practically, however, even in a system of only a few wires, the process rapidly becomes prohibitively laborious and complicated, so that only the first and perhaps the second approximate solutions are practicable. Theoretically, however, the process is straight-forward and the successive approximate solutions form a convergent sequence. Fortunately, in engineering applications the allowable amount of crosstalk is so strictly limited that higher approximations than the second at most are not usually required.

It is an important and valuable property of the solution by successive approximations that the 'datum configuration' is not uniquely fixed, but is at our disposal, within limits. By 'datum configuration' is meant the assumed distribution from which the first approximate solution is derived. In the preceding the datum configuration for the primary wires is taken as

$$\begin{aligned} I_1 &= I^0 = -I_2, \\ Q_1 &= Q^0 = -Q_2, \end{aligned} \tag{38}$$

while in calculating any I_j' $(j = 3, \cdots n)$ it is assumed that the unbalance currents and charges of the other disturbed wires are zero. From the form of the equations this is certainly the natural configuration with which to start. It does not at all follow, however, that this datum configuration results in the optimum first approximate solution.

Another datum configuration which may be taken and which appears to possess practical advantages in certain cases is the following:[11]

$$\begin{aligned} I_1 &= I^0 = -I_2, \\ Q_1 &= Q^0 = -Q_2, \end{aligned} \tag{39}$$

for the primary wires, while in calculating any I_j' $(j = 3, \cdots n)$ it is assumed that the unbalance currents and *potentials* (instead of *charges*) of the other disturbed wires are zero.[12] Higher successive approximate solutions then follow the same scheme of procedure as in the first case.

The foregoing completes the formal analytical theory. The remaining sections of the paper will be devoted to the interpretation of the fundamental mathematical theory and its formulation along more physical and engineering lines, together with applications to representative problems.

[11] This is essentially the basis of the crosstalk formulas developed, in terms of a different mathematical treatment, by Dr. G. A. Campbell of the American Telephone and Telegraph Co., in his early and fundamental work on crosstalk and transposition theory.

[12] See, however, the preceding footnote as to possible modification of the datum configuration.

III

Representation of Impressed Field by Equivalent Electromotive Forces

In the present section we shall start anew with the problem dealt with in Section II, and attack it by a synthetic method, as distinguished from the analytical method employed there. While the results so derived are all deducible from the analytical theory and formulas of Section II, the synthetic or physical mode of attack has important advantages in engineering applications, in giving a physical picture of the phenomena and an intuitive grasp of the problem. In many cases it enables us to deduce results very simply, when the physical picture is well in mind, whereas the purely analytical solution may be laborious.

The essence of this synthetic method consists in replacing the known electric field impressed on the physical system by a set of equivalent electromotive forces; the current at any point in the system can then be calculated when the transfer admittances between that point and the points where the electromotive forces are situated are known or calculable (as is often the case in practical applications). For, considering any linear system containing any number m of electromotive forces inserted at any points $1, \cdots m$, it is known, from the principle of superposition, that the current I_h at any point h is a linear function of all the electromotive forces, that is,

$$I_h = \sum_{k=1}^{m} A_{hk} E_k. \tag{40}$$

The coefficient A_{hk} is called the 'transfer admittance' from k to h, because A_{hk} is equal to the ratio of I_h to E_k when all of the electromotive forces except E_k are zero. If the system contains any unilateral element (such as a one-way amplifier, for instance), A_{hk} is not in general equal to A_{kh}.

Fundamental Set of Equivalent Electromotive Forces: General Formulation

Consider any system of parallel wires situated in an arbitrary impressed field, with any number of localized admittance bridges between wires or between wires and ground. (Evidently, distributed bridged admittance can be analyzed into infinitesimal elements, and these can be regarded as localized.) The cross-sectional dimensions of the wires are assumed to be small enough so that the axial (longitudinal) impressed electric force is sensibly constant over each cross-section.

The electric constituent of the impressed field is assumed to be specified at every point along the wires by the impressed axial electric force and the impressed potential. At any point x in any wire, h, the impressed axial electric force will be denoted by $f_h(x)$ and the impressed potential by $F_h(x)$; these are to be regarded as arbitrary functions of x, and may even be discontinuous.

The following set of electromotive forces is easily seen to be equivalent to the above-specified arbitrary impressed field, in the sense of producing the same currents and charges. This set will be termed the 'fundamental' set of equivalent electromotive forces; for, from the physical viewpoint of this paper, it is in fact the fundamental set.[13]

(A) In each wire a distributed axial electromotive force whose value, per unit length, at each point is equal to the impressed axial electric force there; thus, at any point x in wire h, an electromotive force $f_h(x)dx$ in the differential length dx.

(B) At each point where the impressed potential is discontinuous, an axial electromotive force equal to the decrement in the impressed potential there; thus, at any point of discontinuity $x = u$ in any wire h, an electromotive force equal to $-\Delta F_h(u) = F_h(u-) - F_h(u+)$.

(C) In each bridge an electromotive force equal to the impressed voltage in that bridge; thus, in a bridge at any point $x = b$, from wire h to any other wire k (or to ground), an electromotive force equal to $F_h(b) - F_k(b)$.

(D) In case a point $x = b$ where a bridge is situated coincides with a point $x = u$ where the potential $F_h(x)$ impressed on wire h is discontinuous, the corresponding electromotive forces are as follows: Axial electromotive forces equal to $F_h(b-)$ and $-F_h(b+)$ at points $b-$ and $b+$ respectively in wire h; no electromotive force in the bridge itself, which is connected to the point b situated between $b-$ and $b+$ in wire h.[13a]

[13] For a one-wire line and for a balanced two-wire line, five other sets of equivalent electromotive forces are formulated in a later subsection.

[13a] By supposing points b and u to be not quite coincident, say $b = u-$ or $b = u+$, item (D) can be derived by first applying items (B) and (C) and then applying the 'branch-point theorem' formulated in the second paragraph following equation (75).

A further application of the 'branch-point theorem' yields for item (D) the following alternative set of electromotive forces: Axial electromotive forces each equal to $[F_h(b-) - F_h(b+)]/2$ at $b-$ and $b+$ in wire h; an electromotive force equal to $[F_h(b-) + F_h(b+)]/2$ in the bridge at b. Clearly, this set reduces to (C) when $F_h(x)$ is continuous at $x = b$, and it reduces to (B) when there is no bridge at $x = u$.

A physical verification of the correctness of the foregoing set of equivalent electromotive forces can be obtained by starting with the given system, situated in the specified arbitrary impressed field (but not otherwise energized), and then inserting in the wires and bridges a set of electromotive forces, which will be termed the 'annulling electromotive forces,' such as to annul all currents in the wires and bridges. The resultant axial electric force in the wires will then be zero, and furthermore the wires will be uncharged; hence the inserted axial electric force must be equal and opposite to the impressed axial electric force. Since the wires are uncharged their potentials will be those of the impressed field; hence, since no current flows in the bridges, the electromotive forces inserted in the bridges must be equal and opposite to the voltages of the impressed field at the bridges. Evidently the negatives of the annulling electromotive forces constitute a set of electromotive forces equivalent to the impressed field; for, insertion of the negatives of the annulling electromotive forces restores the system to its original state, in which it is acted on by only the original impressed field.

From the nature of this demonstration it is seen that the 'fundamental set' of equivalent electromotive forces is not limited to a system of parallel horizontal wires. In the general case, where the wires are neither straight nor parallel nor horizontal, x (and hence u and b) is to be interpreted as being the 'intrinsic coordinate' of a point in the particular wire contemplated, that is, the distance measured along that wire from any arbitrary fixed point therein. Thus, for wires h and k respectively, x becomes x_h and x_k, which in general are independent of each other.

For the case of a one-wire line, an analytical derivation of this set of equivalent electromotive forces is given in a later subsection by interpretation of the fundamental differential equations of the line.

A One-Wire Line in an Arbitrary Impressed Field

As indicated by Fig. 1, the line extends from $x = 0$ to $x = s$, and is terminated in impedances Z_0 and Z_s respectively. γ denotes the propagation constant per unit length, and K the characteristic impedance.[14] The direct leakage admittance from the wire to ground, per unit length, is denoted by Y'; this is the generalization of a mere leakage conductance.[31]

The impressed field is specified by the functions $f(x)$ and $F(x)$; $f(x)$ denoting the impressed axial electric force and $F(x)$ the impressed

[14] Given by formulas (12) and (11) of Appendix I.

potential, at any point x of the wire. For generality, $F(x)$ is assumed to be discontinuous at any point $x = u$ by the increment

$$\Delta F(u) = F(u +) - F(u -).$$

The problem is to calculate the current $I(x)$ produced at any point x by the impressed field.

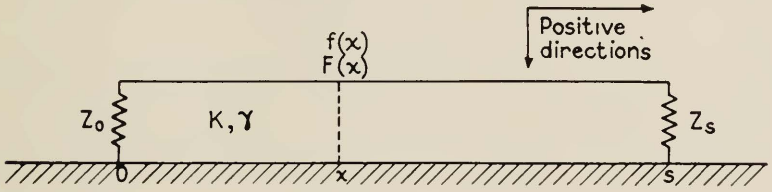


Fig. 1.

Case 1: General Case

The current $I(x)$ is the sum of two constituents: $j(x)$ due to the impressed axial electric force, and $J(x)$ due to the impressed potential. Formulas for these constituents will now be written down by aid of the fundamental set of equivalent electromotive forces formulated in the preceding subsection. Thus [15]

$$j(x) = \int_0^s A(x, y)f(y)dy, \tag{41}$$

$A(x, y)$ denoting the transfer admittance between points x and y. $J(x)$ itself consists of four constituents: $J_0(x)$ and $J_s(x)$, originating in the terminal impedances Z_0 and Z_s respectively, $J_{0s}(x)$ originating in the direct leakage admittance of the whole line 0-s, and $J_u(x)$ originating at the point $x = u$ where $F(x)$ is discontinuous. Thus

$$J_0(x) = - A(x, 0)F(0), \tag{42}$$

$$J_s(x) = A(x, s)F(s), \tag{43}$$

$$J_{0s}(x) = \int_0^s Y'F(y)B(x, y)dy, \tag{44}$$

$$J_u(x) = - A(x, u)\Delta F(u), \tag{45}$$

$B(x, y)$ denoting a current transfer factor representing that fraction

[15] With regard to the analytical evaluation of the integrals, attention should perhaps be called to the fact that the integrand may be discontinuous or may change its functional form at one or more points within the range of integration; whence the integral must be broken up into a sum of integrals.

of the current contribution originating in the direct leakage admittance element $Y'dy$ at y, which reaches point x.

Case 2: Terminal Impedances Equal to Characteristic Impedances
Here we have

$$Z_0 = Z_s = K, \tag{46}$$

$$A(x, y) = \frac{1}{2K} e^{-\gamma|x-y|}, \tag{47}$$

$$B(x, y) = \mp \tfrac{1}{2} e^{-\gamma|x-y|}, \qquad y \lessgtr x, \tag{48}$$

whence

$$j(x) = \frac{1}{2K} \int_0^s e^{-\gamma|x-y} f(y)dy \tag{49}$$

$$= \frac{1}{2K} \int_0^x e^{-\gamma(x-y)} f(y)dy + \frac{1}{2K} \int_x^s e^{-\gamma(y-x)} f(y)dy, \tag{50}$$

$$J_0(x) = -\frac{F(0)}{2K} e^{-\gamma x}, \tag{51}$$

$$J_s(x) = \frac{F(s)}{2K} e^{-\gamma(s-x)}, \tag{52}$$

$$J_{0s}(x) = -\frac{Y'}{2} \int_0^x F(y) e^{-\gamma(x-y)} dy + \frac{Y'}{2} \int_x^s F(y) e^{-\gamma(y-x)} dy, \tag{53}$$

$$J_u(x) = -\frac{\Delta F(u)}{2K} e^{-\gamma|x-u}. \tag{54}$$

A Balanced Two-Wire Line in an Arbitrary Impressed Field

Because the metallic circuit here contemplated is balanced, its treatment can be made formally the same as the treatment of the one-wire line in the preceding subsection.

This fact is immediately evident in two special cases of the impressed electric field (potential and axial electric force): (1) the case where the impressed electric field has equal values at the two wires, and (2) the case where it has equal but opposite values at the two wires.

The general case where the impressed field at the two wires has any values can be treated as a superposition of the two special cases just

mentioned, by the simple device of resolving the impressed field at each wire into two constituents one of which has equal values at the two wires while the other has equal but opposite values at the two wires. This resolution is always possible, for it is merely in accordance with the following pair of algebraic identities:

$$\eta_1 = \tfrac{1}{2}(\eta_1 + \eta_2) + \tfrac{1}{2}(\eta_1 - \eta_2), \tag{55}$$

$$\eta_2 = \tfrac{1}{2}(\eta_1 + \eta_2) - \tfrac{1}{2}(\eta_1 - \eta_2). \tag{56}$$

Although η_1 and η_2 may in general denote any two quantities whatever, in the present application they refer to the impressed electric field at the two wires No. 1 and No. 2 of the contemplated two-wire line. It is convenient to introduce the symbols η_c and η_a defined by the equations

$$\eta_c = \tfrac{1}{2}(\eta_1 + \eta_2), \tag{57}$$

$$\eta_a = \eta_1 - \eta_2, \tag{58}$$

so that the resolutions (55) and (56) of η_1 and η_2 can be written in the more compact forms

$$\eta_1 = \eta_c + \tfrac{1}{2}\eta_a, \tag{59}$$

$$\eta_2 = \eta_c - \tfrac{1}{2}\eta_a. \tag{60}$$

η_c and η_a will be termed respectively the mode-c and mode-a constituents of the impressed field, because they give rise to mode-c and mode-a effects respectively; mode-c effects being defined as those which are equal in the two wires, mode-a effects as those which are equal but opposite in the two wires—as discussed in connection with equations (28). From (57) and (58) respectively it will be noted that the mode-c effects and the mode-a effects depend respectively on the average and on the difference of the impressed fields at the two wires.

As in treating the one-wire line (in the preceding subsection), so also in treating the balanced two-wire line (in the present subsection) it is usually advantageous to deal separately with the axial electric force and the potential of the impressed field. Furthermore, in the case of the two-wire line each of these constituents of the impressed field is to be resolved into two modes, c and a, in the manner represented by equations (59) and (60) together with (57) and (58).

Owing to the balance (bilateral symmetry) of the assumed two-wire line, the mode-c constituent η_c of the impressed field will produce only mode-c effects, and the mode-a constituent only mode-a effects. Thus, η_c will produce equal currents I_c and I_c in the two wires, while η_a will produce equal but opposite currents I_a and $-I_a$ in the two

wires. The total mode-c current $2I_c$ along the two wires in parallel is calculable from η_c through the mode-c parameters (γ_c, K_c, and terminal impedances) of the system; while the mode-a or loop current (I_a and $-I_a$ in the two wires respectively) is calculable from η_a through the mode-a parameters (γ_a, K_a, and terminal impedances). The connection of each current constituent with the corresponding field constituent, through the corresponding parameters, is formally the same as for the one-wire line (treated in the preceding subsection).

Finally, it may be remarked that the assumption of balance (bilateral symmetry) for the two-wire line is essential to the above simplicity; for otherwise each mode of the impressed field would produce components of both modes of effects, instead of only the appropriate single mode of effects.

Illustrative Special Case

For illustration it will suffice to choose the simple case of a balanced two-wire line terminated at each end in its mode-c and mode-a characteristic impedances simultaneously. That is, the line consisting of the two wires in parallel, with ground return, is terminated at each end in the mode-c characteristic impedance K_c; while the loop circuit is terminated at each end in the mode-a characteristic impedance K_a. (Evidently these two modes of terminating can be simultaneously accomplished by means either of a balanced T-network or of a balanced Π-network at each end.)

Let $f_1(x)$ and $f_2(x)$ denote the axial impressed electric forces at any point x in wires No. 1 and No. 2 respectively; and let them be resolved into mode-c and mode-a constituents $f_c(x)$ and $f_a(x)$, respectively, such that

$$f_c(x) = \tfrac{1}{2}[f_1(x) + f_2(x)], \tag{61}$$

$$f_a(x) = f_1(x) - f_2(x), \tag{62}$$

in accordance with equations (57) and (58). Similarly, let $F_1(x)$ and $F_2(x)$ denote the impressed potentials at point x; and let them be resolved likewise, so that

$$F_c(x) = \tfrac{1}{2}[F_1(x) + F_2(x)], \tag{63}$$

$$F_a(x) = F_1(x) - F_2(x). \tag{64}$$

Thus, formulas (49), $\cdots$ (54) of the one-wire line are seen to be formally applicable to the balanced two-wire line, for calculating separately the two modes of currents. This is with the understanding that they give the sum of the mode-c currents in the two wires, hence twice the mode-c current in each wire; and that they give the loop

current (the current circulating in the metallic loop), which is equal to the mode-a current in one of the wires and hence to the negative of the mode-a current in the other wire.

Six Different Sets of Equivalent Electromotive Forces for a One-wire[16] *Line in an Arbitrary Impressed Field*

The physical system here contemplated (Fig. 2 below) is a one-wire transmission line consisting of a uniform horizontal straight wire situated in an arbitrary impressed field and terminated in any arbitrary impedances to ground. The wire extends from $x = 0$ to $x = s$; the arbitrary terminal impedances[17] are denoted by Z_0 and Z_s. The arbitrary impressed field is specified, at each point of the wire, by the impressed axial electric force $f(x)$ and the impressed potential $F(x)$, as previously.

Six different sets of 'equivalent electromotive forces' are formulated in the early part of this subsection; while their derivations are briefly outlined toward the latter part. Set 1 will be recognized as a particular case of the 'fundamental' set already formulated in the early part of Section III. The five remaining sets are derived from Set 1. In the actual formulations of these various sets of equivalent electromotive forces, the impressed potential $F(x)$ is assumed to be a continuous function of x; the extension to the case where $F(x)$ is discontinuous is a simple matter and is formulated in connection with equations (65) and (66).

In the following diagrams (Figs. 2, $\cdots$ 11) it is found convenient to represent any localized electromotive force by the conventional battery-symbol. This symbol is intrinsically directional; the longer of the two plates is to be regarded as at the higher potential, so that there is an internal rise of potential in passing through the symbol from the shorter to the longer plate.

In some of the figures the actual line is represented as replaced by the corresponding artificial line composed of differential elements, each of length dx. (For clearness, the line is represented as composed of only a small number of such elements.)

The letters Z, Y, Y', Y^0 denote certain line parameters per unit length, as follows: Z and Y respectively denote the 'complete series impedance' and the 'complete shunt admittance' or, briefly, the 'series impedance' and the 'shunt admittance.' These may be regarded as defined by the equations

$$Z = \gamma K, \qquad Y = \gamma / K,$$

[16] The case of a balanced two-wire line is outlined in the next subsection.
[17] See also the remarks under the subheading following shortly after equation (67).

γ denoting the propagation constant of the line per unit length, and K the characteristic impedance. Or they may be regarded as defined by the differential equations

$$-dV/dx = ZI, \qquad -dI/dx = YV,$$

characterizing the line when there is no impressed field present. Y' denotes the 'direct leakage admittance' and Y^0 the 'basic shunt admittance,' the latter defined as being the value of Y when $Y' = 0$, whence $Y = Y^0 + Y'$. On referring to equations (12), (11), (8), (1) of Appendix I, and also to equations (2) and (7) in Section I, it is seen that [18]

$$Z = z + i\omega L, \qquad Y = G + i\omega C,$$
$$G = G^0 + Y', \qquad Y^0 = G^0 + i\omega C.$$

The various sets of equivalent electromotive forces remain valid even when the line parameters are functions $Z(x)$, $Y(x)$, etc., of position x along the system. For the 'fundamental' set this fact can be readily seen by reference to the formulation and verification of the fundamental set, in the early part of Section III.

As indicated by the arrows, the positive axial (longitudinal) direction is the direction of increasing x, and the positive vertical direction is downward.

Six Different Sets of Equivalent Electromotive Forces

Set 1 (Fig. 4)

(A) In the wire, a distributed electromotive force, $f(x)dx$ in each differential length dx.

(B) In the distributed direct leakage admittance, a distributed electromotive force, $F(x)$ in each differential element $Y' \cdot dx$ of direct leakage admittance.

(C) In the termial impedances Z_0 and Z_s, electromotive forces $F(0)$ and $F(s)$ respectively.

From the physical viewpoint of the present paper, Set 1 is the fundamental set of equivalent electromotive forces.

This set is particularly simple when there is no direct leakage admittance ($Y' = 0$), for then it reduces to merely the axial constituents (A) and the terminal constituents (C).

[18] Thus Z, unsubscripted, includes the internal impedance $z = z_w + z_g$ of the circuit, and hence is to be sharply distinguished from the double-subscripted Z occurring frequently in this paper; for, as remarked in connection with equation (2), Z_{jj} does not include the internal impedance z_{jj} of wire j, whence it is seen that $Z = Z_{jj} + z_{jj}$ for wire j.

Set 4 (Fig. 8)

(*A*) In the wire, a distributed electromotive force, $[f(x) + (Y'/Y) \times dF(x)/dx]dx$ in each differential length dx.

(*B*) In the terminal impedances Z_0 and Z_s, electromotive forces $(1 - Y'/Y)F(0)$ and $(1 - Y'/Y)F(s)$ respectively.

Set 2 is distinguished by containing no electromotive forces in the shunt admittance even when the direct leakage admittance Y' is not negligible. Thus Set 2 with the direct leakage admittance not negligible is *formally* as simple as Set 1 with the direct leakage admittance negligible. However, in Set 2 the element of axial electromotive force is a much more complicated function than in Set 1.

Set 3 (Fig. 9)

(*A*) In the wire, a distributed electromotive force, $[f(x) + dF(x)/dx]dx$ in each differential length dx.

(*B*) In the distributed basic shunt admittance, a distributed electromotive force, $-F(x)$ in each differential element Y^0dx of the distributed basic shunt admittance.

In Set 3 it should be noted that the electromotive force $-F(x)$ is in the basic shunt admittance element Y^0dx, not in the complete shunt admittance element Ydx.

It will be observed that this set contains no electromotive forces in the terminal impedances.

When the ground is a perfect conductor, so that $f_g(x) = 0$, the differential element of axial electromotive force in this set reduces to merely $-[d\Phi(x)/dt]dx$, as is shown by equation (2) of Appendix I, $\Phi(x)$ denoting the impressed magnetic flux.

Set 4 (Fig. 8)

(*A*) In the wire, a distributed electromotive force, $[f(x) + (1 + Y'/Y)dF(x)/dx]dx$ in each differential length dx.

(*B*) In the distributed complete shunt admittance, a distributed electromotive force, $-F(x)$ in each differential element Ydx of the complete shunt admittance.

(*C*) In the terminal impedances Z_0 and Z_s, electromotive forces, $-(Y'/Y)F(0)$ and $-(Y'/Y)F(s)$ respectively.

In Set 4 it should be noted that the electromotive force $-F(x)$ is in the complete shunt admittance element Ydx.

The differential element of axial electromotive force in this set does not reduce to $-[d\Phi(x)/dt]dx$ when the ground is a perfect conductor ($f_g(x) = 0$) unless also the direct leakage admittance is zero ($Y' = 0$).

Set 5 (Fig. 6)

(*A*) In the wire, a distributed electromotive force, $f(x)dx$ in each differential length dx.

(*B*) In the distributed complete shunt admittance, a distributed electromotive force, $(Y'/Y)F(x)$ in each differential element Ydx of the complete shunt admittance.

(*C*) In the terminal impedances Z_0 and Z_s, electromotive forces $F(0)$ and $F(s)$ respectively.

In Set 5 it should be noted that the electromotive force $(Y'/Y)F(x)$ is in the complete shunt admittance element Ydx.

It will be observed that Set 5 (Fig. 6) is the same as Set 1 (Fig. 4) as regards the axial and the terminal electromotive forces.

Set 6 (Fig. 11)

(*A*) At any arbitrary fixed point $x = a$ in the wire, an axial electromotive force G_a,

$$G_a = \int_0^s \left[f(x) + \frac{dF(x)}{dx} \right] dx.$$

(*B*) In each differential element Ydx of the distributed complete shunt admittance, a distributed electromotive force E_x,

$$E_x = \int_0^x \left[f(x) + \frac{Y'}{Y} \frac{dF(x)}{dx} \right] dx, \qquad x < a,$$

$$E_x = -\int_x^s \left[f(x) + \frac{Y'}{Y} \frac{dF(x)}{dx} \right] dx, \qquad x > a.$$

(*C*) In the terminal impedances Z_0 and Z_s, electromotive forces $(1 - Y'/Y)F(0)$ and $(1 - Y'/Y)F(s)$ respectively.

Set 6 is perhaps mainly of academic interest.

Two limiting cases of Set 6 may be noted, corresponding to $a = 0$ and $a = s$ respectively, each characterized by containing no *internal* axial electromotive force: for when $a = 0$ the axial electromotive force G_a can be combined with the terminal electromotive force $(1 - Y'/Y)F(0)$ in the terminal impedance Z_0, and when $a = s$ it can be combined with $(1 - Y'/Y)F(s)$ in Z_s.

Extension to the Case where the Impressed Potential is Discontinuous

In the foregoing formulations of Sets 1, $\cdots$ 6 of equivalent electromotive forces it has been assumed that the impressed potential $F(x)$ is a continuous function of x throughout the length of the line.

Suppose now, for greater generality, that the impressed potential $F(x)$ is discontinuous at any point $x = u$ by the increment

$$\Delta F(u) = F(u +) - F(u -).$$

Then (as shown in the next paragraph), for the particular differential element which contains the point u, the quantities $f(x)dx$ and $[dF(x)/dx]dx$ must be replaced by $-\Delta F(u)$ and $\Delta F(u)$ respectively; that is,

$$f(u)du = -\Delta F(u), \tag{65}$$

$$\frac{dF(u)}{du}du = \Delta F(u). \tag{66}$$

Equations (65) and (66) can be obtained, by a limiting process, from equation (2) of Appendix I, which for the present purpose will be written in the form

$$f(x)dx = -\frac{dF(x)}{dx}dx - \frac{d\Phi(x)}{dt}dx + f_g(x)dx. \tag{67}$$

It will be recalled, from Appendix I, that this equation was derived by applying the second curl law to a differential rectangle extending from x to $x + dx$; but x may equally well be a point within the differential segment dx, and for the present purpose it will be so regarded. The limiting process now consists in letting $dF(x)/dx$ approach infinity while dx approaches zero, but in such a way that the product $[dF(x)/dx]dx$ approaches a preassigned finite value, denoted by $\Delta F(x)$. Then, in the limit, the last two terms on the right side of (67) vanish so that (67) reduces to

$$f(x)dx = -\Delta F(x).$$

Thus we obtain equations (65) and (66), where u denotes, for distinction, the particular value of x at which $F(x)$ is discontinuous.

Remarks on the Terminal Impedances and the Equivalent Electromotive Forces in Them

The arbitrary terminal impedances Z_0 and Z_s (Fig. 2) need not actually be localized. They may, for instance, be the impedances offered by other lines to which the given line 0-s may be connected, and these other lines may themselves be situated in arbitrary impressed fields; in particular, 0-s may be merely a segment, of any length, forming part of a given line in an arbitrary impressed field.

From this broad view, any 'equivalent electromotive forces' situated in the terminal impedances Z_0 and Z_s may advantageously be regarded as being situated in the ends of the line itself (that is, in the end-points $x = 0$ and $x = s$), these electromotive forces being then regarded as pertaining primarily to the line-segment 0-s rather than to the terminal

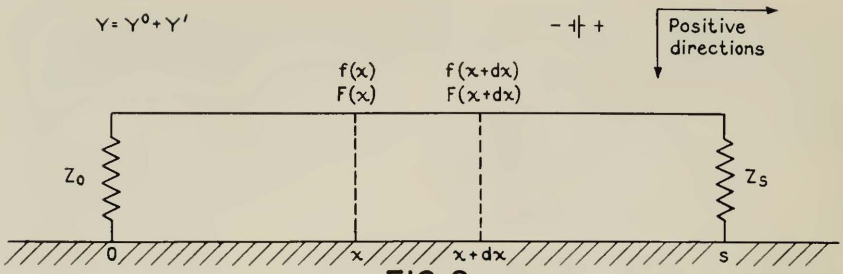


FIG. 2

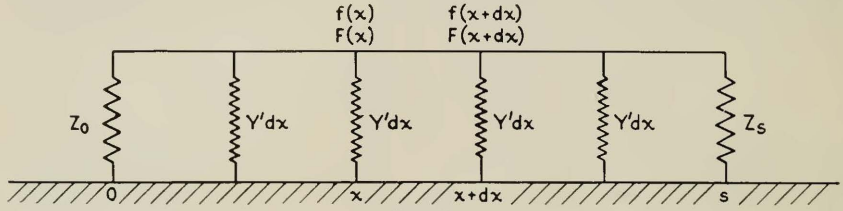


FIG. 3

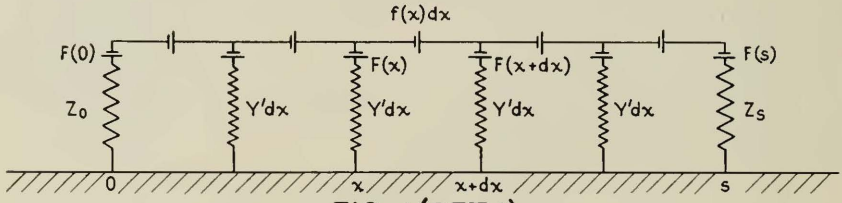


FIG. 4 (SET 1)

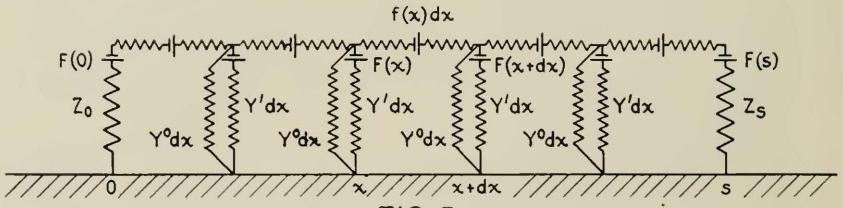


FIG. 5

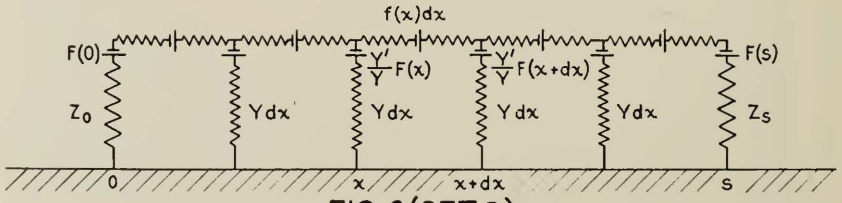


FIG. 6 (SET 5)

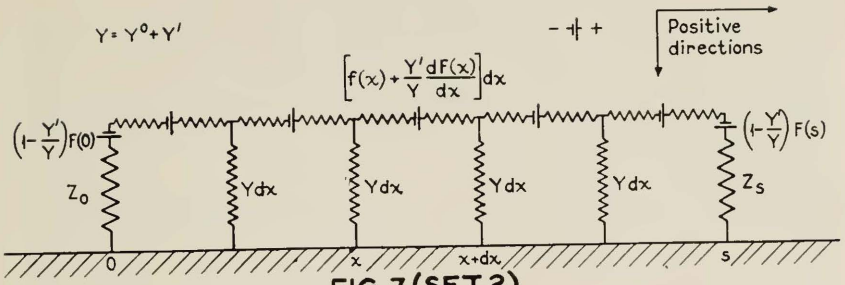


FIG. 7 (SET 2)

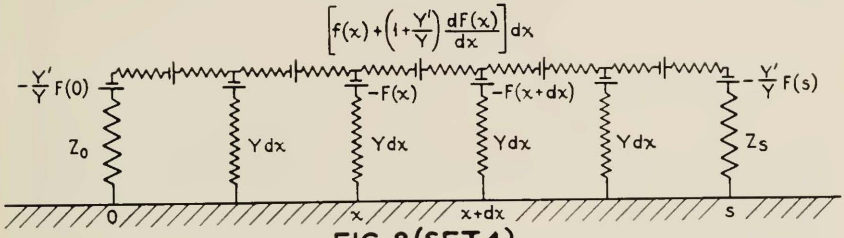


FIG. 8 (SET 4)

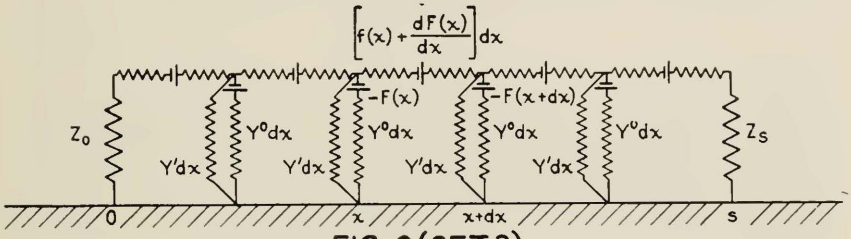


FIG. 9 (SET 3)

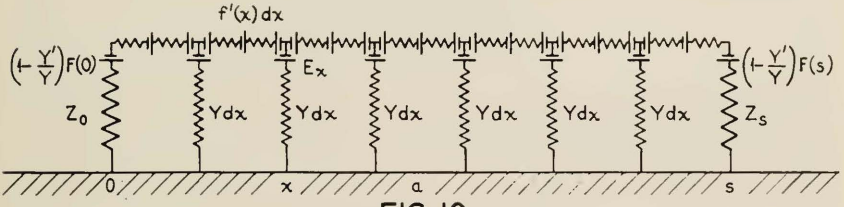


FIG. 10

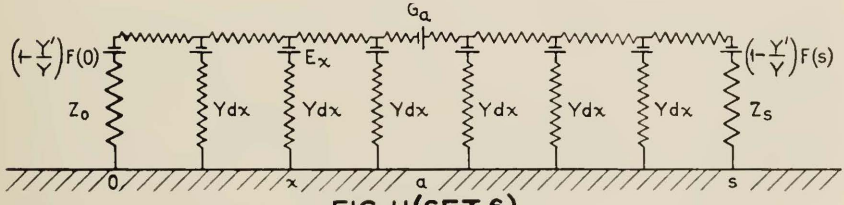


FIG. 11 (SET 6)

impedances. Thus, for instance, in the formulations of Set 1 and Set 5, item (C) would read: '(C) In the ends $x = 0$ and $x = s$ of the line, axial electromotive forces $-F(0)$ and $F(s)$ respectively.' (Observe, here, the negative sign before $F(0)$, in contrast to the positive sign in the original formulation.)

In this way it is readily seen that at a point $x = u$ where the impressed potential $F(x)$ is discontinuous, the equivalent electromotive force is an axial electromotive force equal to the decrement of the impressed potential, that is, equal to $F(u-) - F(u+)$; this agrees with equation (65), and with item (B) in the fundamental set of equivalent electromotive forces formulated in the early part of Section III.

Derivations of Set 1

A synthetic derivation of Set 1 has already been furnished in the early part of Section III. An analytical derivation will now be outlined; it is based on an interpretation of equations (68), (71), (73) below; these equations, in turn, are based on certain equations of Appendix I, as follows:

Combining equations (1) and (2) of Appendix I gives

$$zI + \frac{dV'}{dx} + \frac{d\phi'}{dt} = f, \tag{68}$$

where f denotes f_w; and V' is that part of the potential of the wire due to its charges (and the corresponding opposite charges on the surface of the ground), while ϕ' is that part of the magnetic flux due to the current in the wire (and the corresponding return current in the ground); that is,

$$V' = V - F = Q/C, \tag{69}$$

$$\phi' = \phi - \Phi = LI, \tag{70}$$

so that V' and ϕ' do not include the impressed potential and impressed magnetic flux F and Φ respectively.

By (5) and (7) of Appendix I the equation of current continuity can be written

$$-\frac{dI}{dx} = \frac{dQ}{dt} + \frac{G}{C}Q + Y'F. \tag{71}$$

The actual potential V of the line is of course the resultant of V' and F; that is,

$$V \equiv V(x) = V'(x) + F(x), \tag{72}$$

whence, in particular, at the ends $x = 0$ and $x = s$,

$$\begin{aligned} V(0) &= V'(0) + F(0), \\ V(s) &= V'(s) + F(s). \end{aligned} \tag{73}$$

Returning, now, to a consideration of equations (68), (71), (73), it is seen that they are identically the same as the equations for the same line without any impressed field but containing the set of electromotive forces formulated above under the heading 'Set 1 (Fig. 4)'; for an interpretation of equations (68), (71), (73) yields respectively (*A*), (*B*), (*C*) of Set 1.

It may be noted that equations (68) and (71) can be written in the following more compact forms:

$$-dV'/dx = ZI - f, \tag{74}$$

$$-dI/dx = YV' + Y'F, \tag{75}$$

whose interpretation yields immediately items (*A*) and (*B*) of Set 1.

Outline of Derivations of Sets 2, 3, 4, 5, 6

Synthetic derivations of Sets 2, 3, 4, 5, 6 from Set 1 will now be briefly outlined by aid of the diagrams in Figs. 2, ··· 11. The physical systems represented by these diagrams are all equivalent in the sense that the currents at corresponding points in all of them are equal.

In the derivation-work extensive use is made of an artifice which, for convenience, will be formulated in what may be termed the 'branch-point theorem,' as follows: *In any network of any number of branches the currents will not be affected by inserting at any branch-point a set of equal electromotive forces, one in each branch, directed either all toward or all from the branch-point.*

Fig. 2 represents the given one-wire line in an arbitrary impressed field, as already specified. For generality the line is assumed to have uniformly distributed leakage admittance of amount Y' per unit length.

Fig. 3 is derived from Fig. 2 by lumping the distributed direct leakage admittance into localized admittances each of amount $Y' \cdot dx$ at intervals of length dx.

Fig. 4 is derived from Fig. 3 by replacing the arbitrary impressed field by Set 1 of equivalent electromotive forces.

Fig. 5 is derived from Fig. 4 by replacing the line, exclusive of the direct leakage admittance Y', by its equivalent artificial line having 'complete series impedance' Z and 'basic shunt admittance' Y^0 per unit length. This replacement of the actual line by the corresponding artificial line is permissible now that the impressed field has been replaced by a set of equivalent electromotive forces (Fig. 4).

Fig. 6 is derived from Fig. 5 by replacing the compound shunt

element, consisting of Y^0dx in parallel with $Y'dx$ containing the electromotive force $F(x)$, by the equivalent simple shunt element Ydx containing the electromotive force $(Y'/Y)F(x)$.

Figs. 7, 8, 9 are derived from Figs. 6, 7, 5 respectively by applying the 'branch point theorem.'

Fig. 11 is derived from Fig. 7 by applying the 'branch point theorem' in the manner indicated by Fig. 10, where $f'(x)dx$ denotes, for brevity, the original axial equivalent electromotive force situated between x and $x + dx$ of Set 2 as indicated by Fig. 7, so that

$$f'(x) \equiv f(x) + \frac{Y'}{Y}\frac{dF(x)}{dx}, \tag{76}$$

and a is the coordinate of the contemplated arbitrary point. The E's, of which E_x is typical, are sets of electromotive forces inserted at the branch-points. At first these electromotive forces are arbitrary, except that each set of three accords with the branch-point theorem, so as not to alter the original currents in the system. Next, starting at the ends, it is found that these electromotive forces can be so determined as to annul the original axial electromotive forces $f'(x)dx$ in all of the differential elements dx except in the one containing the point a; the requisite value of E_x and the resulting value of G_a are found to be as formulated in Set 6.

Finally it may be remarked that each of the Sets 2, 3, 4, 5, 6 can be verified against Set 1 by formulating the total current produced at any point x by each of the Sets 2, 3, 4, 5, 6 and then comparing the resulting formula with the sum of formulas (41), (42), (43), (44). Evidently it suffices to do this for the relatively simple case where the terminal impedances are equal to the characteristic impedance of the line; for this case, formulas (41), (42), (43), (44) reduce to (50), (51), (52), (53) respectively.

Sets of Equivalent Electromotive Forces for a Balanced Two-Wire Line in an Arbitrary Impressed Field

The foregoing six sets of equivalent electromotive forces for a one-wire line can be readily extended to a two-wire line after resolving the impressed field into mode-a and mode-c constituents, which are then dealt with separately. For Set 1 this procedure has been fully outlined above in the subsection entitled 'A Balanced Two-Wire Line in an Arbitrary Impressed Field,' and it has found a natural application in the 'Crosstalk Problem' treated below in Section IV.

It is clear that all of the sets of equivalent electromotive forces are immediately applicable to dealing with the mode-c constituent of the

impressed field, since this constituent acts on the circuit consisting of the two wires in parallel with each other, with ground return, which is formally the same as a one-wire line with ground return.

All of the sets of equivalent electromotive forces become applicable to dealing with the mode-a constituent of the impressed field by an appropriate interpretation of the diagrams (Fig. 2, $\cdots$ 11), namely, the following interpretation:

1. In each diagram regard the wire-symbol as representing the outgoing wire of the actual two-wire line, and regard the ground-symbol as representing not the ground but the return wire of the two-wire line. (The presence of the earth is then to be regarded as implied, its effects appearing implicitly in the values of the line parameters.)

2. Hence regard Z_0 and Z_s as denoting the mode-a terminal impedances functioning as though connected directly across the two-wire line at its ends $x = 0$ and $x = s$ respectively.

3. Regard Y', Y^0, Y, Z as denoting the mode-a line constants (including implicitly the effects of the earth).

4. Regard $f(x)$, $F(x)$, and $\Phi(x)$ as denoting the mode-a constituents of the impressed field—that is, as denoting the difference of the actual values impressed at the two wires. (In order to maintain the balanced condition of the two-wire line, $f(x)$ is to be regarded as constituted of $f(x)/2$ in the outgoing wire and $-f(x)/2$ in the return wire; and similarly for $F(x)$ and $\Phi(x)$.)

The Electric Field Due to a System of n Parallel Wires in an Arbitrary Impressed Field

Thus far in the present section of this paper the field impressed on the given physical system has been supposed known and the problem has been to calculate the resulting currents. Actually, however, the impressed field is not usually known but has to be calculated—from a knowledge of the currents and charges producing it.

The present subsection deals with the problem of calculating the electric field impressed on a secondary system consisting of a single horizontal wire j by a primary system π consisting of n wires which are parallel to each other and to j. For generality, the primary and secondary systems are supposed to be in an arbitrary impressed field.[19]

Consider at first any parallel geometrical line i, not necessarily in any of the wires; and let $V_i = V_i(x)$ and $E_i = E_i(x)$ denote the

[19] Of course the field produced by any given system is directly due only to the currents and charges of the system, and does not depend directly on any field that may be impressed on the system; but, assuming the system to be energized only by the impressed field, the currents—and thence the charges—are directly due to the impressed field and can (theoretically, at least) be expressed in terms of it.

potential and the axial electric force at any point x in i. Then V_i is analyzable into three parts ($V_{i\pi}$, V_{ij}, F_i) and E_i into three parts ($E_{i\pi}$, E_{ij}, f_i) due respectively to the primary system π, to the secondary system j, and to the arbitrary impressed field; that is,[20]

$$V_i = V_{i\pi} + V_{ij} + F_i, \tag{77}$$

$$E_i = E_{i\pi} + E_{ij} + f_i. \tag{78}$$

In particular, at the secondary wire j the potential V_j and the axial electric force E_j are analyzable in accordance with the equations

$$V_j = V_{j\pi} + V_{jj} + F_j, \tag{79}$$

$$E_j = E_{j\pi} + E_{jj} + f_j. \tag{80}$$

In the present subsection, the problem to be dealt with is the calculation of $V_{j\pi}$ and $E_{j\pi}$, namely the potential and the axial electric force at any point x in the secondary j due directly to the currents and charges of the primary system π.

The n wires of the primary system π will be numbered 1, 2, 3, $\cdots$ n. The letters h, k, r will be employed generically: each may denote any one of the designation numbers 1, $\cdots$ n—as h in equation (81); or each may run through the whole set 1, $\cdots$ n—as in equation (91). (It is hardly necessary to remark that j is *not* a member of the set 1, $\cdots$ n, in the notation of this Section (III), where j *always* designates the secondary wire.)

The current at any point x in any wire h of the primary will be denoted by $I_h = I_h(x)$; and the charge on wire h, per unit length, by $Q_h = Q_h(x)$.

The potential V_h and the axial electric force E_h at any primary wire h are analyzable in the manner expressed by the equations

$$V_h = V_{h\pi} + V_{hj} + F_h, \tag{81}$$

$$E_h = E_{h\pi} + E_{hj} + f_h, \tag{82}$$

in accordance with the general equations (77) and (78) respectively. It will be found convenient to call $V_{h\pi}$ the 'systemic potential' and $E_{h\pi}$ the 'systemic axial electric force' at wire h, since $V_{h\pi}$ and $E_{h\pi}$ are due only to the system π of which h is a member, and do not include

[20] Regarding the use here of double subscripts, it will be noted that the first subscript designates the line or the wire where the effect occurs, and the second the wire or the system of wires which produce the effect. Thus, $V_{i\pi}$ is the potential produced in line i by the whole primary system π; the contribution of any one wire h would be denoted by V_{ih}.

any contributions from the secondary system or from the impressed field. (More fully, $V_{h\pi}$ may be termed the 'primary systemic potential' at h and $E_{h\pi}$ the 'primary systemic axial electric force' at h.)

For explicit use below, we may here note the formulas for the systemic potential $V_{k\pi}$ and the systemic axial electric force $E_{k\pi}$ at any wire k of the primary system π:

$$V_{k\pi} = \sum_{h=1}^{n} p_{kh} Q_h, \qquad (k = 1, \cdots n), \tag{83}$$

$$E_{k\pi} = -\sum_{h=1}^{n} \left(Z_{kh} I_h + p_{kh} \frac{dQ_h}{dx} \right), \qquad (k = 1, \cdots n), \tag{84}$$

p_{kh} and Z_{kh} being respectively the mutual potential coefficient and the mutual impedance[3] between wires h and k, per unit length. Equation (84) is obtainable by applying the second curl law to a differential rectangle substantially as in deriving equations (1) and (2); see also Appendix I.

As already stated in connection with equations (79) and (80) the problem to be considered in the present subsection is the calculation of the potential $V_{j\pi}$ and the axial electric force $E_{j\pi}$ produced at the secondary wire j by the primary system π. The fundamental formulas for $V_{j\pi}$ and $E_{j\pi}$ are:

$$V_{j\pi} = \sum_{h=1}^{n} p_{jh} Q_h = \sum_{h=1}^{n} V_{jh}, \tag{85}$$

$$E_{j\pi} = -\sum_{h=1}^{n} \left(Z_{jh} I_h + \frac{dV_{jh}}{dx} \right) = \sum_{h=1}^{n} E_{jh} \tag{86}$$

$$= -\sum_{h=1}^{n} \left(Z_{jh} I_h + p_{jh} \frac{dQ_h}{dx} \right), \tag{87}$$

where V_{jh} and E_{jh} are the contributions of wire h to $V_{j\pi}$ and $E_{j\pi}$ respectively.

With regard to applications of the equations (85) and (87) for the potential $V_{j\pi}$ and the axial electric force $E_{j\pi}$ impressed on the secondary j by the primary π, it will be supposed that all the primary currents $I_1, \cdots I_n$ are known. But the primary charges Q_h and their axial gradients dQ_h/dx (where $h = 1, \cdots n$) are usually not known; and therefore ways will now be indicated for expressing them in terms of quantities which may be known. For that purpose, the presence of the secondary will be entirely ignored, in all respects. (This procedure may be regarded as the first-approximation step in a solution by successive approximations.)

The charges Q_h can be expressed in terms of the systemic potentials $V_{k\pi}$ by solving the set of n equations (83). Thus

$$Q_h = \sum_{k=1}^{n} q_{hk} V_{k\pi}, \qquad (h = 1, \cdots n), \tag{88}$$

where q_{hk} is the Maxwell capacity coefficient between wires h and k; in terms of the potential coefficients, its value is

$$q_{hk} = D_{kh}(p)/D(p), \tag{89}$$

$D(p)$ being the determinant of all the potential coefficients (the p's) in the set of n equations (83) and $D_{kh}(p)$ the cofactor of p_{kh} in $D(p)$.

The systemic potentials $V_{k\pi}$, occurring in (88), can be obtained by solving the equations of current continuity, namely the set of n equations [21]

$$-\frac{dI_h}{dx} = \sum_{k=1}^{n} (Y_{hk} V_{k\pi} + X_{hk} F_k), \qquad (h = 1, \cdots n), \tag{90}$$

Y_{hk} and X_{hk} being of the nature of admittances (per unit length), and F_k the impressed potential at wire k; it is thus found that

$$V_{k\pi} = -\sum^{n} W_{kh} \left(\frac{dI_h}{dx} + \sum_{r=1}^{n} X_{hr} F_r \right), \qquad (k = 1, \cdots n), \tag{91}$$

where the coefficient W_{kh} is the same function of the Y's that q_{kh} is of the p's, that is,

$$W_{kh} = D_{hk}(Y)/D(Y). \tag{92}$$

It is seen that W_{kh} is of the nature of an impedance (per unit length), though it is not a simple impedance.

The charges can now be expressed in terms of the impressed potentials F_r and the axial gradients of the currents by substituting (91) in (88).

The axial gradients of the charges can be expressed in various ways. They can be immediately expressed in terms of the axial gradients of the systemic potentials V_k by merely differentiating (88) with respect to x. Also, they can be expressed in terms of the currents I_r and the systemic axial electric forces $E_{k\pi}$ at the wires, by solving the set of n equations (84); thus

$$\frac{dQ_h}{dx} = -\sum_{k=1}^{n} q_{hk} (E_{k\pi} + \sum_{r=1}^{n} Z_{kr} I_r), \qquad (h = 1, \cdots n), \tag{93}$$

[21] Derived in the latter part of Appendix I.

q_{hk} being the Maxwell capacity coefficient given by (89). Furthermore, the systemic axial electric force $E_{k\pi}$ occurring in (93) is expressible in terms of the current I_k in wire k and the axial electric force f_k impressed on wire k, by the simple relation

$$E_{k\pi} = z_k I_k - f_k, \tag{94}$$

z_k denoting the internal impedance of wire k, per unit length; for, the resultant axial electric force at wire k must be equal to $z_k I_k$ and must also be equal to $E_{k\pi} + f_k$. Thus the axial gradients of the charges can be expressed explicitly in terms of the currents and the impressed axial electric forces at the wires, by substituting (94) in (93). The axial gradients of the charges can be expressed still otherwise by differentiating (88) with respect to x after substituting (91).

Substituting into (85) and (87), the various foregoing expressions for the charge Q_h and its axial gradient dQ_h/dx, and in some cases transforming and rearranging the results, gives the following formulas for the potential $V_{j\pi}$ and the axial electric force $E_{j\pi}$ produced at any point x in the secondary wire j by the primary system π, when the presence of the secondary j is entirely ignored in calculating the currents, charges, and potentials of the primary (in accordance with the statement of the paragraph following equation (87)):

$$V_{j\pi} = \sum_{h=1}^{n} p_{jh} Q_h \tag{95}$$

$$= \sum_{h=1}^{n} T_{jh} V_{h\pi} \tag{96}$$

$$= -\sum_{h=1}^{n} T_{jh} \left[\sum_{k=1}^{n} W_{hk} \left(\frac{dI_k}{dx} + \sum_{r=1}^{n} X_{kr} F_r \right) \right], \tag{97}$$

where T_{jh}, which may be termed a 'potential transfer factor' or 'voltage transfer factor,' has the value

$$T_{jh} = \sum_{k=1}^{n} p_{jk} q_{kh}. \tag{98}$$

$$E_{j\pi} = -\sum_{h=1}^{n} Z_{jh} I_h - \frac{dV_{j\pi}}{dx} \tag{99}$$

$$= -\sum_{h=1}^{n} \left(Z_{jh} I_h + p_{jh} \frac{dQ_h}{dx} \right) \tag{100}$$

$$= -\sum_{h=1}^{n} \left(Z_{jh} I_h + T_{jh} \frac{dV_{h\pi}}{dx} \right) \tag{101}$$

$$= -\sum_{h=1}^{n}\left(Z_{jh}I_h - \sum_{k=1}^{n} T_{jh}W_{hk}\left[\frac{d^2I_k}{dx^2} + \sum_{r=1}^{n} X_{kr}\frac{dF_r}{dx}\right]\right) \quad (102)$$

$$= -\sum_{h=1}^{n}\left(Z_{jh}I_h - T_{jh}E_{h\pi} - T_{jh}\sum_{r=1}^{n} Z_{hr}I_r\right) \quad (103)$$

$$= -\sum_{h=1}^{n}\left(Z_{jh}I_h - T_{jh}[z_hI_h - f_h] - T_{jh}\sum_{r=1}^{n} Z_{hr}I_r\right). \quad (104)$$

Formula for $E_{j\pi}$ when the Earth is a Perfect Conductor

When the earth is a perfect conductor, all of the external mutual and self impedances (Z_{jk}, Z_{hk}, Z_{kk}, etc.) are pure reactances and are proportional to the corresponding potential coefficients (p_{jk}, p_{hk}, p_{kk}, etc.), the proportionality factor being merely $i\omega/\tau$, where τ is an absolute constant whose value depends only on the units employed. Thence it can be shown that (103) and (104) respectively reduce to the very simple formulas

$$E_{j\pi} = \sum_{h=1}^{n} T_{jh}E_{h\pi}, \quad (105)$$

$$E_{j\pi} = \sum_{h=1}^{n} T_{jh}(z_hI_h - f_h), \quad (106)$$

with T_{jh} given by (98). It is seen that (105) corresponds exactly to (96).

As at least of some academic interest, it may be remarked that equations (105) and (106) hold even when the earth is imperfect, provided

$$\frac{Z_{jh}}{p_{jh}} = \frac{Z_{kh}}{p_{kh}}, \qquad (h = 1, \cdots n;\ k = 1, \cdots n).$$

IV

Practical Applications

For illustrative purposes, the methods presented in the foregoing sections will now be applied to two practical problems of a rather diverse nature. The first application will be to the wave antenna employed in certain important cases of long-distance radio reception,[22] the second, to a problem in crosstalk.

The Wave Antenna

The wave antenna, in its usual form, may be described as a transmission line with ground return, utilized for the reception of radio waves.

[22] Notably in transoceanic radio telephony.

In its simplest form, as here contemplated, the wave antenna consists of a long straight horizontal wire terminated at each end in its characteristic impedance K, as represented by Fig. 12*a*, which gives

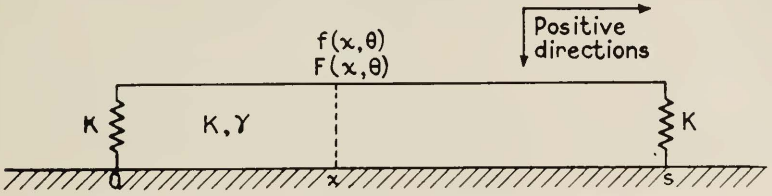


Fig. 12*a*

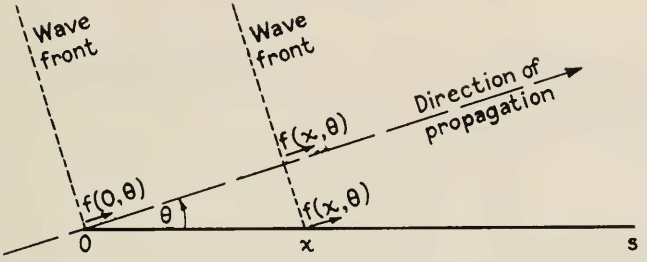


Fig. 12*b*

an elevation view. This is seen to be the same as Fig. 1 when $Z_0 = Z_s = K$; and γ is now the propagation constant, per unit length, of the wave antenna regarded as a transmission line. Hence formulas (46), $\cdots$ (54), pertaining to Fig. 1, are immediately applicable for calculating the current at any point x in the wave antenna of Fig. 12*a*, after the appropriate formulation of the functions $f(x)$ and $F(x)$—namely the impressed axial electric force and the impressed potential, respectively, at any point x in the wave antenna.

These functions can be evaluated by aid of Fig. 12*b*, which gives a plan view representing a train of plane radio waves (whose magnetic component is horizontal) incident on the wave antenna at an arbitrary angle θ measured horizontally from the wave antenna to the direction of propagation of the wave train along the earth's surface. By the 'direction of propagation' is here meant the horizontally specified direction of a vertical plane which is normal to the plane of the wave front. $f(x, \theta)$ denotes the horizontal component of electric force in the impressed waves at any point x of the wave antenna, and $F(x, \theta)$ the potential of the impressed waves there.[23] Then the axial electric

[23] The presence of θ in the functional symbols is of course to allow for a possible

force $f(x)$ and the potential $F(x)$ impressed at point x on the wave antenna by the radio waves are given by the equations

$$f(x) = f(x, \theta) \cos \theta, \tag{107}$$

$$F(x) = F(x, \theta). \tag{108}$$

It is convenient to take one end, say $x = 0$, as a fixed reference point, and then to express $f(x, \theta)$ and $F(x, \theta)$ in terms of their values $f(0, \theta)$ and $F(0, \theta)$ at $x = 0$. For this purpose, it will be assumed that the radio waves are propagated in a simple exponential manner, so that

$$\frac{f(x, \theta)}{f(0, \theta)} = \frac{F(x, \theta)}{F(0, \theta)} = e^{-\Gamma x \cos \theta}, \tag{109}$$

Γ denoting the propagation constant of the radio waves, per unit length measured horizontally along the direction of their propagation. Then the equations (107) and (108) become

$$f(x) = f(0, \theta) \cos \theta \, e^{-\Gamma x \cos \theta}, \tag{110}$$

$$F(x) = F(0, \theta) e^{-\Gamma x \cos \theta}, \tag{111}$$

wherein $f(0, \theta)$ and $F(0, \theta)$ may be supposed known. In this connection it should be remarked that $f(0, \theta)$ and $F(0, \theta)$—and, more generally, $f(x, \theta)$ and $F(x, \theta)$—are not in phase.[24]

On substitution of (110) and (111), equations (49), $\cdots$ (53) now become applicable for calculating the current $I(x)$ at any point x of the wave antenna; this current will be written $I(x, \theta)$ because it depends on the incidence-angle θ, even when $f(x, \theta)$ and $F(x, \theta)$ are independent of θ. In the engineering of wave antennæ, we are usually concerned merely with the current $I(s, \theta)$ received at the end $x = s$. In general there will be four constituents of $I(s, \theta)$, corresponding to equations (50), (51), (52), (53) when $x = s$. From the discussion of the corresponding more general equations (41), (42), (43), (44), it will be recalled that the current-constituent $j(s, \theta)$ is due to the impressed axial electric force acting throughout the length of the wave antenna, $J_0(s, \theta)$ is due to the impressed voltage $F(0, \theta)$ acting at the end $x = 0$, $J_s(s, \theta)$ is due to the corresponding impressed voltage $F(s, \theta) = F(0, \theta) e^{-\Gamma s \cos \theta}$ acting at the end $x = s$, and $J_{0s}(s, \theta)$

dependence on θ. It may be noted that, in the calculation of the ordinary polar diagram representing the directional selectivity of a wave antenna, the functions $f(x, \theta)$ and $F(x, \theta)$ are regarded as independent of θ, in accordance with the very definition of the directional selectivity.

[24] The ratio of the horizontal electric force $f(x, \theta)$ to the vertical electric force $F(x, \theta)/H$—where H here denotes the height of the wave antenna above the earth's surface—is a complex number whose value depends on the conductivity, dielectric constant, and permeability of the ground, and on the frequency.

is due to the distributed impressed voltage acting in the leakage admittance from the wave antenna to ground (this leakage admittance being regarded as uniformly distributed). By substituting the values $f(y)$ and $F(y)$ given by (110) and (111) when x is replaced by y, then carrying out the indicated integrations, and finally transforming the results somewhat, the constituents corresponding to (50), (51), and (52) are found to have the following formulas:

$$j(s, \theta) = \frac{sf(0, \theta) \cos \theta}{2K} \frac{\sinh [(\gamma - \Gamma \cos \theta)s/2]}{(\gamma - \Gamma \cos \theta)s/2} e^{-(\gamma + \Gamma \cos \theta)s/2}, \quad (112)$$

$$J_0(s, \theta) = -\frac{F(0, \theta)}{2K} e^{-\gamma s}, \quad (113)$$

$$J_s(s, \theta) = \frac{F(0, \theta)}{2K} e^{-\Gamma s \cos \theta}. \quad (114)$$

The fourth constituent, $J_{0s}(s, \theta)$, corresponding to (53), will be omitted, because it is relatively unimportant and also because its formula is found to be somewhat lengthy.

The valuable directional selectivity of a wave antenna resides mainly in the directional properties of the admittance $j(s, \theta)/sf(0, \theta')$ whose value is found by dividing equation (112) through by $sf(0, \theta')$, where θ' denotes some fixed value of θ (usually $\theta' = 0$). This ratio may properly be termed a 'directional admittance.' The corresponding admittances obtained by dividing (113) and (114) through by $sf(0, \theta')$ are not usefully directional, the former being entirely non-directional, and the latter only directional as regards its phase angle—not as regards its absolute value. By suitable choice of the length of the wave antenna, the constituent represented by (112) can be made to have high directional selectivity, while the constituents corresponding to (113) and (114) become relatively unimportant (except over a few narrow ranges of the incidence angle θ).[25]

A Crosstalk Problem

This problem is concerned with the derivation of formulas for the first-order crosstalk between two simple open-wire telephone circuits of which one is non-transposed and the other is once-transposed, as represented in plan view by Fig. 13.

The once-transposed circuit is taken as the primary, and the non-transposed as the secondary. Each extends from $x = 0$ to $x = s$; and the primary is transposed at its mid-point $x = s/2$.

[25] For a detailed study of the wave antenna, the reader is referred to the well-known paper by Beverage, Rice, and Kellog entitled 'The Wave Antenna' in *J. A. I. E. E.* beginning with March, 1923.

The primary is energized by an alternating electromotive force E_0 inserted at $x = 0$. Stated precisely, the problem here contemplated is the derivation of formulas for the currents produced in the two ends, $x = 0$ and $x = s$, of the secondary circuit by the primary circuit, when all reactions of the secondary on the primary are neglected.

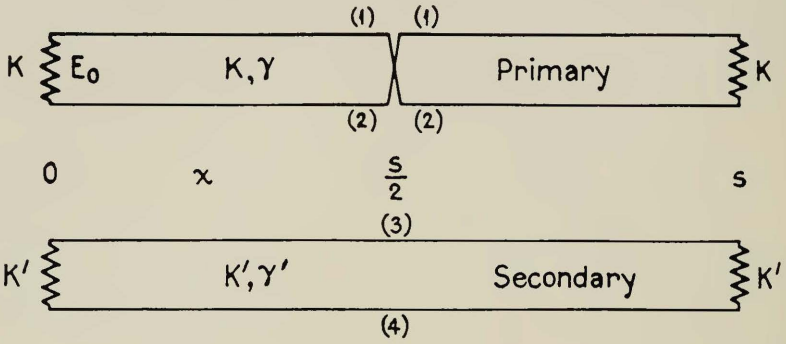


Fig. 13

The wires of the secondary circuit are numbered 3 and 4. The primary wires are numbered 1 and 2, in the sense that 1 and 2 designate the *positions* the wires would occupy if non-transposed; in this sense, wires 1 and 2 are each discontinuous at $x = s/2$, the transposition cross thus being regarded as extraneous to the wires.[26]

Each circuit is terminated at each end in its mode-a characteristic impedance—K for the primary, K' for the secondary. The mode-a propagation constants of the primary and secondary, per unit length, are denoted by γ and γ' respectively.

The earth is assumed to be a perfect conductor. This assumption is effectively a good approximation because the contemplated circuits are such that the distance between the two wires of each circuit is small compared with their height above ground; at the same time the assumption greatly facilitates and simplifies the solution.

Evidently the first step is to formulate the primary current and the primary systemic potential at any point x. The second step is to formulate the electric field impressed on the secondary by the primary. The third and final step is to formulate the currents produced in the secondary by the impressed field of the primary; this third step will be carried through by means of the synthetic method, employing the set of equivalent electromotive forces formulated in the early part of Section III.

[26] The transposition cross may conveniently be regarded as merely a particular kind of transducer (four-terminal network) inserted in the primary, namely a reversing transducer.

Let $I_r(x) = I_r$ and $V_r(x) = V_r$ denote the current and the potential, respectively, at any point x of wire r, where $r = 1, 2, 3, 4$. Then, evidently, for the primary currents and potentials we have:

$$I_1 = -I_2, \qquad V_1 = -V_2, \tag{115}$$

$$V_1 - V_2 = KI_1. \tag{116}$$

For $x \leqq s/2$:

$$I_1 = \pm \frac{E_0}{2K} e^{-\gamma x}, \tag{117}$$

$$V_1 = \pm \frac{E_0}{4} e^{-\gamma x}. \tag{118}$$

Thus the primary currents, $I_1(x)$ and $I_2(x)$, and the corresponding primary potentials, $V_1(x)$ and $V_2(x)$, are each discontinuous at $x = s/2$ (by reason of the transposition there).

The electric field impressed on the individual wires of the secondary circuit by the primary circuit can be formulated by means of equations (106) and (96). Thus

$$E_3 = (T_{31} - T_{32})zI_1, \tag{119}$$

$$E_4 = (T_{41} - T_{42})zI_1, \tag{120}$$

$$V_3 = (T_{31} - T_{32})V_1, \tag{121}$$

$$V_4 = (T_{41} - T_{42})V_1, \tag{122}$$

where $z = z_1 = z_2$ is the internal impedance of each wire of the primary, per unit length, and the T's are 'voltage transfer factors' given by (98).

Evidently the secondary circuit constitutes a balanced two-wire line in an arbitrary impressed field (the field due to the primary), and hence is amenable to the treatment already fully described and formulated in the subsection following equation (54). Thus the current at any point x in the secondary consists of two modes, a and c. However, as already indicated, we shall ultimately be concerned only with the currents in the ends $x = 0$ and $x = s$ of the secondary; evidently these are mode-a currents, for at each end the mode-c currents must be zero, since the circuit is insulated from ground at each end.

As we shall be concerned only with the mode-a currents produced in the secondary, the next step is to formulate the mode-a constituents of the electric field impressed by the primary. If E' and V' denote the mode-a constituents of the axial electric force and of the potential

impressed by the primary, then

$$E' = E_3 - E_4 = 4TzI_1, \tag{123}$$

$$V' = V_3 - V_4 = 4TV_1, \tag{124}$$

where

$$T = (T_{31} - T_{32} - T_{41} + T_{42})/4. \tag{125}$$

Remembering that I_1 and V_1 are discontinuous at $x = s/2$, in accordance with equations (117) and (118), it is seen that E' and V' are discontinuous at $x = s/2$ in accordance with the following equations.[27] For $x \lesseqgtr s/2$:

$$E' = \pm E_0 T \frac{2z}{K} e^{-\gamma x}, \qquad V' = \pm E_0 T e^{-\gamma x}. \tag{126}$$

We are now prepared to formulate the mode-a currents produced in the secondary (3, 4) by the field arising from the primary (1, 2). Since the secondary constitutes a balanced two-wire line in an arbitrary impressed field, it is amenable to the treatment formulated in the subsection following equation (54); thence equations (41), $\cdots$ (45) and equations (50), $\cdots$ (54) are formally applicable.

If $I_3(x)$ and $I_4(x)$ denote the mode-a currents at any point x in the secondary wires 3 and 4 respectively, then

$$I_3(x) = -I_4(x) = I(x), \text{ say.} \tag{127}$$

On referring to the subsection containing equations (41), $\cdots$ (45) and applying it to the mode-a effects in the present problem, it will be seen that $I(x)$ is the sum of the five mode-a constituents $j(x)$, $J_0(x)$, $J_s(x)$, $J_{0s}(x)$, $J_u(x)$, corresponding to equations (41), (42), (43), (44), (45) respectively. From the discussion in connection with those equations and from the analysis of the impressed field into two modes, a and c, as described and formulated in the subsection following equation (54), it will be seen that $j(x)$ is due to the mode-a axial electric force $E_3(y) - E_4(y)$ acting at all points y of the secondary, $J_0(x)$ is due to the mode-a impressed voltage $V_3(0) - V_4(0)$ acting at the end $y = 0$, $J_s(x)$ is due to the mode-a impressed voltage $V_3(s) - V_4(s)$ acting at the end $y = s$, $J_{0s}(x)$ is due to the mode-a impressed voltage $V_3(y) - V_4(y)$ acting at all points y in the leakage admittance[28] between the secondary wires, and $J_u(x)$ is due to the discontinuity $V'(u -)$

[27] From mere physical considerations, it is evident that the whole primary field is reversed at $x = s/2$.

[28] That is, the 'mutual' leakage admittance (equal to the 'direct' leakage admittance between wires plus one half of the 'direct' leakage admittance from each wire to ground).

$-\ V'(u+)$ in the mode-a impressed voltage $V'(y) \equiv V_3(y) - V_4(y)$ at $y = u = s/2$. (In what follows, the constituent $J_{0s}(x)$ will be omitted because it is relatively unimportant.)

Thus, for the formulation of the mode-a effects the functions $f(y)$ and $F(y)$, representing the electirc forces and potentials in equations (41), $\cdots$ (45) and (50), $\cdots$ (54), have the following mode-a values:

$$f(y) = E_3(y) - E_4(y) = E'(y), \tag{128}$$

$$F(y) = V_3(y) - V_4(y) = V'(y), \tag{129}$$

whence, in particular,

$$F(0) = V'(0), \tag{130}$$

$$F(s) = V'(s), \tag{131}$$

$$F(u-) - F(u+) = V'(u-) - V'(u+). \tag{132}$$

Substituting these values into equations (50), (51), (52), (54), and carrying out the indicated integrations[15] when $x = 0$ and when $x = s$, and finally dividing each equation by the value of the primary current $I_1(0) = E_0/2K$ at $x = 0$, we obtain the following formulas (134), $\cdots$ (137) for the four mode-a current ratios at $x = 0$, and the formulas (141), $\cdots$ (144) for those at $x = s$. Also, there are included formulas for $J(x)/I_1(0)$ at $x = 0$ and at $x = s$, $J(x)$ denoting the sum of the mode-a current constituents due to the impressed potential, that is,

$$J(x) = J_0(x) + J_s(x) + J_u(x) \tag{133}$$

since $J_{0s}(x)$ is neglected.

At $x = 0$ the formulas for the four current-ratios are

$$\frac{j(0)}{I_1(0)} = T\frac{K}{K'}\frac{sz}{K}\frac{[1 - e^{-(\gamma+\gamma')s/2}]^2}{(\gamma+\gamma')s/2}, \tag{134}$$

$$\frac{J_0(0)}{I_1(0)} = -T\frac{K}{K'}, \tag{135}$$

$$\frac{J_s(0)}{I_1(0)} = -T\frac{K}{K'}e^{-(\gamma+\gamma')s}, \tag{136}$$

$$\frac{J_u(0)}{I_1(0)} = 2T\frac{K}{K'}e^{-(\gamma+\gamma')s/2}. \tag{137}$$

The sum of the last three is

$$\frac{J(0)}{I_1(0)} = -T\frac{K}{K'}[1 - e^{-(\gamma+\gamma')s/2}]^2. \tag{138}$$

On dividing (134) by (138) the ratio of $j(0)$ to $J(0)$ is found to have the simple value

$$\frac{j(0)}{J(0)} = -\frac{z}{K(\gamma + \gamma')/2}. \tag{139}$$

In particular, when the two circuits have equal propagation constants ($\gamma' = \gamma$),

$$\frac{j(0)}{J(0)} = -\frac{z}{Z}, \tag{140}$$

where $Z = \gamma K$ is the mode-a 'complete series impedance'[18] of the primary, per unit length; it will be recalled that z is the 'internal impedance' of each primary wire, per unit length. The ratio z/Z is ordinarily a very small fraction.

At $x = s$ the formulas for the four current-ratios are

$$\frac{j(s)}{I_1(0)} = T\frac{K}{K'}\frac{sz}{K}\frac{[1 - e^{-(\gamma-\gamma')s/2}]^2}{(\gamma - \gamma')s/2}e^{-\gamma's}, \tag{141}$$

$$\frac{J_0(s)}{I_1(0)} = -T\frac{K}{K'}e^{-\gamma's}, \tag{142}$$

$$\frac{J_s(s)}{I_1(0)} = -T\frac{K}{K'}e^{-\gamma s}, \tag{143}$$

$$\frac{J_u(s)}{I_1(0)} = 2T\frac{K}{K'}e^{-(\gamma+\gamma')s/2}. \tag{144}$$

The sum of the last three is

$$\frac{J(s)}{I_1(0)} = -T\frac{K}{K'}[1 - e^{-(\gamma-\gamma')s/2}]^2 e^{-\gamma's}. \tag{145}$$

On dividing (141) by (145) the ratio of $j(s)$ to $J(s)$ is found to have the simple value

$$\frac{j(s)}{J(s)} = -\frac{z}{K(\gamma - \gamma')/2}. \tag{146}$$

When the absolute value of $(\gamma - \gamma')s/2$ is small compared to unity, equations (141) and (145) become approximately

$$\frac{j(s)}{I_1(0)} = T\frac{K}{K'}\frac{sz}{K}\frac{(\gamma - \gamma')s}{2}e^{-\gamma's}, \tag{147}$$

$$\frac{J(s)}{I_1(0)} = -T\frac{K}{K'}\left[\frac{(\gamma - \gamma')s}{2}\right]^2 e^{-\gamma's}. \tag{148}$$

Thus:

$$\text{When} \quad \gamma' = \gamma: \quad j(s) = 0, \qquad J(s) = 0.$$

For some cases, particularly those where the attenuation is neglected, it is advantageous to express the square-bracketed factors in equations (134), (138), (141), (145) partially in terms of hyperbolic sines.

Appendix I

Derivations of Equations (25) and (90)

Equation (25)

Let the primary or impressed field of force be specified by an electric intensity f_a parallel to the axis of the wire (and to the surface of the earth), and an electric intensity f_n normal to the surface of the earth and measured downward. We denote by f_w the value of f_a at the axis of the wire,[29] and by f_g its value at the surface of the ground in the plane which is normal to the ground and which includes the axis of the wire. The impressed or primary potential F of the wire, due to the impressed field, is then

$$F = \int_0^h f_n dy,$$

where h is the height of the wire above ground and the integral is taken along the vertical from the wire ($y = 0$) to ground ($y = h$).

Due to the impressed field, specified above, a current I flows in the wire and a corresponding *superposed* current distribution is induced in the ground. The *resultant* axial electric intensity at the surface of the wire is then $z_w I$ (where z_w is the internal impedance of the wire, per unit length); correspondingly the *resultant* electric intensity along the surface of the ground is $f_g - z_g I$. Application of the second curl law to a contour composed of two verticals from the wire to ground and the line segments dx in the surfaces of the wire and ground gives

$$(z_w + z_g)I - f_g + \frac{dV}{dx} = -\frac{d\phi}{dt},$$

which is preferably written as

$$zI - f_g + \frac{dV}{dx} = -\frac{d\phi}{dt}, \tag{1}$$

where $z = z_w + z_g$ is the internal impedance of the circuit, per unit length; V is the *resultant* potential of the wire; and ϕ is the *resultant* magnetic flux threading the contour, per unit length. But we have

[29] f_w is assumed to be sensibly constant over the cross-section of the wire.

also

$$f_w - f_g + \frac{dF}{dx} = -\frac{d\Phi}{dt}, \tag{2}$$

where Φ denotes the impressed magnetic flux threading the contour, per unit length. Subtracting (2) from (1) and observing that

$$\begin{aligned} V - F &= \frac{1}{C}Q, \\ \phi - \Phi &= LI, \end{aligned} \tag{3}$$

we get

$$zI + L\frac{dI}{dt} + \frac{1}{C}\frac{dQ}{dx} = f_w, \tag{4}$$

where, of course, Q is the charge, C the capacity to ground and L the external inductance, per unit length of the wire.

To eliminate Q from (4) we make use of the equation of current continuity, namely

$$-\frac{dI}{dx} = \frac{dQ}{dt} + I', \tag{5}$$

where I' is the leakage current per unit length of the wire. If the wire is embedded in a homogeneous leaky medium, then

$$I' = \frac{G^0}{C}Q = G^0(V - F), \tag{6}$$

where G^0 is proportional to the conductivity of the medium.[30] If, furthermore, there is direct leakage admittance from the wire to ground (as at poles and insulators) of amount Y' per unit length,[31] when regarded as uniformly distributed, then

$$I' = \frac{G^0}{C}Q + Y'V = \frac{G}{C}Q + Y'F, \tag{7}$$

where

$$G = G^0 + Y'. \tag{8}$$

On substituting the last value of I' into (5), setting $d/dt = i\omega$, then differentiating with respect to x, and finally substituting the resulting value of dQ/dx into (4), we get

$$(z + i\omega L)I - \frac{1}{G + i\omega C}\frac{d^2I}{dx^2} = f_w + \frac{Y'}{G + i\omega C}\frac{dF}{dx}, \tag{9}$$

[30] A formula for G^0 is equation (18) derived below.

[31] While G^0 is merely a pure conductance, Y' is in general an admittance (leakage admittance), because the insulators and poles have capacity as well as conductance. Hence G, defined by (8), is an admittance.

which can be written

$$\frac{K}{\gamma}\left(\gamma^2 - \frac{d^2}{dx^2}\right) I = f_w + \frac{Y'K}{\gamma}\frac{dF}{dx}, \tag{10}$$

where

$$K = \sqrt{\frac{z + i\omega L}{G + i\omega C}}, \tag{11}$$

and

$$\gamma = \sqrt{(z + i\omega L)(G + i\omega C)}. \tag{12}$$

Thus K is the characteristic impedance and γ the propagation constant of the transmission system composed of the overhead wire with ground return; it is to be noted that $G = G^0 + Y'$, in accordance with (8), and hence that G is in general an admittance—not a pure conductance.

If we define f' by the equation

$$f' = f_w + \frac{Y'K}{\gamma}\frac{dF}{dx}, \tag{13}$$

then (10) becomes

$$\frac{K}{\gamma}\left(\gamma^2 - \frac{d^2}{dx^2}\right) I = f', \tag{14}$$

which is formally the same as equation (25) of the text. There, however, it is assumed that the term $(Y'K/\gamma)dF/dx$ is negligible; probably this is usually the case but circumstances may arise where it is not negligible. Its inclusion, however, introduces no formal modification of the analysis.

The foregoing derivation has been given in detail because prior derivations known to the writers have not been entirely satisfactory. Their chief defect has been that no explicit consideration was given to the finite conductivity of the ground (except that it produces a tangential component f_a). In the derivation given above, the effect of ground conductivity is expressly recognized and in the final equation appears implicitly in the values of K and γ. These parameters, it will be observed, are experimentally determinable, and are the only parameters besides Y' appearing in the final differential equation.

A formula for the quantity G^0 occurring in equations (6) and (7) can be derived by application of Gauss' theorem, as follows: Let E_r denote the radial component of the total electric force at the surface of the wire, dS a differential element of the surface of the wire, and σ and ϵ the conductivity and specific inductive capacity of the medium, which is homogeneous and isotropic by assumption. Then the leakage

current I' flowing outward, per unit length of the wire, is given by

$$I' = \int \sigma E_r dS = \frac{\sigma}{\epsilon} \int \epsilon E_r dS, \tag{15}$$

the surface integral being taken over the unit length of the wire. But, by Gauss' theorem (ξ being a constant whose value depends only on the units),

$$\int \epsilon E_r dS = 4\pi Q/\xi, \tag{16}$$

the resultant axial electric flux from the ends of the element being negligible compared with the radial electric flux from the lateral surface. Thus

$$I' = \frac{4\pi\sigma}{\epsilon\xi} Q, \tag{17}$$

and comparison of this equation with (6) gives the result

$$G^0/C = 4\pi\sigma/\epsilon\xi. \tag{18}$$

In this connection it may be noted that the leakage current represented by (15) does not directly depend on the impressed field, but only on the field produced by the wire itself. This is because the assumed medium is homogeneous and isotropic; hence σ in (15) can be taken outside the sign of integration, and then the conclusion follows from (15) by noting that one of the constituents of E_r is the impressed radial electric force f_r, and that

$$\int f_r dS = \int \operatorname{div} f \cdot dv = 0,$$

since the divergence of the impressed electric force must be zero. The conclusion would not follow, in general, if the medium were either heterogeneous or æolotropic. It may be noted that a homogeneous isotropic medium surrounding a wire and containing direct leakage admittance paths from the wire to ground may be regarded as a heterogeneous æolotropic medium.

The value given for δ in equation (9) of the text, namely $\delta = 4\pi\sigma/\epsilon\xi$, is readily derivable by combining equations (4), (10), (5) of the text with (17) of this Appendix.

Equations (90)

If there were no impressed potential at the primary wires ($F_k = 0$), the equations of continuity would be merely

$$-\frac{dI_h}{dx} = \sum_{k=1}^{n} Y_{hk} V_{k\pi}, \qquad (h = 1, \cdots n), \tag{19}$$

where, in accordance with equations (7) in Section I,

$$Y_{hk} = g_{hk} + i\omega q_{hk}. \tag{20}$$

It should here be remarked that Y_{hk} depends not only on the geometry of the system and on the conductivity of the medium but also on any direct leakage admittance existing between the wires themselves and also on any between the wires and ground. The direct leakage admittance between wires h and k, per unit length, will be denoted by Y'_{hk} and that between wire h and ground, by Y'_{hh}; these are regarded as being uniformly distributed along the system.

When there is present an impressed potential, the existence of the direct leakage admittances gives rise to the following supplementary terms for the right side of equation (19):

$$F_h Y'_{hh} + \sum_{\substack{k=1 \\ k\neq h}}^{n} (F_h - F_k) Y'_{hk} = \sum_{k=1}^{n} X_{hk} F_k,$$

where

$$\begin{aligned} X_{hk} &= -Y'_{hk} \quad \text{for} \quad k \neq h, \\ X_{hh} &= \sum_{k=1}^{n} Y'_{hk}. \end{aligned} \tag{21}$$

It is seen that Y_{hk} and X_{hk} are of the nature of admittances (per unit length), although they are not 'direct admittances.' Their precise meanings are readily deducible from equations (90).

When the medium itself is of zero conductivity, g_{hk} reduces to X_{hk}.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

Direct Determination of Hydrocarbon in Raw Rubber, Gutta-Percha, and Related Substances.[1] A. R. KEMP. Iodochloride in glacial acetic acid is shown to be a suitable reagent to determine the unsaturation of the hydrocarbon in rubber or gutta-percha. The influence of time, temperature, sunlight, and reagent concentration upon the reaction is shown.

Comparisons are made between iodochloride, iodobromide and bromine relative to their reactions with raw rubber and some of the terpenes.

Results of analyses of several rubber and gutta-percha samples are given.

The effects of mastication and heat upon the unsaturation of raw rubber are shown.

Microtomic Preparation of Soft Metals for Microscopic Examination.[2] F. F. LUCAS. This paper outlines the apparent limitations of polishing methods for preparing specimens of soft metals for metallographic examination. A microtome method has been developed and its successful application to the study of lead cable sheath alloys illustrated. Much time and labor are saved and results have been obtained which were impossible by polishing methods.

In lead-antimony cable sheath alloys a widened grain boundary phenomenon was disclosed by the new method and the probable nature of the structural changes determined. Changes in structure due to aging are shown and those which accompany thermal or mechanical treatment of the metal may be followed clearly.

Distribution of Energy in Worked Metals.[3] LYALL ZICKRICK and R. S. DEAN. The purpose of this paper is to give experimental results that seem to be in accord with the theory, that in the deformation of a crystal the energy supplied is distributed to the atoms of the lattice, probably by the forced formation of molecules.

[1] *Journ. Ind. and Engr. Chem.*, Vol. 19, p. 531, 1927.
[2] Institute of Metals Division, A. I. M. E., February 15, 1927.
[3] *Wire*, Vol. 2, p. 161, 1927.

Modern Developments in Inspection Methods.[4] E. D. HALL. This extensively illustrated article describes inspection methods as carried out at the Hawthorne plant of the Western Electric Company. The number of individual piece parts manufactured is more than 100,000 and the number of inspection gauges employed totals more than 25,000. Machine testing and gauging for certain parts is described and cost savings resulting therefrom are given. One of the machines described tests a porcelain protector block containing a carbon insert. The machine is adjusted to accept blocks from which the recess distance of the carbon lies between 0.0024 and 0.0032 inch and to reject, when the distance is 0.0023 inch or less or 0.0033 inch or more. The machine performs its operation at the rate of 2,800 blocks per hour. It is used to test about 4,500,000 blocks per year and represents an annual saving of approximately $2,500.

A Direct Comparison of the Loudness of Pure Tones.[5] B. A. KINGSBURY. The loudness of eleven pure tones was studied by adjusting the voltage applied to a telephone receiver to make these tones as loud as certain fixed levels of a 700-cycle tone. The average results of 22 observers, 11 men and 11 women, were arranged as contour lines of equal loudness through the normal auditory sensation area in terms of r.m.s. pressure in car canal as a function of frequency. Frequencies from 60 to 4,000 cycles were used and intensities from threshold of audibility to 90 T. U. above the 700-cycle threshold. It was found that if the amplitudes of pure tones are increased in equal ratios the loudness of low frequency tones increases much more more rapidly than that of high frequency tones. For frequencies above 700 cycles the rate is nearly uniform.

As a loudness unit the least perceptible increment of loudness of a 1,000-cycle tone was employed. In absolute magnitude this varies from level to level, but in the ordinary range of loudness it becomes constant. This unit takes into account the subjective character of loudness.

The variability of the data from which the averages were computed was separated into a factor expressing dissimilarity of ears and another expressing errors of observers' judgment. There was no level at which the variances were a minimum. Dissimilarity of ears causes more variation than errors of observers' judgment. The variances showed no significant sex difference.

[4] *Mechanical Engineering*, Vol. 48, p. 1435, 1926.
[5] *Physical Review*, Vol. 29, p. 588, 1927.

The Scattering of Electrons by a Single Crystal of Nickel.[6] C. DAVISSON and L. H. GERMER. Preliminary announcement is made in this note of the discovery that a beam of swiftly moving electrons in its reaction with a single crystal of nickel behaves in some respects as if it were a beam of wave radiation such as light or x-rays. As the speed of the electron beam is increased a series of critical speeds is found at which sharply defined beams of scattered electrons issue from the crystal. This is similar to what is observed when a beam of monochromatic x-rays is sent into a crystal—as the wave-length of the x-rays is decreased a series of critical wave-lengths is found at which sharply defined beams of scattered x-rays issue from the crystal. This x-ray phenomenon is quantitatively accounted for as due to the interference of waves scattered by the regularly arranged atoms of the crystal. In fact, it was this phenomenon discovered by Laue, Friedrick and Knipping in 1913 that established the wave nature of x-radiation, and it is from measurements based on this phenomenon that the lengths of x-ray waves are determined.

The analogous electron phenomenon is less simple, and yet it is simple enough and of such a nature as to leave little doubt that a beam of swiftly moving electrons is in some sense equivalent to a beam of wave radiation. The wave-length of the equivalent radiation can be measured, and is found to be in satisfactory agreement with requirements of the new theory of wave or undulatory mechanics: namely, that the wave-length of the equivalent radiation shall be equal to h/mv, where h represents Planck's universal constant of action, and mv the momentum of an individual electron.

Structure of a Protective Coating of Iron Oxides.[7] RICHARD M. BOZORTH. It is shown that the Bower-Barff protective coating, produced by the action of steam on iron at about 700° with subsequent cooling in air, is built up of layers of ferrous oxide, magnetite and ferric oxide, arranged in this order (the order of oxidation) upon the iron base. The thicknesses of these layers are estimated to be of the order of 10^{-2}, 2×10^{-4} and 2×10^{-5} cm., respectively. The data on which the above conclusions are based are the positions and intensities of lines on powder photographs taken with molybdenum, iron and copper $K\alpha$ X-rays. The iron and copper $K\alpha$ X-rays penetrate the coating to different depths and give information about different parts of its structure because their wave-lengths are, respectively, a little greater and a little less than the critical-absorption wave-length of the iron which forms the greater part of the coating.

[6] *Nature, 119,* 558 (1927).
[7] *J. Amer. Chem. Soc.,* Vol. 49, pp. 969–976 (1927).

Contributors to this Issue

J. G. Ferguson, B.S., University of California, 1915; M.S., 1916; research assistant in physics, 1915–16; Engineering Department of the Western Electric Company and Bell Telephone Laboratories, 1917–. Mr. Ferguson's work has been in connection with the development of methods of electrical measurement.

J. J. Gilbert, A.B., University of Pennsylvania, 1909; Harvard, 1910–11; Chicago, 1911–12; E.E., Armour Institute, 1915; instructor of electrical engineering, Armour, 1912–17; Captain Signal Corps, 1917–19; Engineering Department, Western Electric Company, 1919; Bell Telephone Laboratories, Inc., 1925–. Mr. Gilbert has worked primarily on submarine cable problems.

A. A. Clokey was employed before the war in the Engineering Department of the Western Union Telegraph Co. During the war, he served as captain in the Signal Corps working on experimental investigations to improve cable communication. In 1919, he joined Bell Telephone Laboratories and has since been in charge of the development of terminal equipment for permalloy loaded cables.

A. M. Curtis came to the Engineering Department of the Western Electric Company in 1913 after having spent several years as radio engineer for the Brazilian Government. During the war he was commissioned and sent to France to serve in the Division of Research and Inspection of the Signal Corps. In 1919, he returned to Bell Telephone Laboratories and has since been engaged with the applications of vacuum tube amplifiers to submarine cables.

E. Peterson, Cornell University, 1911–14; Brooklyn Polytechnic, E.E., 1917; Columbia, A.M., 1923; Ph.D., 1926; Electrical Testing Laboratories, 1915–17; Signal Corps, U. S. Army, 1917–19; Engineering Dept., Western Electric Co., 1919–24; Bell Tel. Labs., 1924–.

H. P. Evans, B.S. in electrical engineering, University of Wisconsin, 1923; M.S., 1927. Mr. Evans was with Bell Telephone Laboratories from 1923 to 1925, returning to the University of Wisconsin as instructor in electrical engineering and then as research assistant in physics.

Mr. Evans' contributions to the present paper were made while still a member of the Laboratories staff.

Edward C. Molina, Engineering Department of the American Telephone and Telegraph Company, 1901–19, as engineering assistant; transferred to the Circuits Design Department to work on machine switching systems, 1905; Department of Development and Research, 1919–. Mr. Molina has been closely associated with the application of the mathematical theory of probabilities to trunking problems and has taken out several important patents relating to machine switching.

John R. Carson, B.S., Princeton, 1907; E.E., 1909; M.S., 1912; Research Department, Westinghouse Electric and Manufacturing Company, 1910–12; instructor of physics and electrical engineering, Princeton, 1912–14; American Telephone and Telegraph Company, Engineering Department, 1914–15; Patent Department, 1916–17; Engineering Department, 1918; Department of Development and Research, 1919–. Mr. Carson's work has been along theoretical lines and he has published several papers on theory of electric circuits and electric wave propagation.

Ray S. Hoyt, B.S. in electrical engineering, University of Wisconsin, 1905; Massachusetts Institute of Technology, 1906; M.S., Princeton, 1910; American Telephone and Telegraph Company, Engineering Department, 1906–07; Western Electric Company, Engineering Department, 1907–11; American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919–. Mr. Hoyt has made contributions to the theory of transmission lines and associated apparatus, and more recently to the theory of crosstalk and other interference.

The Bell System Technical Journal

October, 1927

Television [1]

By HERBERT E. IVES

Synopsis: The chief problems presented in the accomplishment of television are discussed. These are: the resolution of the scene into a series of electrical signals of adequate intensity for transmission; the provision of a transmission channel capable of transmitting a wide band of frequencies without distortion; means for utilizing the transmitted signals to re-create the image in a form suitable for viewing by one or more observers; arrangements for the accurate synchronization of the apparatus at the two ends of the transmission channel.

Introduction

THIS paper is to serve as an introduction to the group of papers following, which describe the apparatus and methods used in the recent experimental demonstration of television over communication channels of the Bell System. In that demonstration television was shown both by wire and by radio. The wire demonstration consisted in the transmission of images from Washington, D. C., to the auditorium of the Bell Telephone Laboratories in New York, a distance of over 250 miles by wire. In the radio demonstration, images were transmitted from the Bell Laboratories experimental station at Whippany, New Jersey, to New York City, a distance of 22 miles. Reception was by two forms of apparatus. In one, a small image approximately 2 in. by 2½ in. was produced, suitable for viewing by a single person, in the other a large image, approximately 2 ft. by 2½ ft., was produced, for viewing by an audience of considerable size. The smaller form of apparatus was primarily intended as an adjunct to the telephone, and by its means individuals in New York were enabled to see their friends in Washington with whom they carried on telephone conversations. The larger form of receiving apparatus was designed to serve as a visual adjunct to a public address system. Images of speakers in Washington addressing remarks intended for an entire audience, and of singers and other entertainers at Whippany, were seen by its use, simultaneously with the reproduction of their voices by loud speaking equipment.

[1] Presented at the Summer Convention of the A. I. E. E., Detroit, Mich., June 20–24, 1927.

Characteristic Problems of Television

The problem of television in its broad outlines is that of converting light signals into electrical signals, transmitting these signals to a distance, and then converting the electrical signals back into light signals. Given means for accomplishing these three essential tasks, the problem becomes that of developing these means to the requisite degree of sensitiveness, speed, efficiency, and accuracy, in order to re-create a changing scene at a distant point, without appreciable lapse of time, in a form satisfactory to the eye.

A convenient starting point for the discussion of television is the human eye itself. In this an image is formed upon the retina, a sensitive screen, consisting of a multitude of individual light-sensitive elements. Each of these elements is the termination of a nerve fibre which goes directly to the brain, the entire group of many million fibres constituting the optic nerve. A theoretically possible television system could be made by copying the eye. Thus a large number of photosensitive elements could be connected each with an individual transmission channel leading to a distant point, and signals could be sent simultaneously from each of the sensitive elements to be simultaneously used for the re-creation of the image at the distant point. The number of wires or other communication channels demanded in a television system of this sort would be impractically large. For practical purposes, reduction of the number of transmission channels is made possible by the fact that, while in vision all parts of the image on the retina are simultaneously and continuously acting to send nerve impulses, the inertia of the visual system is such that a sensation of continuity is obtained from discontinuous signals, provided these succeed each other rapidly enough. Due to the phenomenon of persistence of vision, it is immaterial to the eye whether the whole view be presented simultaneously or whether its various elements be viewed in succession, provided the entire image be traversed in a sufficiently brief interval, which for purposes of discussion may be taken as 1/16th of a second or less.[2] We thus have available in television the same artifice which is used in the much less exacting problem of transmission of pictures over a telephone line, that is, of *scanning*, or running over the elements of the image in sequence, instead of endeavoring to transmit all of the elementary signals simultaneously. The development of a television system therefore

[2] This figure of 1/16th of a second, commonly quoted in discussions of this sort, is a convenient one, although the frequency of image repetition necessary to extinguish "flicker" is actually proportional to the logarithm of the field brightness. A somewhat higher rate of image repetition was used in the final television apparatus.

necessitates, at an early stage, the design of some scanning system by which the image to be transmitted may be broken up into sequences of signals. In the simplest case, where one transmission channel is to be used, the whole image will be resolved into a single series of signals; if more than one transmission channel is to be utilized, the resolution may, by parallel scanning schemes, or their equivalent, be broken up into several series for simultaneous transmission.

Like the eye, an artificial television system must have some light-sensitive element or elements by means of which the light from the object shall produce signals of the sort which can be transmitted by the transmission system to be used. For a television system to operate over electrical transmission lines this means some photoelectric device. It is obvious that this photoelectric device must be extremely rapid in its response, since the number of elements of any image to be transmitted must be some large multiple of the fundamental image repetition frequency, that is 16 per second. The response should, of course, be proportional to the intensity of the light, and finally, the device must be sufficiently sensitive so that it will give an electrical signal of manageable size with the amount of light available through the scanning system. This latter requirement, that of sensitiveness, is one which, it was realized, from studies made with our earlier apparatus for the transmission of still pictures over wires,[3] would be extremely difficult to meet. In the picture transmission system a very intense beam of light from a small aperture is projected through a transparent film and on to a photoelectric cell. In practical television, the system must be arranged to handle light reflected from a natural object, under an illumination which would not be harmful or uncomfortable to a human being. Actual experiment showed that the greatest amount of light which could be collected from an image, formed by a large aperture photographic lens on the small scanning aperture of the picture transmission apparatus, was less by a factor of several thousand times than the light projected through it for still picture transmission purposes. Assuming the same kind of photoelectric cell to be used, the additional amplification required over that used in the picture transmission system, taking into account also the higher speed of response demanded, would bring us at once into the region where amplifier tube noise and other sources of interference would seriously affect the result. This indicated clearly that some more efficient method of gathering light from the object than the commonly assumed one of image formation by a

[3] "Transmission of Pictures over Telephone Lines," Ives, Horton, Parker and Clark. *Bell System Technical Journal*, Vol. IV, No. 2, April, 1925.

lens was required, unless some much more sensitive type of photoelectric cell should be found.

Assuming that means could be developed for producing an electrical signal proportional to the intensity of the light, of sufficient quickness to follow a rapid scanning device, and of sufficient strength either as directly delivered from a photosensitive device or as amplified, the next problem is that of its transmission over an electrical communication system. We may quickly arrive at an understanding of certain of the transmission problems by reviewing the requirements for the transmission of photographs. In the system of still picture transmission now in use by the American Telephone & Telegraph Company, a picture 5 in. by 7 in. in size, divided into the equivalent of 10,000 elements per square inch or 350,000 elements, is transmitted in approximately seven minutes. This requires the transmission of a frequency band of about 400 cycles per second on each side of the carrier frequency. If we plan, in the transmission of television, to transmit images of the same fineness of grain, it would mean that what is now transmitted in seven minutes would have to be transmitted in a 16th of a second, which in turn means that the transmission frequency range would have to be nearly 7000 times as great. That is, a band approximately 3,000,000 cycles wide would be required. Bearing in mind that wire circuits are ordinarily not designed to utilize frequencies higher than 40,000 cycles per second, and that with radio systems uniform transmission of wide signal bands becomes extremely difficult, it is seen at once that either an image of considerably less detail than that which we have been considering must suffice, or else some means for splitting up the image so that it may be sent by a large number of channels is indicated.

A further theoretical requirement must also be given consideration. This is that the complete television signal will consist of all frequencies up to the highest above discussed, and down to zero, that is, an essential part of the signal is the direct current component, furnished by those parts of the scene which do not change. The problem of handling the very low frequency components, presents difficulties both in the vacuum tube amplifier system adjacent to the photosensitive device, and in ordinarily available transmission channels.

In any case certain fundamental transmission requirements must be met. These are that the attenuation of the signals must be uniform over the whole frequency range and that the speed of transmission of all frequencies must be the same. Also, as in the transmission of sound signals, the amount of interference or noise must be kept down sufficiently not to impair the quality of the signal or picture.

Assuming the undistorted transmission of the signals to a distant point, the next fundamental problem of television is the reconstruction of the image, or the translation of the electrical signal back into light of varying intensity. Just as at the sending end we have seen that the production of a useful electrical signal with the amount of light available from a naturally illuminated object is a major problem, so at the receiving end the converse problem, that of securing an adequately bright light from the electrical signal, presents great difficulty. The nature of the problem may be understood by assuming that it is to be done by projecting the received image on a screen similar to an optical lantern projection screen. If the spot of light which is to build up this image scans the whole area in the same way that the object is scanned, we find that the amount of light which can be concentrated into a small elementary spot will, when distributed by the scanning operation over the whole screen, reduce the brightness of the screen in the ratio of the relative areas of the elementary spot and the whole screen. The amount of this reduction will, of course, depend upon the number of elements into which the picture is divided, but will in any event be a factor of several thousand times. It is doubtful whether any light source exists of sufficient intensity such that an image projected by it can be spread out by a scanning operation over a large screen and give an average screen brightness which would be at all adequate. It is possible to imagine optical systems by which such a thing as the crater of an arc could be projected upon the screen, but the motion of this image and its variation in intensity would involve the extremely rapid motion of lenses, mirrors and apertures of a size such as to render the operation mechanically impracticable. It appears from these considerations that the only promising means of reconstructing the image would be those in which a light source, whose intensity can be controlled with great rapidity, is directly viewed.

Another element of a television system upon whose solution success depends as much as any other is that of synchronization; the reconstruction of the image, postulated in the last paragraph, is only possible if the reconstructed elements fall in exactly the right positions at the right times, to correspond with the signals as generated at the analyzing end. The criterion for satisfactory synchronization will be expressed in terms of variation from identity of speed by figures which will depend on the fineness of grain of the image which it is planned to send. No element of the image must, of course, be out of place by a considerable fraction of the size of the element.

General Outline of Means Employed in the Present Television System

It has been pointed out above that if the goal which we set in television is the transmission of extended scenes, with a large amount of detail and hence made up of an exceedingly large number of elementary areas, we meet with the necessity for transmission channels of a character which are not now available. In the present development it was decided at the start to restrict our experiments to a size and grain of picture which, if the scanning and re-creating means were developed, would be capable of transmission over practical transmission channels, either wire or radio. This restriction fortunately leaves us with the possibility of meeting what was felt to be the typical problem of a Telephone Company, namely, the transmission of a human face in a television system used as an adjunct to a telephone system. Taking, as a criterion of acceptable quality, reproduction by the halftone engraving process, it is known that the human face can be satisfactorily reproduced by a 50-line screen. Assuming equal definition in both directions, 50 lines means 2500 elementary areas in all. 2500 elements transmitted in 1/16 second is 40,000 elements per second. The frequency range necessary to transmit this number of elements per second with a fidelity satisfactory for television cannot be calculated with assurance in advance. An approximate value can however be arrived at from a study of the results obtained in still picture transmission. In pictures transmitted by the system already referred to, individual faces contained in a square space ½ inch on a side are quite recognizable.[4] Taking the ratio of this area to the area of the whole picture, and using the frequency range figure already deduced, for a complete 5 in. by 7 in. picture, it appears that a band of 20,000 cycles would be sufficient to transmit such an image in 1/16 second.[5] These considerations led to the choice of a 50-line (2500-element) image as one which would be both satisfactory as to detail rendering, for our purposes, and as calling for frequency transmission requirements sufficiently low to give a good margin of safety in existing single communication channels.

As a method of scanning, the method which is probably mechanically simplest, namely, that of rotating disks with spirally arranged holes, proposed by Plotnow [6] in 1884, was chosen. In accordance with the

[4] Cf. Fig. 18 of the paper referred to (Reference 2).

[5] A factor which this analogy does not cover is that if the image is moving so that it falls on several discrete scanning elements in rapid succession a very material apparent increase in the fineness of the image structure results. This effect is similar to that by which the relatively coarse-grained individual images in a motion picture film fuse to give smooth appearing pictures.

[6] Plotnow, D. R. P. 30105, 6.1, 1884.

choice of grain above indicated, the disks were perforated with 50 apertures.

For the second element of the problem, the light-sensitive means, the alkali metal photoelectric cell was chosen as possessing the qualities of proportionality of response and quickness of reaction. The currents produced by it are at best quite small, but they lend themselves to the process of amplification by the three-electrode vacuum tube amplifier.

The problem of securing a large enough signal, which is intimately associated with that of securing enough light from the object, was, in our development work, postponed in the earlier stages, our first experimental work having been done by concentrating light through photographic transparencies.[7] The solution of the problem of securing adequate light was subsequently attained by reversing the light path and projecting a narrow beam of light through the scanning disk upon the object. By this means only the element of the object which was being scanned was illuminated at any one time, thereby reducing the average illumination enormously, and the problem of increasing the signal strength could be attacked by increasing the amount of photosensitive surface as well as by increasing the brightness of the scanning light.[8]

The problem of amplifying the photoelectric currents to sufficient value for transmission was solved by a practical compromise which at the same time met one of the transmission difficulties. This compromise consisted in amplifying and transmitting only the fluctuating or alternating current components of the signal, leaving the direct current component, which determines the general tone value of the image, for empirical reintroduction at the receiving end. By this scheme, stable amplifier constructions were made available, and the transmission channels, particularly the wire channels, could be utilized in their normal working form.

At the receiving end, the problem of securing a sufficiently bright image was solved, as indicated earlier, by the use of self-luminous surfaces of much higher intrinsic brightness than it is possible to secure by illumination of a surface by any light source which can be rapidly controlled as to its intensity. The self-luminous surfaces

[7] As one step in the development work moving picture film, projected by a commercial projector in synchronism with the scanning disks, was successfully transmitted.

[8] A still further advantage is obtained by limiting the scanning light to the region of the spectrum to which the photoelectric cells are sensitive (blue and violet). This is unnecessary where one-way transmission only is used but is of value where in two-way transmission a transmitted image is to be viewed by a person being scanned.

employed were glow lamps containing neon gas, the brightness of which changes with sufficient rapidity to follow the incoming signals.

The problem of synchronization was postponed in our earlier development work by mounting the scanning and receiving disks upon the same axle. It was later solved for the demonstration apparatus by the utilization of synchronous motors controlled by two frequencies, a low frequency, that of the image repetition period, and a high frequency, chosen of such a value that the fraction of the cycle through which transient phase displacements occurred amounted in angular displacement to less than half the angular extent of a single disk aperture. The synchronization control therefore called for the transmission of additional currents for synchronization purposes over and above the picture current.

In order to transmit and synchronize the image signals it is necessary to transmit three different frequency bands, one for the image, and two for the high and low frequency synchronization controls. In the demonstration of April 7, 1927, the images were sent in the wire demonstration over a high quality open wire line. The synchronization control was sent over two separate carrier channels of a second telephone line. In addition to these lines, another line was used for conveying the telephone conversation. In the radio demonstration two different wave-lengths were used respectively for the image signals and for the synchronization signals which were, as in the wire demonstration, carried on two different carrier frequencies. A third channel was used for the voice. In the case of both wire and radio transmission, it is quite possible to put all of the different signals upon the same transmission channel, using different carrier frequencies.

It will aid toward a clear understanding of the reasons for the success of the system of television described in the following papers if we summarize at this point the chief novel features to which that success is due. They may be listed as follows:

1. Choice of image size and structure such that the resultant signals fall within the transmission frequency range of available transmission channels.[9]
2. Scanning by means of a projected moving beam of light.
3. Transmission only of alternating current components of image.
4. Use of self-luminous surfaces of high intrinsic brilliancy for reconstruction of the image.
5. High frequency synchronization.

[9] As the succeeding papers show, the margin between the frequency range required by the scanning apparatus and that which could be made available was quite liberal. It appears in the light of our experience that apparatus with 60 or 70 scanning holes instead of 50 might be used with the transmission facilities which were at our disposal.

Applications and Future Developments

It is not easy at this early date to predict with any confidence what will be the first or the chief uses for television, or the exact lines that future development may take. It must be clearly understood that television will always be a more expensive service than telephony, for the fundamental reason that it demands many times the transmission channel capacity necessary for voice transmission. This expense will inevitably increase in proportion to the size and quality of the transmitted image.

The kinds of service which are naturally thought of upon consideration of the services now rendered in connection with sound transmission are: first, service from individual to individual, parallel in character to telephone service, and as an adjunct thereto; second, public address service, by which the face of a speaker at a distant point could be viewed by an audience while his voice was transmitted by loud speaker; third, the broadcasting of scenic events of public interest, such as athletic contests, theatrical performances and the like.

The first two types of service just mentioned lie within the range of physical practicability, with apparatus of the general type already developed. The third type, because of the uncontrolled conditions of illumination, and the much finer picture structure which would be necessary for satisfactory results, will require a very considerable advance in the sensitiveness and the efficiency of the apparatus, to say nothing of the greatly increased transmission facilities. For all three types of service, wire or radio transmission channels could be utilized, for while the problems incident to securing distortionless transmission over wide frequency bands, or multiple transmission channels, are different in detail in the two cases, they appear to be equally capable of solution by either means. However, the very serious degradation of image quality produced by the fading phenomena characteristic of radio indicates the practical restriction of radio television to fields where the much more reliable wire facilities are not available.

The Production and Utilization of Television Signals[1]

By FRANK GRAY, J. W. HORTON and R. C. MATHES

Synopsis: The design of a television system, once the fundamental principles are understood, involves a detailed consideration of the methods by which the several important functions are to be performed.

(1) In the present system the initial signal wave is obtained by sweeping a spot of light over the subject in parallel lines completely scanning it once every 18th of a second. The light reflected is collected by large photoelectric cells which control the transmitted current. At the receiving station the picture current controls the brightness of a neon lamp from which the received image is built up by means of a small aperture moving in synchronism with the spot of light at the transmitting station. For presentation to a large audience television images may be produced by a neon lamp in the form of a grid having a large number of separate electrodes. A high frequency excitation controlled by the picture current is distributed to the successive electrodes in synchronism with the spot of light at the transmitting station.

(2) Space and time variations in the reflecting power of the subject are translated into time variations in signal strength. For design purposes these time variations are represented by component frequencies, a minimum band of which must be properly transmitted to insure an adequate reproduction of the image. Within this band there must be maintained a certain degree of uniformity in the efficiency of transmission of the separate components. Also, their phases must not be permitted to shift unduly in relation to each other.

(3) The design of the terminal amplifiers is based on the quantitatively determined characteristics of the photoelectric cells and of the neon lamps as well as on the limits imposed by the transmission study and by the characteristics of available transmission media, whether telephone line or radio system. The circuits employed at the transmitting station furnish an amplification such that the power delivered to the transmission medium is 10^{15} times the power received from the photoelectric cells.

Section I. Apparatus for the Analysis and Synthesis of the Image

THE introductory paper to this series of articles on television explained principles along which any television system must operate to transmit an image over a single pair of wires or other channel of communication. As the first step in such a transmission, the space variations in brightness from point to point in the view must be translated into time variations in an electrical current that can be sent over the channel of communication. This translation may be accomplished by a scanning process that operates on the view to produce the same effect as if the view were cut up into a single long strip and passed rapidly in front of a light-sensitive cell to generate an electrical current varying with the brightness along the strip. To eliminate flicker in the reconstructed image and also to follow moving

[1] Presented at the Summer Convention of the A. I. E. E., Detroit, Mich., June 20–24, 1927.

subjects in a view, the scanning process must be repeated and a new picture transmitted at least every sixteenth of a second.

Many purely theoretical methods could be, and have been, devised to accomplish such a scanning process and to translate a view into electrical currents or signals. Unfortunately, however, a practical system of television must operate with materials and conditions as they exist, and these practical limitations constitute the serious problems of television.

The high speeds and relatively large amplitudes with which any television scanning mechanism must move, and the necessity for synchronizing the transmitting and receiving apparatus lead to the use of synchronously rotating machines as apparently the only practical solution of the scanning and receiving problems. Consequently, the

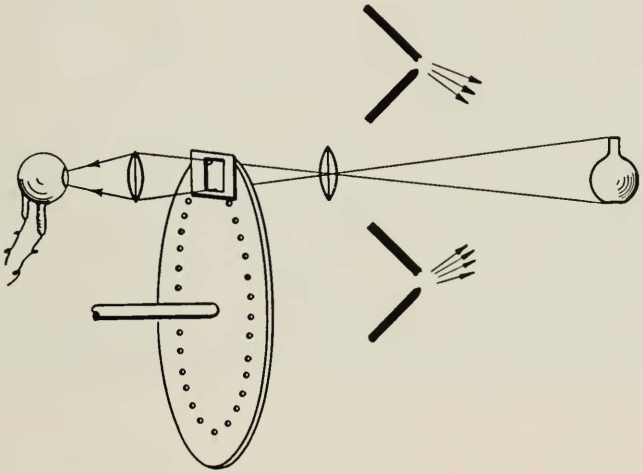


Fig. 1—Several light sources illuminate the subject; a lens forms an image which is scanned by a spiral of apertures, through which the light falls on a single photo-electric cell.

present television system has been designed to operate with continuously rotating mechanical parts.

The efficiency that must be secured in the optical part of any scanning method is fixed by the three following factors—the amount of picture detail that is to be transmitted, the efficiency of the light-sensitive cell, and the practical limit to amplifier systems. The first of these factors decides the area from which light can be collected at any one instant. In the present case this was fixed in an initial survey of the entire television problem when it was decided to confine the first attempt to the transmission of pictures as if they were made

up of 2500 small elemental areas; that is, to scan the view in a series of fifty parallel lines. The second factor is determined by the sensitivity of the potassium hydride photoelectric cell. This cell is, at the present time, the most efficient light-sensitive cell that can follow the rapid variations in light intensity without a time lag. The third factor, the limitation of amplifier systems, results from the extraneous currents that are present in metallic conductors and

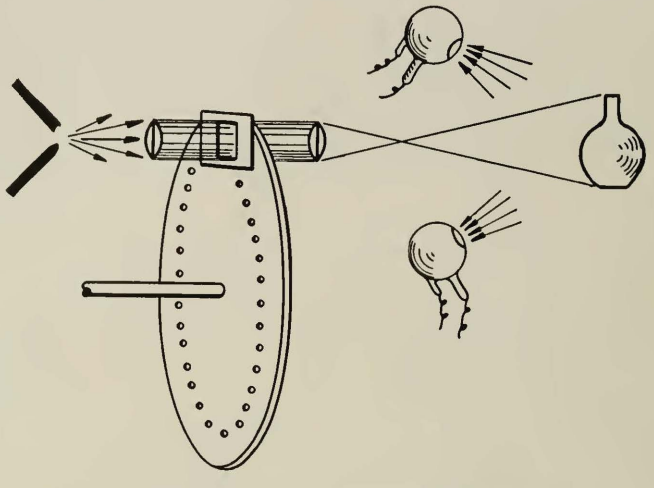


Fig. 2—Light from a single source is projected as a small moving spot on the subject; the reflected light is received by several photoelectric cells

amplifier tubes. The thermal agitation of the electrons in any input resistance generates such currents; and rapid variations in the number of electrons emitted from the hot filament of an amplifier tube also generate disturbing voltages. For successful amplification, the initial photoelectric current must be considerably larger than these extraneous currents. Consequently, the optical arrangement must be such that at any one instant it collects enough light from an elemental area of the view to generate this minimum permissible output current from the photoelectric cell.

The operation and advantages of the scanning method actually used in the present process for transmitting television images may be better understood by first considering a simple and analogous method illustrated by Fig. 1. The subject is illuminated by lights placed in front of it as shown. A lens forms an image of the subject on the rotating disk. This disk is pierced with a series of small holes or apertures arranged in the form of a spiral; and, as the disk rotates,

the apertures trace across the image one after the other in a series of parallel lines. The frame limits the size of the image and prevents more than one aperture being in the image at one time. Light, passing through an aperture as it travels across the image, falls in the light-sensitive cell and generates a picture current proportional to the brightness of the image from point to point along strips taken one after the other across the image.



Fig. 3—Illustrative transmitting apparatus. Light from the arc lamp is condensed on the disk, which is driven by a high frequency synchronous motor. The disk carries a spiral of pin hole apertures, each of which in turn projects a moving spot of light on the subject. Reflected light is collected by three large photoelectric cells.

In any system such as that outlined above, which depends upon scanning an image of the view as formed by a lens, the efficiency of the system is ultimately limited, for any given size of image that can be scanned, by the ratio of aperture to focal length of the best lens that can be secured. Experiments show that, with the best lens available to form a one-inch-square image, it would be necessary to illuminate a subject with a 16,000-candle power arc at a distance of about four feet in order to secure an image bright enough for a photoelectric cell to give an output current above the noise level in an amplifier system. In other words, television would apparently be extremely inconvenient to the subject if it were to be carried out from an image formed by a lens.

In the system actually used for television transmission, this apparent limitation has been evaded by reversing the entire optical system of Fig. 1 and arranging it as shown diagrammatically in Fig. 2. Instead of scanning an image of the subject, the actual subject is scanned directly by a rapidly moving spot of light. An illustrative laboratory set-up, Fig. 3, shows the arrangement of parts in such a transmitting station. A fifteen-inch disk rotating approximately eighteen times per second carries a series of fifty small apertures arranged in the form of a spiral. A beam of light is condensed by a lens from a 40-ampere Sperry arc to intensely illuminate a limited area in the path of the moving apertures; and a slender, intense beam of light passes through each aperture as it moves across the illuminated area. A frame in front of the disk permits light to emerge from only one aperture at a time and the lens in front of the disk focuses an image of this moving aperture on the subject. As a result of this arrangement the subject is completely scanned in a series of successive, parallel lines by a rapidly moving spot of light, once for each revolution of the disk; and on account of the transient nature of the illumination the subject is scarcely aware that he is being exposed to it.

As the spot of light traces across the subject, light is diffusely reflected or scattered from the subject in all directions, and some of the light that is reflected forward passes into three large photoelectric cells placed just in front of the person who is being viewed. The current outputs from the three photoelectric cells operate in parallel into a common amplifier system. As the beam of light passes, for instance, across a person's eyebrow less light is reflected to the photoelectric cells, and as the beam passes across his forehead more light is reflected. Since the current output from the photoelectric cells is proportional to the received light, the current follows accurately the brightness of the various elemental areas of the subject's features as he is traced over by the scanning beam. This fluctuating current is unidirectional.

The actual operation of such an optical system, its influence on the lighting effects and quality of the reproduced image, may best be understood by noting that optically the system is identically the same as if all of the rays of light were reversed in direction to give an optical system equivalent to Fig. 1. The television apparatus sees the subject exactly as if rays of light came out of the photoelectric cells to illuminate the subject; the lens formed an image of the subject on the disk; and the apparatus scanned this image and reproduced it at the receiving end. The lights and shadows seen in the image are the same as if the subject were illuminated by three large lights in

the positions of the photoelectric cells and looked at from the position of the lens. It also follows from the above considerations that, within its range of resolving power, this scanning method will not only reproduce a plane subject, such as a drawing, but that it will also faithfully reproduce three-dimensional figures with sharp edges and elevations and depressions, just as well as they could be reproduced in a photograph.

In addition, because the light passes in an approximately parallel beam through a disk aperture, the slender beams of light sweeping across the region in front of the transmitter just barely overlap each other even at a considerable distance from the apparatus. Consequently, it is not necessary that the subject be at the exact positions at which the small apertures are sharply focussed; and within wide limits no confusion results as the subject moves toward or away from the apparatus. The brightness as well as the size of the received image decreases as the subject moves away from the photoelectric cells; and for good transmission of the human features, which reflect very little blue light to which the photoelectric cells are sensitive, a person should not be more than a few feet away from the cells.

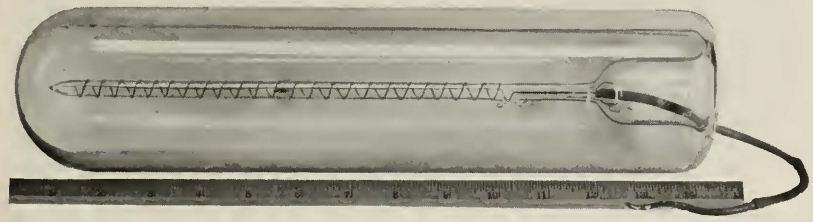


Fig. 4—Large photoelectric cell. The cell presents forty square inches of photosensitive surface to receive light reflected from a subject

This method of scanning permits two very large gains to be made in the amount of light available for producing photoelectric currents. The transient nature of the light permits a very intense illumination to be used without inconvenience to the subject. Furthermore, the optical efficiency of the system is not limited by the apertures of available lenses; but can be increased by using large photoelectric cells and more than one cell connected in parallel.

The photoelectric cells of the potassium hydride, gas-filled type used in the transmitting stations, were specially constructed for the purpose and are probably the largest photoelectric cells that have ever been made, Fig. 4. Three of these cells present an aperture of 120 square inches to collect the reflected light.

With this large collecting area and the strong light intensity that can be used for the transient illumination, the cells give an electrical output that, though still extremely small, is safely above the noise level of an amplifier system.

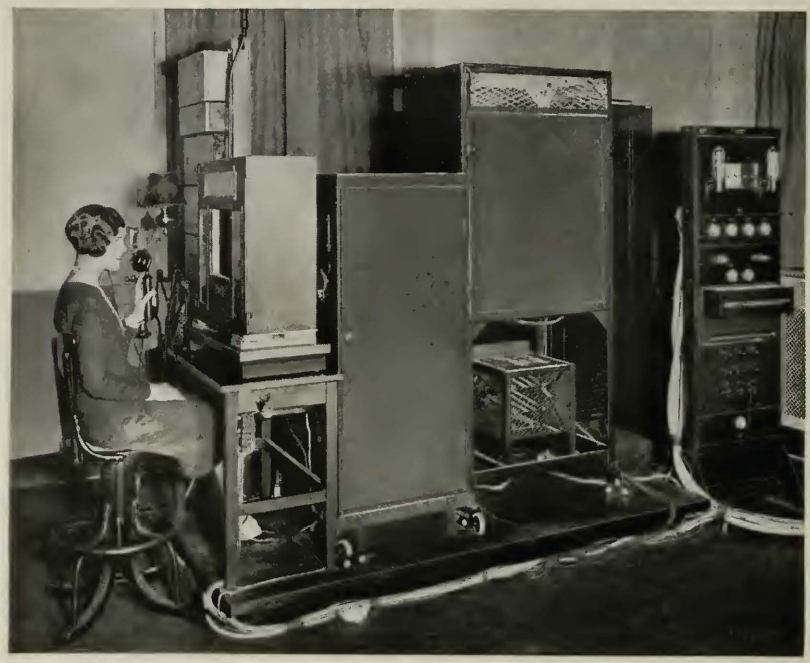


Fig. 5—Television transmitting apparatus. Sweeping beams of light pass out through the tunnel-like opening in the photoelectric cell case; light reflected from the subject is collected by three large photoelectric cells behind the screened openings.

A photograph, Fig. 5, shows the details of a television transmitting station as it is operated in the field. The arc, rotating disk and photoelectric cells are contained in separate cabinets and aligned as shown in the photographs. The three photoelectric cells and first stages of amplification are mounted in a shielded, sound-proof case. The slender, sweeping beam of light coming from the disk cabinet passes through the tunnel-like opening in the photoelectric cell case and scans the subject seated in front of it. The apparatus sees the person from light reflected back into the three large cells located just behind the screened openings in the case.

The variations of the feeble picture currents delivered from these photoelectric cells are highly amplified and transmitted over a wire or radio channel of communication by circuits described elsewhere in

this series of articles. At the receiving station this current shape is reamplified, impressed on a direct current, and finally produces an image in the receiving apparatus.



Fig. 6—Illustrative receiving apparatus. A neon lamp operated from the picture current illuminates a series of small apertures as they pass across the field of view; the observer sees an image reproduced in the frame.

A photograph, Fig. 6, shows an illustrative arrangement of the parts in one type of television receiver. An essential part of this type of receiver is a disk similar to the one at the transmitting station and also provided with fifty small apertures arranged in the form of a spiral. The driving motor rotates the disk in exact synchronism with the one at the transmitting station. The observer looks at a small rectangular opening or frame in front of the disk. This frame is of such dimensions that only one aperture can appear in the field of view at a time. As the disk rotates, the apertures pass across the frame one after the other in a series of parallel lines, each displaced a little from the preceding one until in one revolution of the disk the entire field has been covered. Beyond the disk is a special form of neon glow lamp shown in detail by Fig. 7. In this lamp, the cathode is a flat metal plate of a shape and area sufficient to entirely fill the field defined by the frame in front of the disk. The anode of the

glow lamp is a similar metal plate separated from the cathode by only a very small space (about one millimeter). At the proper gas pressure this space between the plates is within the "cathode dark space" where no discharge can pass. As a consequence, the glow-discharge develops on the outer surface of the cathode, where it shows as a perfectly uniform, thin, brightly glowing layer.

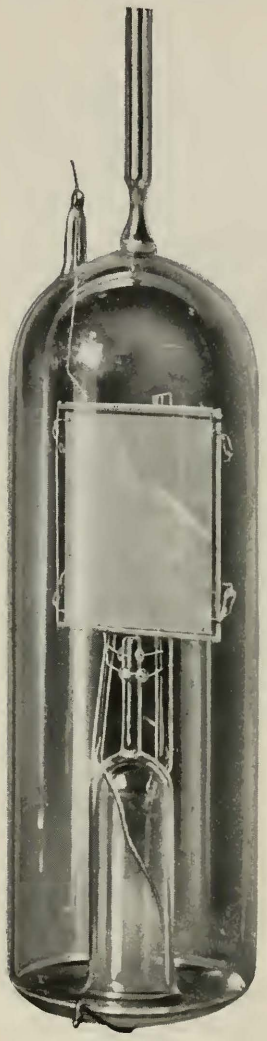


Fig. 7—Neon receiving lamp. The rectangular cathode is covered by a uniform layer of glow slightly larger than the field of view on a television disk

As an aperture in the disk moves across the field, the observer, looking through at the neon lamp behind the disk, sees the aperture as a bright point. When the disk is rotating at high speed, the observer, owing to the persistency of vision, sees a uniformly illuminated area in the frame, provided that a constant current is flowing through the lamp. (The line structure that would otherwise appear in the field is largely eliminated by using apertures that slightly overlap in their paths across the field.)

The brightness of the neon lamp is directly proportional to the current flowing through it; and when a picture is being received, the lamp is operated directly from the received picture current. As a result of the system just described, there is at any instant, in the field of view at the receiving station, a small aperture illuminated proportionally to the brightness of a corresponding spot on the distant subject. Consequently, the observer sees an image of the distant subject reproduced in the frame at the receiving station.

Fig. 8 shows the external appearance of the disk type of receiver in which the images appear. The disk rotates inside of a rectangular cabinet and the observer views the image through the shielding window. The largest disk, three feet in diameter, gives a 2 in. by 2½ in. rectangular image. Each television receiver is also equipped with a telephone receiver and transmitter; and it is possible for the observer to both see and converse with a distant person at the same time.

Considering the limited number of picture elements, a surprising amount of detail can be transmitted with this television system. A distant person can be seen and easily recognized and his motions can be plainly followed as he talks into a transmitter, turns the pages of a magazine and goes through other similar motions. Large-sized pictures in a magazine can be seen as the subject turns the pages and looks at them himself.



Fig. 8—Disk receiving apparatus. The observer looks through the shielding window at a picture on the 36-inch disk

An auxiliary television receiving system also accompanies each transmitting set and enables the operator to see that he is sending a satisfactory picture current out over the channel of communication. This auxiliary or pilot picture is formed on the scanning disk itself. A small fraction of the outgoing picture current is tapped off and amplified to operate a neon lamp, which is placed behind the disk ninety degrees around from the scanning beam. An image of the subject may thus be seen on the scanning disk just as at a receiving

station. To correct for the ninety-degree phase shift, the spiral of apertures on the transmitting disk is continued by additional apertures a quarter of a turn beyond the starting point. The first turn alone of the spiral is used for scanning; and the last turn alone, to form the pilot image; consequently, this image appears exactly in frame. A small mirror on the front of the motor cabinet reflects this image to the operator and enables him to see the character of the picture that he is sending out over the channel of communication.

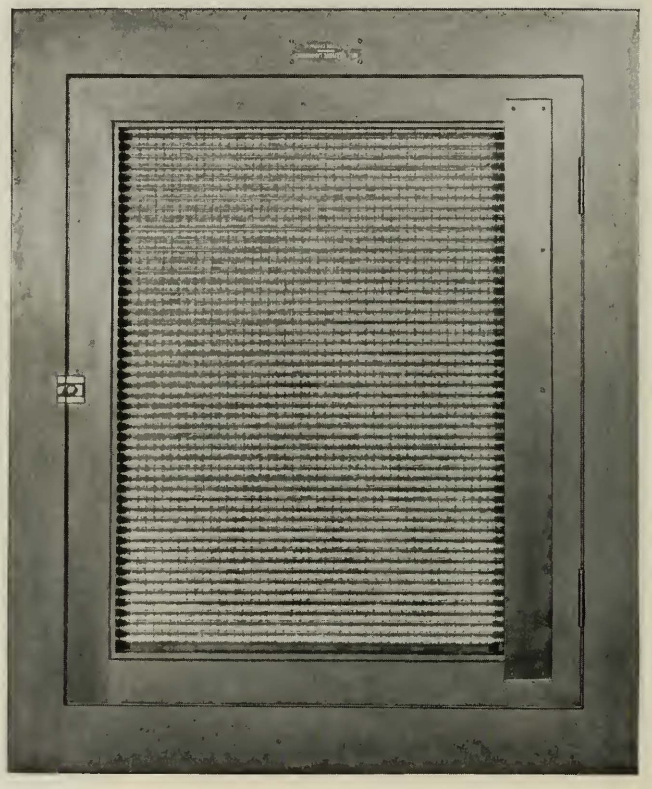


Fig. 9—Large grid. The large grid is a neon lamp with 2500 electrodes on a tube bent back and forth to form a luminous screen that is visible throughout a large auditorium.

When it is desirable to present television images to a large audience, a special grid type of receiver is used. The grid has the appearance of an illuminated screen and can be seen throughout a large auditorium. The image is not projected on the screen from a lantern like a moving picture; such optical projection would be inefficient and demand

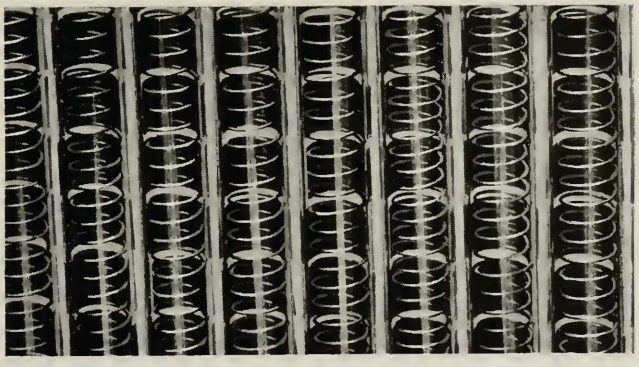


Fig. 10—Detailed structure of the grid. The exterior electrodes are pieces of metal foil cemented to the outside of the tube. The interior electrode is a long spiral of wire.



Fig. 11—Distributor and wiring. High frequency current is distributed by 2500 wires to successive electrodes of the grid from 2500 bars on a high speed distributor.

the electrical control of an impractical amount of light. The picture current itself is distributed by a commutator to successive elemental areas of a large neon lamp. This lamp, as shown in Fig. 9, consists of a single, long, neon-filled tube bent back and forth to give a series of fifty parallel sections of tubing. The tube has one interior electrode and 2500 exterior electrodes cemented along the back side of the glass tubing, Fig. 10. A high frequency voltage applied to the



Fig. 12—Details of the distributor. The bars are arranged in four rows each displaced with respect to the other three. The sliding brush is a strip of thin sheet metal.

interior electrode and any one of the exterior electrodes will cause the tube to glow in front of that particular electrode. The glow discharge actually passes to the inside wall of the glass tubing and the high frequency current flows by a capacity effect out through the glass wall to the exterior electrode. The high frequency voltage is commutated to the electrodes in succession from 2500 bars on a distributor, Fig. 11, with a brush, Fig. 12, rotating synchronously with the disk at a transmitting station. Consequently, a spot of light moves rapidly and repeatedly across the grid in a series of parallel lines one after the other and in synchronism with the scanning beam at the transmitting station. With a constant exciting voltage the

grid appears as a uniformly illuminated screen; but, when the high frequency voltage is modulated by the received picture current, an image of the distant subject is produced on the screen and his motions can be followed just as in the smaller images formed on a disk.

This method of presenting television images to a large audience permits a very efficient use of the available energy to reproduce a picture. The modulated current produces a glow discharge that exactly covers an elemental area of the picture on the screen and is viewed directly by the audience; consequently, there is absolutely no loss of energy after the picture current has been converted into light. In addition, each illuminated area of the screen responds to the picture current in the same manner; the exterior electrodes are exactly alike, and the use of a single tube assures the same pressure and purity of neon throughout the grid.

Fig. 9 shows such a screen set up for demonstration in an auditorium. A loud speaker is mounted just below the screen and it is thus possible for a large audience to both see and listen to a distant person at the same time.

Section II. The Television Signal Wave

So far it has been assumed that the electrical signal wave is perfectly transmitted between the conversion devices which transform the light variations into electrical variations and back again. Perfect transmission is, however, impossible with practical apparatus. There are certain requirements placed upon the generated signal wave by the characteristics of practical communication channels, and reciprocally certain demands are made upon a transmission system by the inherent nature of an adequate television signal. In addition to exploring these mutual requirements experimentally it is desirable to analyze them in such a way that, as far as possible, quantitative expression may be given to them. This expression in the case of the signal wave is best made by the methods of the Fourier analysis; considering the signal as made up of many sine wave components of various frequencies. The requirement on the signal may then be described in terms of these components and the requirements on the connecting transmission system in terms of attenuation and phase characteristics over a band of frequencies. These requirements will now be discussed as a basis for the subject matter of the succeeding section of this paper and of the following companion papers of this group on "Wire Transmission Systems for Television" and "Radio Transmission Systems for Television."

The problems to be discussed may be conveniently considered under three headings:

(*a*) The Character of the Television Signal.

(*b*) Requirements upon the Signal Wave Set by the Characteristics of Available Transmission Channels.

(*c*) Requirements which the Transmission Channels must meet in order to carry Television Signals.

(*a*) *The Character of the Television Signal.* As we have seen, the voltage produced across a resistance in series with the photoelectric cell is a fluctuating unidirectional potential. The generated signal therefore has frequency components beginning at and including zero frequency. The value of the voltage at any instant is roughly proportional to the average reflected illumination at that instant from an illuminated spot whose size depends upon the apertures in the scanning disk. At any point where there is a sudden change in the tone value of the subject there will also be a sharp change in the generated voltage. It will, therefore, be seen that but for the limits of speed of action of the photoelectric cell and its connected circuits the generated signals would tend to include components over the whole frequency range up to infinity. Since it is possible to effectively transmit but a limited range of these components, the width and location of the frequency band necessary for the acceptable reproduction of a given size and structure of image must be determined. It is convenient to consider first the low frequency end of the band.

In the early experimental work it was soon found that in attempting to amplify the lower frequencies by the use of direct current amplifiers, unstable conditions of operation were reached before sufficient amplification was obtained to operate the receiving apparatus. Experiments were then made with resistance-condenser coupled amplifiers which showed that, if the efficiency of such an amplifier at the frequency equal to the number of pictures sent per second was not more than about two T U below its average efficiency for the transmitted range, acceptable reproduction of the picture was secured together with stable operation of the amplifiers. When the low frequency cut-off of the amplifier was set much above this, spurious shadows were introduced into the picture. That there will be a critical lower frequency for the transmission of an unchanging scene is obvious since the Fourier series into which the signal may be analyzed starts with a constant term and the sine wave terms begin with the picture frequency and include a vast number of its harmonics. If the constant component (d-c.) is removed, the lowest frequency which remains to be transmitted is therefore the picture frequency.

The effect of removing the d-c. component of the signal can be qualitatively traced in a simple manner. Imagine three types of still

pictures or scenes to be transmitted by the system. Let the first be quite dark in general effect and require fluctuations in the signal current of a certain average amount for its transmission. Such a picture would have a low direct current component. Let the second picture consist largely of medium grays and require about the same fluctuations in signal intensity for its delineation. Such a picture will have a medium direct current component. Let the third picture be very light in general effect with such difference in light and shadow as would require the same fluctuations in signal intensity as the other two pictures. Such a picture would have a relatively high direct current component. In passing through a resistance-condenser coupled amplifier, the signals for all three types of pictures would be changed from fluctuations superimposed upon direct current to alternating currents, all of about the same average value.

At the receiving end of the circuit the direct current component may be reinserted by superimposing the alternating current fluctuations upon a fixed value of direct current such as the steady state current in the last amplifier tube. This direct current component would give the best average results if it corresponded to that suitable for the gray picture, which would, of course, then be most nearly correctly reproduced. However, most of the detail of the dark and light scenes would also be reproduced though the tone values would be distributed about a medium gray. Fortunately a change in character of this kind has proven for the most part unimportant. Where it is important it can be taken care of very simply by providing, at the receiving end, means, either manual or automatic, for changing, in accordance with the type of scene being transmitted, the magnitude of the unidirectional current upon which the received alternating current is superposed, which amounts simply to the restoration of the direct current.

In the case of scenes which are changing, however, frequencies lower than picture frequency will in general be generated and their suppression may be expected to affect to some degree the perfection of the picture. In effect, these frequencies are analogous to changes in tone values in the case of still pictures and their elimination results in fluctuations in the apparent brightness of the image. This effect is not disturbing with many types of subjects, as for example in the reproduction of the face.

One remarkable result of not transmitting the direct current component of the signal in the case of the reflected beam method of scanning is that the television transmitting apparatus can be located and operated in a well-lighted room, for if this general illumination is

constant it simply increases the direct current component of the signal. Similarly if the scene itself contains a source of steady light, this will be visible only in so far as it reflects the scanning beam.

Turning now to the upper part of the frequency range, experimental data on the highest necessary components were obtained by the use of circuits with low pass instead of high pass characteristics. With the television terminal apparatus operating at 17.7 pictures per second, it was found that a filter whose phase distortion had been corrected over practically all of its pass band of 15,000 cycles produced a degradation in image quality which was just detectable when the human countenance was being transmitted. Since the electrical terminal apparatus without the filters would efficiently transmit frequencies higher than this, the experiment showed either that frequencies higher than this were not present in the generated signal, that they were not effectively reproduced, or that they contribute little to the appearance of the image. This upper limit to the useful frequency range for this apparatus is rather lower than was anticipated from the initial survey, but because of psychological factors (decreased discrimination of tone values for fine details, apparent improved resolution when the subject is moving, etc.) it proves satisfactory for television purposes.

It is of importance, however, to know where the limitation in frequency range occurs in the apparatus and how it might be modified. Considerable information on this point is obtained by studying the nature of the distortion introduced by the aperture in the optical system and that introduced by frequency limitation in the electrical part of the system. It is convenient to consider them together as the type of distortion turns out to be similar for the two cases. This distortion may be considered most simply in relation to the type of signal corresponding to a sudden unit change in tone value at some point in the subject. With an ideal television system in which the instantaneous values of signal current are at all times proportional to the tone values of the points being scanned, the resulting signal would be represented by the graph of Fig. 13. Such a consideration involves no real loss in generality as any signal shape may be considered as the result of infinitesimal abrupt changes in intensity.

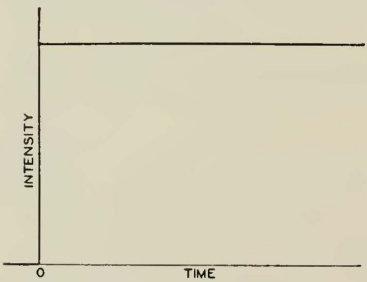


Fig. 13—Elementary signal change

It is readily seen that if a square aperture passes with uniform velocity over a part of the picture having an abrupt change from dark to light the result is that we get a signal from the photoelectric cell which, instead of building up instantaneously, builds up linearly during a time, T, Fig. 14, which is the time required for the aperture to pass a given point.[1] The net effect is an apparent sluggishness in the response of the system. The dotted curve of Fig. 14 shows the integrated illumination passing through a circular aperture of a diameter corresponding to the same time, T, for the condition of Fig. 13. Due to the simpler analysis the discussion will be carried out in terms of the square aperture though the sluggishness due to the round one is seen to be slightly less.

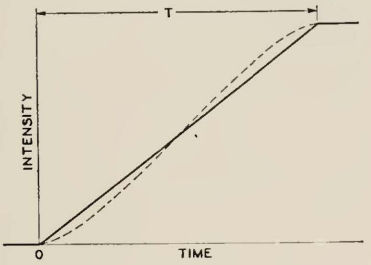


Fig. 14—Elementary signal change as distorted by a square aperture

Now this kind of sluggishness in response is quite similar to that introduced in the electrical part of the system when the upper frequencies are cut out or not transmitted as efficiently as the lower ones. The effect of frequency limitation can be investigated theoretically in a fairly simple fashion if we make the ideal assumption that all frequencies are transmitted without distortion up to a cut-off frequency, f_c, and extinguished beyond it. In Appendix I, it is shown how the signal of Fig. 14 is affected by a frequency limitation of this type. We can then plot a set of curves as shown on Fig. 15

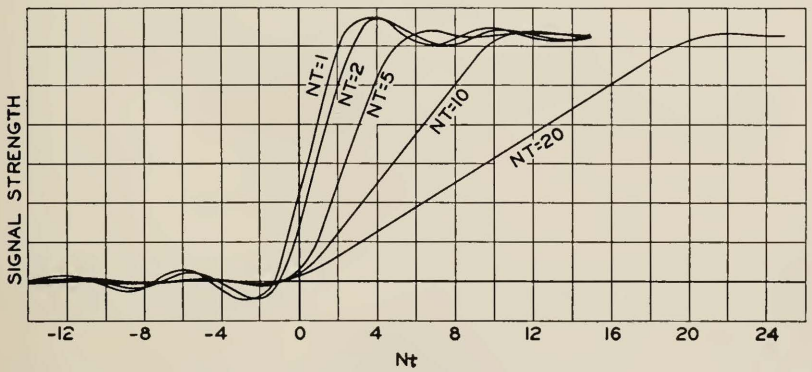


Fig. 15—Elementary signal change as distorted by a square aperture and by ideal frequency restriction

[1] This effect of aperture distortion was pointed out in the paper "Transmission of Pictures over Telephone Lines" by Ives, Horton, Parker and Clark, *B. S. T. J.*, April, 1925.

from which we can measure the total time of rise due to both the aperture and frequency limitation. The abscissa is the product of $N = 2\pi f_c$ and the time, t. Any one curve serves for a wide range of values of N and T as long as their product is the same. Call the new time of rise τ. Then we can plot a relation as on Fig. 16 between

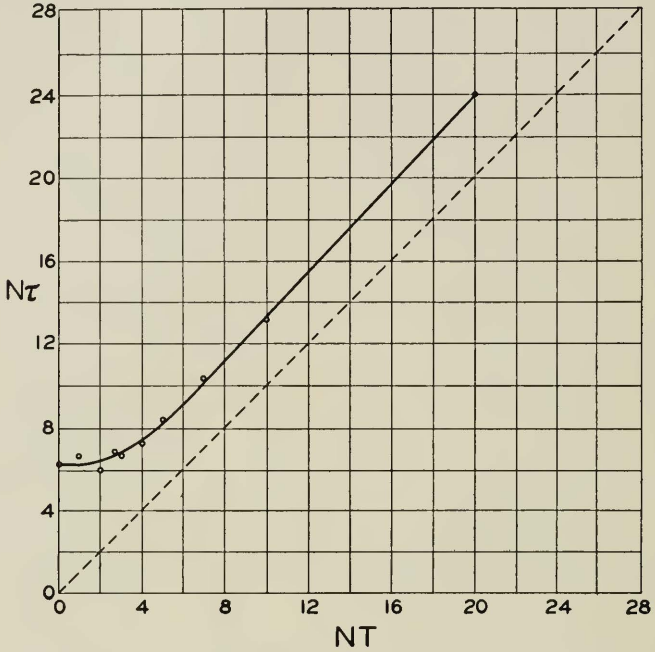


Fig. 16—Sluggishness due to distortion as a function of the aperture width and frequency restriction

$N\tau$ and NT from which we can draw conclusions as to the relative effects of aperture and frequency distortion.

Below the knee of this curve we have approximately

$$N\tau = 2\pi$$

$$\tau = \frac{1}{f_c}$$

and the frequency cut-off determines the whole distortion. Similarly above the knee

$$N\tau = NT + \pi$$

$$\tau = T + \frac{1}{2f_c}$$

and the controlling influence is that of the aperture.

Unless one effect is much more easily remedied than the other, the knee of the curve appears a reasonable point to select for operation. At the knee $NT_k = 2\pi f_c T_k = \pi$ and $T_k = 1/2f_c$. At this point the total lag is not much greater than that due to the frequency restriction alone and is $1/f_c$ or twice T_k. That is, at this point, the additional lag in the time of rise of signal due to the restricted frequency range is equal to that due originally to the aperture, though the additional lag due to the aperture is not much greater than that due to the frequency restriction alone. For a square aperture in a square picture of 2500 elements sent 16 times a second $T = 1/40{,}000$ of a second, and $f_c = 20{,}000$ cycles at the knee of the curve. The point on the curve where the effect of frequency restriction introduces a sluggishness in following light changes comparable to that introduced by a square aperture is the same frequency as that arrived at as the upper limit to useful frequencies by considerations from still picture transmission, in the introductory paper by Mr. Ives. Its value is equal to one half the number of picture elements.

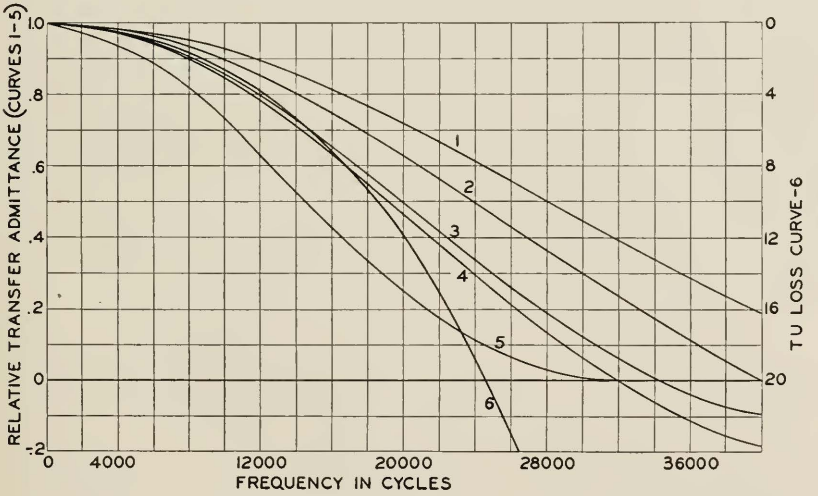


Fig. 17—Equivalent transfer admittance of various apertures

It has furthermore been found possible to determine ideal electrical transmission characteristics or equivalent transfer admittances of circuits which produce exactly the same distortions as various types of apertures. While it appears impossible at present to construct a physical circuit which will produce such characteristics over the whole frequency range, the problem is not difficult if we limit ourselves to the most important frequency band. This is of interest as it points out

the possibility of compensating for the effect of the aperture by putting in an electrical network with frequency transmission characteristics the inverse of those so determined. Within the range of important frequencies it turns out that the effect of the aperture is the same as that of a network which changes merely the relative amplitudes of the frequencies into which the picture signal may be analyzed. Neglecting constant multiplying factors, the relative variation over the frequency range for a square aperture is given by the factor $\frac{\sin T\omega/2}{\omega}$ and for a round aperture by $\frac{J_1(T\omega/2)}{\omega}$, where, as before, T is the maximum time for the aperture to pass a given point and J_1 is the Bessel's function of the first order. The derivation of these factors is given in Appendix II. On Fig. 17, Curve 1 gives the relative values of the equivalent transfer admittance for the square aperture and Curve 2 for an inscribed circular aperture, both in case of a 50-line scanned picture which is square and sent 16 times per second. T then is equal to 1/40,000 sec.

In the system as set up for demonstration the image is rectangular with the vertical and horizontal dimensions in about the ratio 5 to 4. The circular aperture is about 1¼ times 1/50 of the vertical height and the scanning is done 17.7 times a second. T is then 3.53×10^{-5} seconds and Curve 3 gives the corresponding frequency characteristics. Curve 4 shows that a square aperture of the same area as the circular aperture for Curve 3 gives a fairly good approximation to Curve 3. Curve 5 gives the combined effect of the two circular apertures, sending and receiving, corresponding to Curve 3. Curve 6 is Curve 5 plotted in terms of T U on the right hand scale.

An inspection of this last curve indicates that this frequency attenuation characteristic of the aperture introduces a considerable loss at 15,000 cycles and leaves little of the signal components above 20,000 cycles. To see if an electrical circuit of characteristics inverse to those of the aperture would materially improve the resolution of the image, the circuit,[1] which, together with its frequency characteristics, is shown in Fig. 18, was inserted between the sending and receiving amplifiers. It was designed to compensate for most of the aperture distortion and its phase distortion was made small below 20,000 cycles. On the fan-shaped test pattern of Fig. 19 a noticeable improvement was observed, the black and white angles being resolved closer to the tip of the pattern. In the case of faces the improvement appeared to be very little but could be detected

[1] This is a constant resistance type of corrective network or equalizer. See Chap. XVIII, "Transmission Circuits for Telephonic Communication," K. S. Johnson.

in the slightly better definition of sharp narrow lines such as the frames of horn-rimmed spectacles. When a system of considerable attenuation is employed between the sending and receiving terminals, it would in general be preferable to split the equalizing between the sending and receiving ends to make the best use of the sending end power in riding over interference.

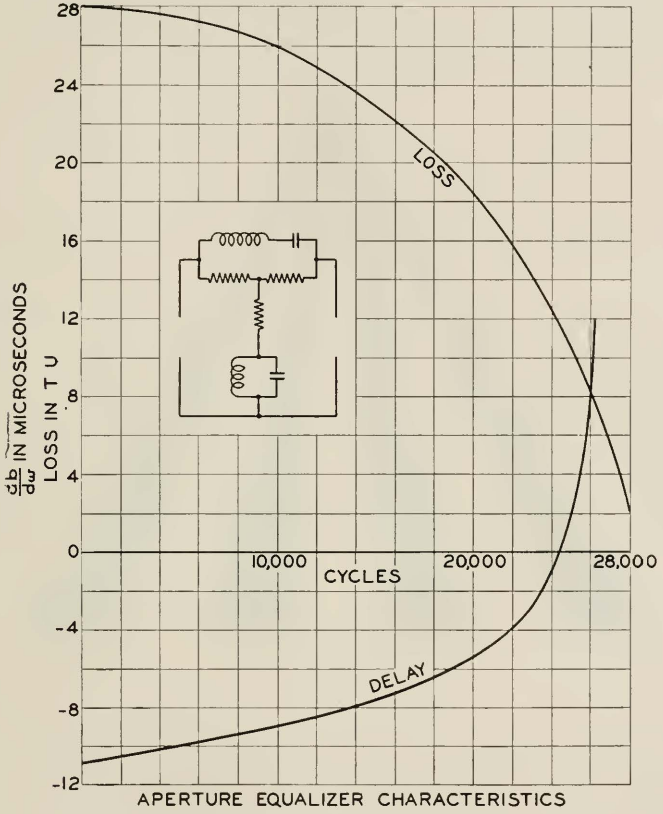


Fig. 18—Circuit for equalizing the aperture effect and its amplitude and phase characteristics

In arriving at the amount of electrical equalization which shall be adopted in any particular case it must of course be borne in mind that as the aperture is made narrower the amount of distortion introduced by it becomes less. As we narrow the aperture, however, the available illumination becomes less and the signal generated by the photoelectric cell becomes smaller. A limit is therefore soon reached at which the difficulties of amplification become greater than the

difficulties of equalization and a minimum practical aperture width is thereby determined. If the distortion is corrected by narrowing the aperture, it is apparent that the apparatus will generate, at but little lower than the correct relative efficiency, frequencies much higher than those thought necessary from the more general considerations in Mr. Ives' introductory paper. Decision as to the desirable frequency transmission band for the connecting communication channel would be no different for either method of reducing the distortion due to the aperture.

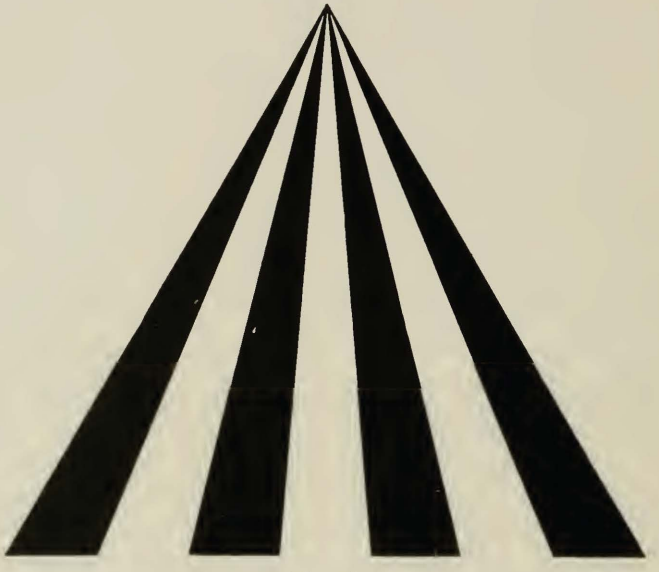


Fig. 19

In summary, then, we may say that experiment and theory show that the lowest frequency essential to satisfactory results is the picture frequency, and the highest frequency required is approximately one half the number of picture elements scanned per second.

(*b*) *Requirements upon the Signal Wave Set by the Characteristics of Available Transmission Channels.* The limitations upon the signal wave set by present available communication channels are:

1. The magnitude of the signal necessary to override the interference to which such channels are subject.

2. The frequency range which such channels can transmit.

The first of these is self-explanatory. It determines the required amplification and load capacity of the transmitting apparatus. In the companion paper on *Wire Transmission Systems for Television* are

the data on interference and on permissible signal to noise ratio which were used in the design of the terminal transmitting amplifiers to be described in the latter part of this paper.

In considering the frequency range of lines, it was apparent in the beginning that the wire channel might include sections of cable. With existing loading systems for such cables a frequency range of not over 40,000 cycles appeared available. The terminal apparatus was therefore designed to deliver a generated signal whose essential components lay well within this limit, and the laboratory tests mentioned in the preceding section showed that this requirement was met.

A lower frequency limit was imposed by the necessity of a transformer for joining the transmission line to the terminal equipment. Fortunately it proved possible to design transformers as described in the final part of this paper in which this limit was at or below the essential low frequency limit found in the preceding discussion of the signal wave.

(*c*) *Requirements Which Transmission Channels Must Meet in Order to Carry Television Signals.* We have shown that a certain band width of frequency components is essential to the adequate reproduction of the image. This sets the frequency limits of the transmission channel which must be provided. It is essential, however, that within these transmission limits the channel should present a reasonably uniform attenuation, and that the phase relations should be fairly accurately maintained. The problem as presented to the transmission engineers of wire, radio and terminal equipment for the recent demonstration was to meet the following requirements:

First, transmission must be provided for frequencies between about 10 cycles and 20,000 cycles.

Second, the amplitude frequency characteristics within this range should be uniform to about $\pm$ 2 T U.

Third, the phase shift through the range should be maintained so that the slope of its characteristic as a function of frequency is constant to $\pm$ 10 or 20 micro-seconds over all but the lowest part of the frequency range. There, about 50 times this limit was considered the maximum permissible.

These requirements were arrived at by considerations based on theory and experiments on television and analogy to similar requirements in telephotography. The first requirement follows directly from the discussion of the essential frequencies in the signal. The following paragraphs are intended to illustrate the significance of the remaining requirements.

As we have as yet no quantitative measure of the goodness of

reproduction of the image, the matter of the second and third transmission requirements on received amplitude and phase characteristics over the frequency scale is one which had to be decided largely on the basis of the experimental results and judgment based on general considerations. We have already seen that the removal of the very lowest frequencies simply changes the tone value of the whole picture. It may be similarly reasoned that departures from the average efficiency of transmission in the lower part of the frequency range would result in the appearance of diffuse shadows or high lights. Likewise, it may be concluded that broad deviations from the average efficiency of transmission in the uppermost part of the signal frequency range would result in the accentuation or the fading out of the finer detail of the scene. Steep slopes in the amplitude-frequency curve would result in the superposition of oscillations upon signals representing sudden changes in intensity. To reduce these effects every reasonable effort was made to keep the variations in the amplitude characteristic with frequency as slight as possible, aiming to hold these characteristics for the separate parts of the demonstration system to within $\pm$ 2 T U or better.

In addition to transmitting the component frequencies with the same relative efficiency as regards amplitude, it is also particularly essential in television to send them through the system with small relative phase shifts; that is, with constant velocity or what is equivalent, a phase shift proportional to frequency. It has long been known in optical theory that the envelope of a group of waves of nearly the same wave-length and nearly the same frequency may travel along with a "group velocity" somewhat different from the phase velocities of the component elements. If the system has but small departures from a flat amplitude-frequency characteristic and from a linear phase shift frequency characteristic, it can be shown that the time of group transmission or "envelope delay" is given by $db/d\omega^2$, the slope of the curve obtained by plotting the phase shift, b, for the system, against the angular velocity, $\omega = 2\pi f$. The time of transmission of a crest for any sine wave component of frequency $\omega/2\pi$ is, of course, given by b/ω. If $b = c\omega$, $b/\omega = c$ and $db/d\omega = c$. Then the phase and envelope times of transmission are equal and all frequencies as well as their group envelopes get over in the same time. If b is given in radians, $db/d\omega$ is given in seconds. In general a knowledge of b as a function of ω is necessary and sufficient to determine the phase distortion. A knowledge of $db/d\omega$ as a function of ω is not sufficient to determine all factors in signal distortion. It is, however, often easier to measure with the needed accuracy and in transmission

systems such as have been used for still pictures and television has proven a useful index of phase characteristics.

After a preliminary estimate from experience with still pictures that the limit on $db/d\omega$ should be $\pm$ 10 microseconds, an electrical network consisting of five sections of a simple lattice structure was used for testing the effect of phase distortion with television apparatus. This network introduced negligible amplitude distortion and a drift in the value of $db/d\omega$ of 50 microseconds over the frequency range of 0 to 20,000 cycles. Its effect was perceptible in blurring the image of a face and it decidedly affected a sharp pattern of two parallel lines of such width and spacing as to be just within the resolving power of the apparatus. This variation of $db/d\omega$ was about 2½ times greater than that postulated. Hence $\pm$ 10 microseconds was agreed on as a desirable limit for $db/d\omega$, though it was felt that this limit might be exceeded by a factor of two in restricted parts of the frequency band.

When this network was combined with a filter the slope of whose envelope delay curve was in the opposite direction so that over the greater part of the frequency range the combined delay of the two

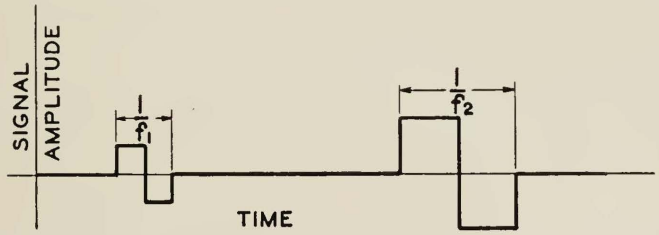


Fig. 20—Signal details of concentrated frequency spectrum for illustrating the effect of envelope delay

circuits was constant and equal to 140 microseconds, this time delay effect was very graphically brought out. Every time the combined circuit was cut in, the undistorted received image jumped to a new position a little over 10 per cent of the width of the picture to one side in the direction of scanning.

To see why $db/d\omega$ should be maintained at a constant value, consider two sharply defined details near together in the picture which would produce a variation in signal intensity with time as indicated in Fig. 20. Imagine each to be cyclically continued so that the small detail defines a frequency f_1 and the other defines a frequency f_2. It is then known from Fourier analysis that the frequency spectra of the two details are chiefly concentrated around the frequencies f_1 and f_2. If $db/d\omega$ is appreciably different at the frequencies f_1 and f_2 for any part of the

system, the two details will be displaced relatively to each other along the line of scanning and, in most cases, if this shift is appreciable, some change in the shape of the signal wave defining each detail results with further increase in the distortion. The same relative shift would occur if the narrow detail were located upon the broader one, in which case such a shift would be more apparent. It would seem reasonable to expect then that differences in the envelope time of transmission comparable to a whole picture element (about 28 microseconds in the demonstration apparatus) would be noticeable.

In most images very few details will have signal shapes, as in this special case, in which the frequency components are concentrated in narrow frequency bands. An abrupt change in signal strength, for instance, is represented by components distributed over the whole frequency range. We can imagine these frequencies divided into any arbitrary number of groups, each of which determines a wave form. When these wave forms are added together, they will reproduce the original abrupt change in signal strength. If, however, they are sent through a system in which the envelope delays for the different groups are unequal, the individual wave forms will be relatively displaced and will no longer combine correctly. As a result the image is blurred. For some types of phase distortion the effect appears as an oscillatory transient following sudden changes in intensity.

It was furthermore found by experiment that the limit of ± 10 microseconds was not necessary for the lower frequencies. Reference to the delay characteristics of the transformers described in the latter part of this paper shows that in the lower part of the frequency scale deviations from the nearly uniform value of delay at the upper frequencies appear of magnitude greater than 100 microseconds. When the signal was sent through these transformers, however, there was no observable distortion of the image. The requirements are therefore much more lenient at the low frequencies.

In the terminal apparatus the problem of meeting the above outlined phase transmission requirements was not a very serious one. The circuits involved are such that when a flat amplitude-frequency characteristic had been secured the phase distortion was also negligible.

Section III. Terminal Circuits for Sending and Receiving Television Signals

The preceding sections have discussed the methods by which an object, the image of which is to be transmitted, is made to control the time variations in a light, thus giving a luminous signal wave, and the means by which the image may be reconstructed with the aid of an

electric signal wave corresponding to this initial luminous wave in its relative instantaneous amplitudes. Certain important relations between the characteristics of the signal wave and the resulting image have been pointed out. There remains the question of obtaining an electric signal wave suitable for long distance transmission and of providing for the control of the illumination at the receiving terminal by the electric signal wave as delivered by the transmission medium.

In the use of wire lines for television it is fortunately true that a suitably prepared open-wire circuit possesses a frequency range sufficient for the transmission of all the essential components of the signal wave. Details regarding the characteristics of the wire circuits are given in a companion paper by Messrs. Gannett and Green, from whose work are obtained data essential to the design of the terminal equipment. These data fix the power level at which the signal should be delivered to the line and the power level which will be available at the receiving end. When the transmission is by radio it is, of course, necessary to effect a frequency translation in order to secure a wave suitable for radiation and transmission through the ether. In this case, however, the radio system, which is described in a paper by Mr. E. L. Nelson, when considered as a whole may be conveniently taken as a system capable of the transmission of a signal wave occupying the same frequency range as that supplied to the wire circuits. In fact the design of the radio system is such that it may be used interchangeably with the wire line in so far as the remaining electrical terminal equipment is concerned.

The terminal circuits, then, fall into two groups: first, those used at the transmitting terminal for building up the wave controlled by the time variations in light to the power level required by the line; and second, those used at the receiving terminal to bring the wave delivered by the line to the proper form for controlling the luminous sources from which the received picture is built up.

Transmitting Circuits

Starting with the photoelectric cell in which the initial luminous signal wave is converted to an electric signal wave, we are interested in the magnitude of various pertinent constants. The cell may be considered for our purposes as an impedance, the value of which is determined by the quantity of light reaching it. With no illumination at all this impedance is almost entirely a capacitance of the order of 10 m.m.f. When the cell is illuminated this capacitance becomes effectively shunted by a very small conductance which is roughly proportional to the square of the voltage between the electrodes.

For a fixed potential the magnitude of this conductance is nearly a linear function of the illumination. With a suitable potential in series with the cell, then, there is obtained a current the amplitude of which is proportional to the quantity of light reaching the cell.

In order to connect the photoelectric cell to the amplifier, there is introduced in series with the cell and its polarizing battery a pure resistance the voltage drop across which is used to control the grid potential of the first tube. It is desirable, of course, to make this resistance high in order to have available as much voltage as possible. Its value is, however, limited by two considerations. The added series conductance must not be so low that it appreciably disturbs the linear relation between the illumination and the total conductance of the circuit. The voltage drop must also be so small, in comparison with the total potential in the circuit, that the photoelectric cell operates at an approximately constant polarizing potential.

In view of the extremely small voltage of the electric signal wave as delivered by the photoelectric cell circuit, it is essential that great care be taken to prevent such interference as may enter the initial amplifier stages from approaching a comparable magnitude. The most troublesome sources of interference are electrostatic induction, electromagnetic induction, mechanical vibration, and acoustic vibration. By mechanical vibration is meant disturbances transmitted through the supports as the result of building vibrations and similar phenomena. By acoustic vibrations are meant impulses transmitted through the air which strike the several elements of the amplifier and cause motion which results in variations in their electrical constants. Electrical disturbances are reduced to a minimum by placing the amplifier as close as possible to the photoelectric cells, thereby keeping the leads short, which avoids electrostatic pick-up and also prevents the formation of closed loops of any appreciable size, thus avoiding electromagnetic induction. The amplifier is provided with a very complete electrical shield and both the shielded amplifier and the photoelectric cells are placed in a carefully shielded cabinet.

The tubes used, namely, the so-called "peanut" tubes, are, under ordinary conditions, remarkably free from any microphonic action. At the very low signal levels used, however, certain extra precautions have to be taken against this effect. In addition to lining the amplifier box with sound-absorbing material, the tubes themselves have been wrapped in felt and placed within a heavy lead case. This prevents such acoustic disturbances as reach the interior of the amplifier container from having any noticeable effect on the tube. The lead container is supported entirely by an elastic suspension and thus

serves a dual function, as the heavy mass, supported in this way, is capable of little response to such mechanical vibrations as may be transmitted through the cabinet and the walls of the amplifier shield. With these precautions it has been found possible to make the effect of all external disturbances of about the magnitude of the thermal disturbances referred to in the first part of the paper.

A schematic diagram of the amplifier tubes directly associated with the photoelectric cell is given in Fig. 21. Attention has already been

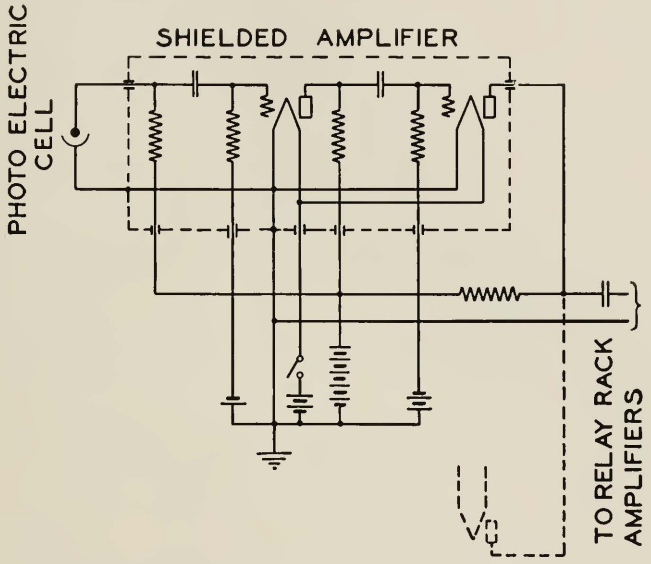


Fig. 21—Schematic of vacuum tube amplifier used with photoelectric cells

called to the fact that the initial signal, that is, the time variation of the light reflected from the scanned object, contains a direct current component. The amplification of this direct current component is, as has been stated, out of the question in any amplifier intended for continued operation over long periods of time. The requirements as to the range of frequencies to be transmitted, as discussed in the preceding section, make it necessary to provide a circuit having practically uniform efficiency from 10 cycles to above 20 kilocycles. The relative phase shift of the several components must also be kept very small. In view of the large amplification and consequent large number of stages necessary, it has been thought impracticable to use transformer coupling between all stages as the aggregate frequency and phase distortion might well be greater than could be tolerated. The so-called resistance capacitance coupling has therefore been used.

The arrangement of the several photoelectric cells in their cabinet, as shown in Fig. 3, is such that one amplifier can be connected directly to two of the cells leaving the third to operate a second amplifier. The outputs of these two amplifiers are then connected in parallel to the common battery supply equipment shown at the bottom of the two vertical cells.

By the use of two stages of amplification in the photoelectric cell amplifier, the signal is brought to such a level that it may be carried by suitably shielded leads to other amplifiers outside the photoelectric cell cabinet. This permits of using the convenient relay rack form of mounting. The signal level is, however, still low and may be adequately handled in amplifier units which differ but little from those used with the photoelectric cell.

The remaining requirements placed on the amplifiers at the transmitting terminal are those set by the telephone line. One of primary importance is that which determines the amount of energy needed. In order that the signal wave shall be of such magnitude that any interference present in the line may be negligible in comparison, it is desired that the alternating current delivered by the final amplifier stage shall be at least 4 milliamperes into an impedance of 600 ohms. The energy to be supplied is, therefore, approximately 0.01 watt, which determines the choice of the last amplifier stage. To build up the signal to a value sufficient to operate this output tube it has been found that eight stages of the small-sized tubes and one stage of greater load-carrying capacity must be used. The total amplification given by these ten stages is approximately 130 T U. It is through this known gain of the amplifiers that we get our only accurate quantitative data as to the magnitude of the initial signal wave. This comes out to be about 10^{-15} watts or, with a 100,000-ohm resistance in series with the photoelectric cell, the potential available at the first tube is roughly 10 microvolts.

The characteristics of the line also determine the means by which it shall be coupled to the final amplifier stage. In order to secure the proper impedance matching and to prevent the line from being unbalanced with respect to ground, it was felt desirable to use transformers if possible rather than to attempt the design of a tube circuit capable of meeting the requirements directly. The problem included both output and input transformers, and specified an amplitude-frequency characteristic constant to within $\pm$ 0.5 T U from 10 cycles to 25,000 cycles. The input coils intended for use at the receiving terminal had the additional requirement that a minimum of interference current should be induced in the secondary due to potentials

between the line and ground. The success with which this problem has been solved is shown by the curves of Fig. 22. Curve 1 is the

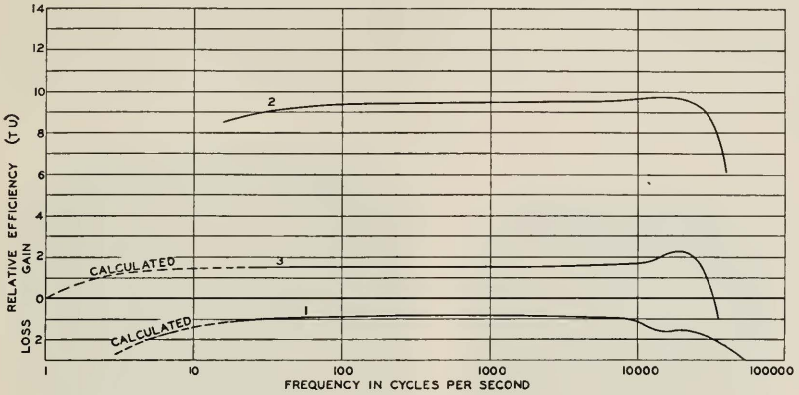


Fig. 22—Transmission characteristics of iron core transformers

1. Output transformer connected between impedances of 2000 ohms and 600 ohms.
2. Input transformer having voltage step-up of 6.5 connected between 600-ohm line and vacuum tube.
3. Input transformer having voltage step-up of 2.5 connected between 600-ohm line and vacuum tube.

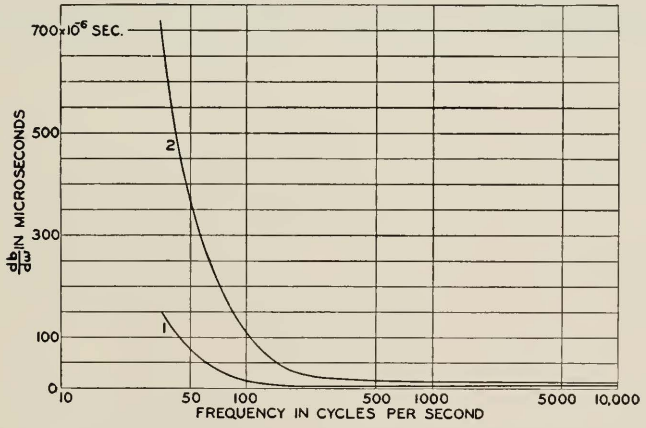


Fig. 23—Envelope delay characteristics of transformers

1. Output transformer
2. High ratio input transformer

transmission characteristic of the output transformer which is designed to work between impedances of 2000 ohms and 600 ohms when connected between generator and load circuits having these values.

Curves 2 and 3 show the effective transmission gain of transformers having voltage step-ups of 6.5 and 2.5 respectively, when used to connect the first stage of the vacuum tube amplifier to a 600-ohm generator impedance. The envelope delay curves for the output transformer and for the high ratio input transformer are given in Fig. 23. Photographs of the coils are given in Fig. 24. A large

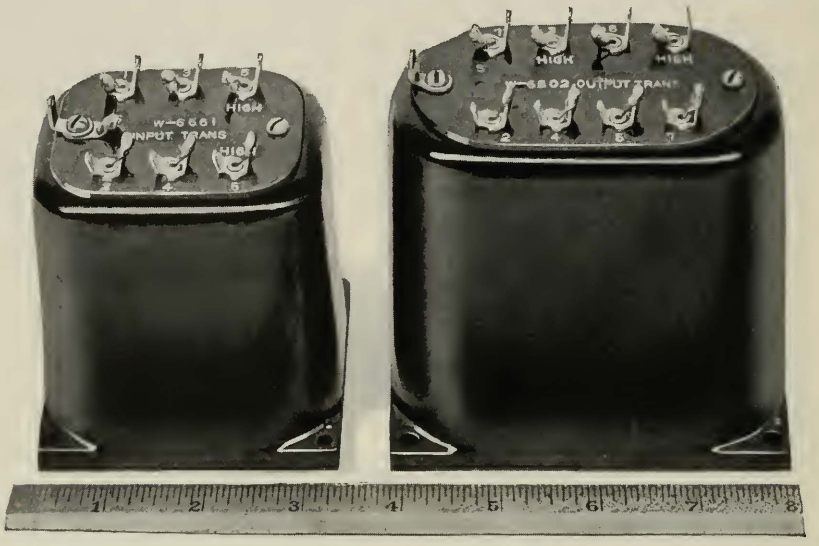


Fig. 24—Transformers used for coupling amplifier circuits to long distance telephone lines

factor in being able to get coils of this type lay in the availability of permalloy for the core material. The output transformer is connected to the amplifier through a blocking condenser in order to avoid possible saturation in the core due to the passage of direct current.

Measurements made on the several elements of the amplifier system have shown that its overall frequency characteristic is constant to within $\pm$ 2 T U from 10 to 20,000 cycles.

In an amplifier having as much gain as that just described it is apparent that a slight change in the potential of the power supply will cause a considerable change in the overall efficiency. Moreover, variations in the intensity of the light source used with the scanning system will cause corresponding changes in the intensity of the initial luminous signal wave. To insure that the energy level supplied to the line is at all times of the proper magnitude a level indicator has been provided to permit continuous observations of the output of the amplifier. This consists of an amplifier-rectifier circuit so arranged

that the space current of the last tube is a function of the alternating current voltage impressed on the first, being roughly proportional to the square of its amplitude. By means of a direct current milliammeter, therefore, it is possible to keep a very accurate check on the amplitude of the signal delivered to the line.

Receiving Circuit

Coming now to the receiving terminal equipment we find that the signal wave which was delivered to the line at a power level of 10 milliwatts may, under some conditions, be reduced to a level 50 T U below this, or to 0.1 microwatt. It is, therefore, necessary first of all to provide amplification to bring the signal to a level where it may operate the circuits controlling the illumination from which the image is to be reconstructed. In view of the fact that several types of receiving equipment are to be operated and also since the signal may be derived from any of several sources, either wire line, radio or local transmitting station, it is desirable to fix some one energy level as a reference point and to bring all signals to this value so that they may be supplied interchangeably to the several receiving systems. A convenient reference level is that already set as the proper input to a telephone line, namely, 10 milliwatts. At the receiving terminal, therefore, amplifiers have been provided which are similar to the final stages used at the transmitting terminal. These include units containing the small-sized tubes and terminate in units identical with that supplying current to the line except that the output transformer is omitted. The first stage is, as mentioned in the preceding section, connected to the line through an input transformer. The amplifiers associated with the several incoming signals are each provided with a level indicator of the type already described. These terminal amplifiers and the several receiving circuits are all terminated in jacks, exactly like telephone circuits, and it is possible, therefore, to connect any receiving machine to any desired transmitting station simply by patching the proper jacks together, exactly as telephone circuits are connected at the central office.

Before describing the final stages of the amplifier circuits it is necessary first to examine the properties of the light source which is to be controlled. In the case of the disk receiving machines described in the first section of this paper it is recalled that a single neon lamp is used having a rectangular electrode the entire area of which glows at each instant with an intensity proportional to the intensity of the initial luminous signal. The current voltage characteristic of a typical neon lamp is given in Fig. 25. It will be seen that no current flows

until the voltage across the lamp reaches the breakdown potential which, in the example shown, is about 210 volts. From this point on the current increases linearly with respect to voltages in excess of a value somewhat below the breakdown point. It will also be seen

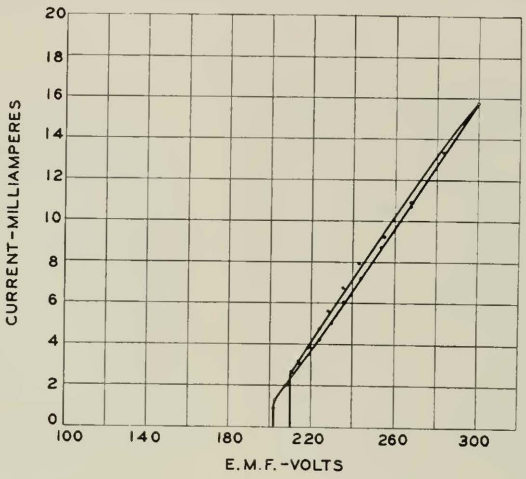
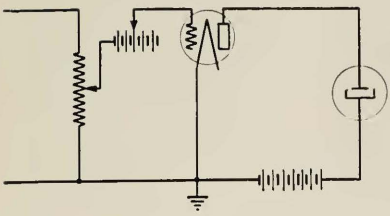


Fig. 25—Current voltage characteristic of typical neon lamp

from the curve that the value of current depends somewhat upon the direction in which the voltage is changing. In most cases, however, the function comes sufficiently close to being single valued for our present purposes. In view of the well-established linear correspondence between the intensity of the illumination resulting from the glow discharge and the current, it is required to so arrange the circuits that the current through the lamp is at all times proportional to the illumination at the transmitting terminal.

It will be recalled that the electric signal wave as transmitted through the various amplifier circuits differs fundamentally from the initial luminous wave in that the direct current component has been eliminated. It is necessary, therefore, to restore this component before the changes in light intensity at the receiving terminal will follow those at the transmitting terminal. The several factors entering at this point may perhaps best be examined in terms of an elementary circuit such as given in Fig. 26. In this case the neon lamp is connected in series with the plate circuit of a vacuum tube and its polarizing battery. The circuit may be considered for the present as equivalent to one in which the neon tube is replaced by an ohmic resistance and in which the potential of the polarizing battery is



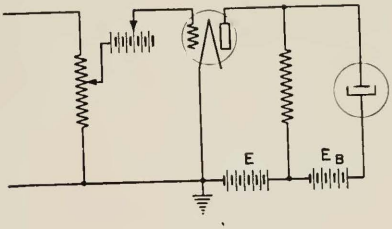


Fig. 26—Circuit schematic and operating characteristic of neon lamp amplifier

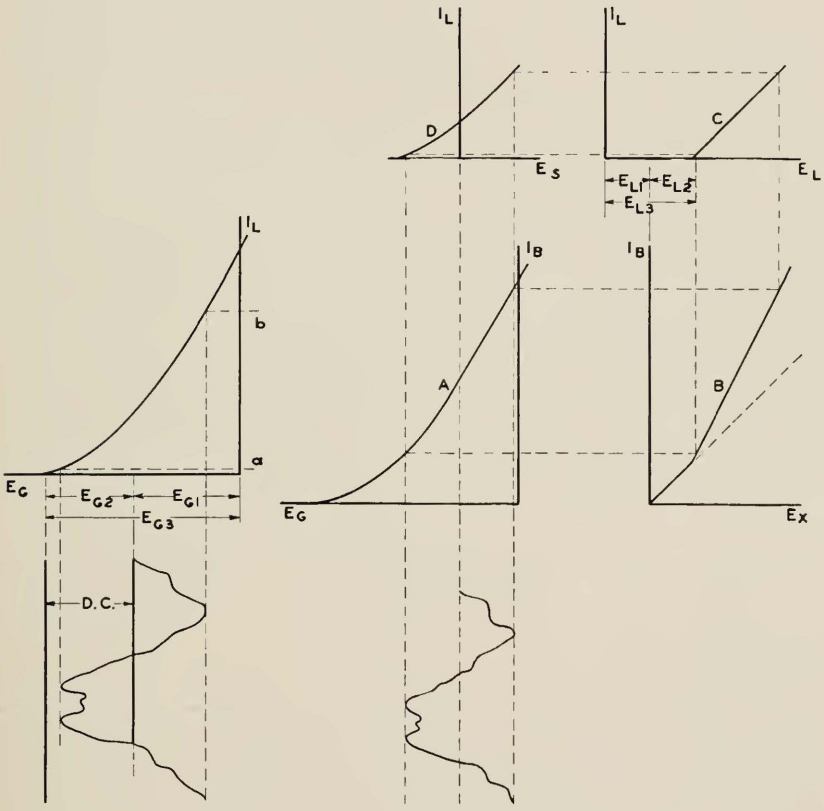


Fig. 27—Circuit schematic and operating characteristics of circuit arranged for linear operation of neon lamp

reduced by an amount corresponding to the back e.m.f. of the lamp. Under these conditions the relation between current—and therefore illumination—and the voltage on the grid of the vacuum tube is as shown by the curve given with the figure. This curve takes into account the change in potential between the plate and filament of the vacuum tube due to the voltage drop in the lamp resistance. If the reactances in the circuit are negligible, this curve may be taken as the dynamic characteristic of this portion of the system.

Let us assume that to properly build up the desired image at the receiving terminal the light is to be varied between the limits set by the two horizontal lines a and b. It is apparent that two adjustments are necessary in the grid circuit. The amplitude of the impressed alternating current must be such that the difference between its positive and negative maxima is equal to the difference between the grid voltages corresponding to these currents. This is taken care of by suitable adjustments of the amplification. It is further necessary that the bias introduced by the grid polarizing battery be such that the positive and negative peaks coincide with these same values of grid voltage. Under these conditions the grid battery must be looked upon as supplying two absolutely distinct biases, one the bias for the tube and the other the bias for the signal. For example, if the signal wave as delivered to the grid circuit contained the original d-c. component properly amplified, it would be necessary to adjust the system so that zero current would be obtained with no impressed signal. To accomplish this the tube would require the negative grid bias E_{G3}. Variations in signal voltage would then be considered as taking place about this value of grid potential as the origin. Thus E_{G3} is the operating bias of the tube. To properly locate the signal wave, however, it is necessary to add the positive bias E_{G2}. It will be seen from the curve that this bias corresponds exactly to the direct current component which is to be restored to the signal. The sum of these two biases, obviously, gives the actual bias, E_{G1}, with which the tube is operated.

In the circuit as shown the well-known curvature of the vacuum tube prevents us from obtaining a linear relation between the current through the neon lamp and the signal voltage. This condition may be overcome by a number of circuit modifications of which that shown in Fig. 27 is typical. Instead of connecting the neon lamp and the vacuum tube directly in series, a resistance is provided across which is set up a potential, E_X, proportional to the current through it. Across this resistance is shunted the neon lamp and a biasing battery, E_B. The adjustment of this circuit is indicated by the curves shown.

Curve A expresses the relation between the grid potential of the vacuum tube and its plate current. Curve B shows the relation between this same plate current and the voltage across the external resistance. When no current is flowing through the vacuum tube, the potential of the biasing battery is insufficient to break down the neon lamp and no current flows through the circuit containing the neon lamp and the plate circuit resistance. As the current through the vacuum tube is increased from zero, the total current flowing is that through the resistance branch. When, however, the potential drop across this resistance reaches such a magnitude that, together with the potential of the biasing battery, it is sufficient to break down the neon lamp, the latter will begin to draw current which thereafter increases linearly with further increases in the voltage, E_X, across the external resistance. The voltage across the neon lamp itself differs from that across the resistance by the amount of the battery E_B. The relation between the neon lamp current and the voltage across it, as given by Curve C, may therefore be plotted directly above the characteristic just discussed by displacing the vertical axis an amount corresponding to E_B. This amount is shown as E_{L1}. Here again we have two separate biases controlled by a single adjustment. The potential E_{L2} is fixed by the minimum plate current which can be taken from the tube without departing too seriously from the linear portion of the tube characteristic. It is, therefore, an operating bias of the circuit which is unaffected by any characteristic of the neon lamp. The latter, however, must be operated with a bias E_{L3} corresponding to its effective back e.m.f. As in the case of the grid circuit bias just considered, the bias E_{L1} actually introduced into the circuit is the difference between these two independently determined biases.

By projecting values of lamp current horizontally and plotting their intersections with vertical projections through the corresponding grid potentials on the vacuum tube characteristic we obtain Curve D, which expresses the relation between the instantaneous value of the signal and of the current in the neon lamp as derived from the characteristics of the several elements of the circuit. Inasmuch as the intensity of the illumination is proportional to the lamp current, it will be seen that we have approached the desired linear correspondence between the instantaneous values of the signal and of the light.

It will be noted that care has to be exercised to insure that the alternating current as impressed on the last vacuum tube is of the proper polarity. If it is not, the received image will be a negative instead of a positive. This may be controlled either by the connections to any one of the transformers or by the number of vacuum

tube stages. With an even number of stages the polarity will be reversed from that given by an odd number. This is because an increase in negative potential on the grid of a vacuum tube causes a decrease in the space current and hence a decrease in the negative potential applied to the grid of the next tube.

In the case of the grid type of lamp with the individual external electrodes, the impedance to which energy must be supplied differs materially from that presented by the rectangular electrode lamp already described. For low voltages the impedance between any electrode and the central helix is effectively a capacitance of the order of 6 m.m.f. When, however, the voltage gradient in the interior of the tube becomes sufficient to break down the gas and cause a discharge to take place, the capacitance is increased to about 15 m.m.f. In fact, the tube may be looked upon as consisting of two capacitances connected in series. When the applied potential is sufficient to break down the gas and cause a glow discharge, that capacitance corresponding to the portion of the path inside the tube is effectively shunted by an ohmic resistance. The minimum discharge potential has been found to be independent of frequency over a wide range, but the current between electrodes is inversely proportional to the frequency because of the presence of the capacitance between the electrode and the glowing gas. Now, the brightness of the discharge is a function of the current sustaining it so that it becomes desirable to use high frequencies in order to get sufficient light without going to prohibitively high potentials. It is also desirable to operate at such a portion of the frequency scale that the percentage difference between the limits of the range shall be small, thus avoiding signal distortion due to the effect referred to above. There is, however, a definite upper limit to the frequency beyond which it would be impossible to operate because of the stray capacitances in the cable connecting the grid to the distributor. It has been found feasible to operate at a frequency of the order of a half million cycles.

The circuit problem, therefore, involves the production of a high frequency wave which varies in amplitude in accordance with the amplitude of the received picture signal. The solution has been conveniently obtained by using a radio broadcast transmitter the voice frequency circuits of which have been so modified that the extended range of frequencies required might be handled with minimum distortion.

The envelope of the 500-kilocycle wave modulated by the picture signal, as shown in Fig. 28, is proportional to the signal amplitude plus a direct current biasing component of such magnitude that when the

envelope reaches 160 volts the tube fails to light. This corresponds to a black area in the picture. When no picture signal is being received, the amplitude of the unmodulated carrier wave causes the tube to light at average brightness, corresponding to the locally introduced d-c. component of the signal. It follows, then, that the amplitude of

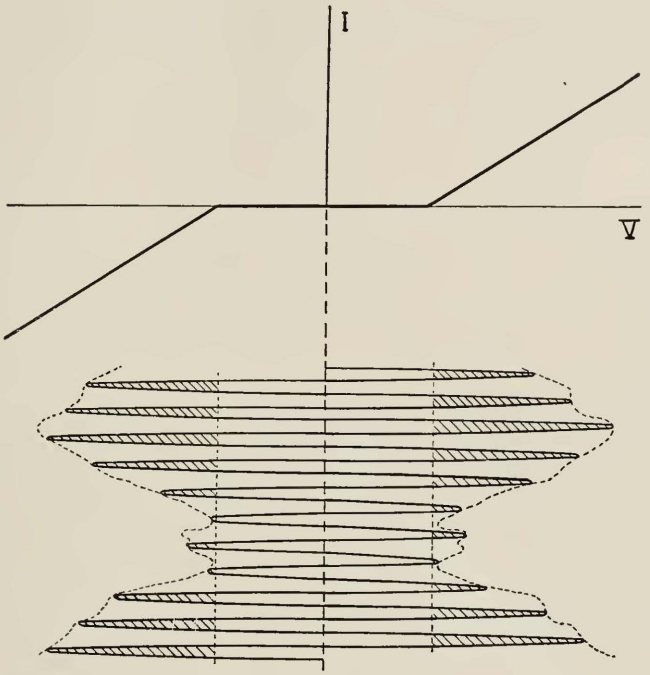


Fig. 28—Diagrammatic representation of relation between modulated high frequency wave impressed on grid type neon lamp and lamp characteristics. Intensity of glow is proportional to shaded area.

the unmodulated carrier is fixed, as in a previous example, by the joint requirements of two biases, that of the lamp and that of the signal bias.

There is a slight distortion inherent in this method due to the fact that the light, which is proportional to the shaded area of the curve of Fig. 28, is not strictly proportional to the amplitude of the envelope with respect to the 160-volt limit. This is, of course, because these peaks are portions of a sine wave and hence the time variation of the glow resulting from any given carrier cycle is a function of its amplitude. The effect is small, however, being most noticeable at low values of illumination.

In the case of the grid-lamp receiver the signal amplitude is adjusted,

as for the disk receiver, by a potentiometer in the low frequency portion of the circuit. The carrier amplitude, however, is adjusted by varying the plate potential applied to the oscillating tube. The coupling to the lamp is made by connecting the central helix and the distributor brush across a portion of the condenser of the oscillating circuit.

The frequency-amplitude relation of the envelope has been made practically constant by employing resistance capacitance coupling in the signal input amplifiers, by providing extremely high inductance retard coils for the modulator—which is of the Heising type—and by inserting resistance in the oscillating circuit to provide sufficient damping. The relations between the original picture signal and the envelope of the high frequency wave, with respect to both amplitude and phase shift, were observed over the signal frequency range by means of a Braun tube and found to be satisfactory. The impedance of the connecting leads to the commutator was also measured and found to have a negligible effect on the frequency and damping of the oscillating circuit.

It has been found that there may be a lag between the time when the potential is applied to an electrode and the time when the gas breaks down. This is especially true following an interval during which there has been no discharge within the tube. Because of this those electrodes which are the first to be connected in any one of the parallel portions of the tube may fail to light. To overcome this effect a small pilot electrode is kept glowing at the left-hand end of each tube, thus irradiating the branch in such a way that the illumination of all electrodes follows immediately upon the application of potential. These pilot electrodes, which are obscured from view of the audience by the frame of the grid, are supplied by means of an auxiliary connection to the oscillator with a potential somewhat lower than that ordinarily impressed upon the picture segments.

Appendix I

The signal of Fig. 13 in the body of the paper may be represented as follows:

$$\left.\begin{aligned} f(t) &= 0 && \text{for } t < 0 \\ &= \frac{t}{T} && \text{for } 0 < t < T \\ &= 1 && \text{for } t > T \end{aligned}\right\} \tag{1}$$

or by a Fourier integral in the form

$$f(t) = \frac{1}{\pi}\int_0^\infty d\omega \int_{-\infty}^{\infty} f(\lambda)\cos\omega(t-\lambda)d\lambda, \qquad (2)$$

where λ is an auxiliary variable of integration and ω is 2π times the frequency. To get the effect of sending this signal through a system which transmits all frequencies without phase or amplitude distortion up to a cut-off frequency f_c it is only necessary to replace the upper limit of the first integral sign by N where $N = 2\pi f_c$. Thus:

$$F(t) = \frac{1}{\pi}\int_0^N d\omega \int_{-\infty}^{\infty} f(\lambda)\cos\omega(t-\lambda)d\lambda.$$

Then from (1):

$$F(t) = \frac{1}{\pi}\int_0^N d\omega \int_0^T \frac{\lambda}{T}\cos\omega(t-\lambda)d\lambda + \frac{1}{\pi}\int_0^N d\omega \int_T^\infty \cos\omega(t-\lambda)d\lambda$$

$$= \frac{1}{\pi NT}\Big\{\cos Nt - \cos N(t-T)$$

$$+ Nt[Si(Nt) - Si(Nt-NT)]\Big\} + \frac{1}{\pi}\left[\pi + Si(Nt-NT)\right].$$

If we write $Nt = x$, $NT = z$, and $\pi F(t) = y(x)$, then

$$y(x) = \frac{1}{z}\Big\{\cos x - \cos(x-z) + x[Si(x) - Si(x-z)]\Big\}$$

$$+ \frac{\pi}{2} + Si(x-z),$$

where

$$Si(x) = \int_0^x \frac{\sin x}{x}dx.$$

A series of graphs of $y(x)$ for different values of the product NT is given in Fig. 15 in the body of the paper. These are generalized curves, the time scale depending on the particular value of cut-off frequency used. From these curves we can get the additional lag in the time, τ, in the rise of these curves over the original time T in Fig. 14.

Appendix II

Let $f(t)$ be the instantaneous intensity of the picture, and let it be represented by a Fourier integral:

$$f(t) = \int_0^\infty A(\omega)\cos[t\omega + \Phi(\omega)]d\omega. \qquad (1)$$

Let T = time required for the aperture to pass a given point, Fig. 29.

Let $\varphi(t_1)$ be height of aperture at distance t_1 from its center.

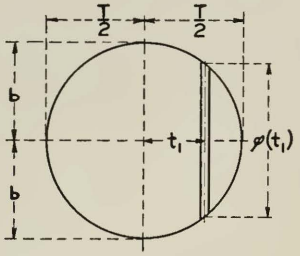


Fig. 29

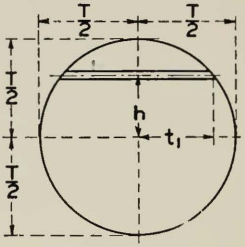


Fig. 30

Analysis of the aperture

The instantaneous amount of light passing through the aperture is

$$\begin{aligned} F(t) &= \int_{t-T/2}^{t+T/2} \varphi(t_1) f(t_1) dt_1 \\ &= \int_{t-T/2}^{t+T/2} \varphi(t_1) dt_1 \int_0^\infty A(\omega) \cos [t_1\omega + \Phi(\omega)] d\omega \\ &= \int_0^\infty A(\omega) d\omega \int_{t-T/2}^{t+T/2} \varphi(t_1) \cos [t_1\omega + \Phi(\omega)] dt_1. \end{aligned} \tag{2}$$

In the case of the rectangular aperture

$$\varphi(t_1) = \text{a constant} \tag{3}$$

and, except for a negligible constant factor,

$$\begin{aligned} F(t) &= \int_0^\infty A(\omega) d\omega \int_{t-T/2}^{t+T/2} \cos [t_1\omega + \Phi(\omega)] dt_1 \\ &= \int_0^\infty A(\omega) \left\{ \frac{\sin [(t + T/2)\omega + \Phi(\omega)]}{\omega} \right. \\ &\qquad \left. - \frac{\sin [(t - T/2)\omega + \Phi(\omega)]}{\omega} \right\} d\omega \\ &= 2 \int_0^\infty A(\omega) \frac{\sin T\omega/2}{\omega} \cos [t\omega + \Phi(\omega)] d\omega. \end{aligned} \tag{4}$$

The transformation from $f(t)$ to $F(t)$ amounts merely to changing the relative amplitude of the Fourier components of $f(t)$ by a factor proportional to $\dfrac{\sin T\omega/2}{\omega}$.

In the case of the circular aperture we can divide the aperture up into narrow elements parallel to the direction of motion, as shown in Fig. 30. Elements at a distance h from the middle line of the strip have lengths

$$2t_1 = 2\sqrt{T^2/4 - h^2}. \tag{5}$$

Each element considered as an independent *rectangular* aperture has the frequency characteristic

$$\frac{\sin t_1\omega}{\omega} = \frac{\sin \omega\sqrt{T^2/4 - h^2}}{\omega}.$$

The mean of all of these elementary frequency characteristics is

$$\begin{aligned}\frac{1}{T}\int_{-T/2}^{T/2}\frac{1}{\omega}\sin\left[\omega\sqrt{T^2/4 - h^2}\right]dh &= \frac{2}{T\omega}\int_0^{T/2}\sin\left[\omega\sqrt{T^2/4 - n^2}\right]dh \\ &= \frac{1}{\omega}\int_0^{T/2}\sin\left[\omega T/2\sqrt{1 - \frac{4h^2}{T^2}}\right]\frac{2dh}{T} \\ &= \frac{1}{\omega}\int_0^1\sin\left[T\omega/2\sqrt{1 - x^2}\right]dx \\ &= \frac{\pi}{2\omega}J_1(T\omega/2),\end{aligned} \tag{6}$$

where J_1 indicates a Bessel function of the first order. In place of the amplitude variation function $\frac{\sin(T\omega/2)}{\omega}$ for the square aperture, we have $\frac{J_1(T\omega/2)}{\omega}$ as such a factor. From the very nature of the physical processes under consideration it follows that this average value of the elementary frequency characteristics is effectively the frequency characteristic of the aperture as a whole.

Synchronization of Television [1]

By H. M. STOLLER and E. R. MORTON

SYNOPSIS: Synchronization of Television is the problem of holding two scanning disks so that their phase displacement is always less than four and one third minutes of arc. A 240-pole synchronous motor of the variable reluctance type is used as a basis. Coupled to it a direct current motor carries the steady component of the load. Hunting is eliminated by a condenser in series with the two synchronous motors whose capacitance is slightly less than that required to tune the circuit.

As the motor might lock into step in any of 120 possible angular positions, only one of which would give the proper phase relations, a two-pole motor, with only one locking position, was provided by tapping the armature of the direct current motor at two points and bringing out the leads to slip rings. This was used for synchronizing while the 240-pole motor, connected subsequently, held the close synchronism required. The disks rotate at 1062.5 r.p.m. which gives 17.7 cycles on the two-pole and 2125 cycles on the 240-pole motor.

For transmission the synchronizing current is attenuated to a level of .6 milliwatt and amplified at the receiving end. The 17.7-cycle current is an undesirably low frequency for transmission over telephone cables and so is used to modulate a 760-cycle current through a polarized relay. This is demodulated at the receiving end, where a polarized relay by interrupting a local battery current gives a rectangular wave which acts through vacuum tubes on the field of the direct-current motor.

THE problem of synchronization involved in television transmitting and receiving equipment is similar in principle to any synchronous motor problem but the requirements are of such a special nature that it is necessary to employ unusual features of motor design and control circuits to secure the required results.

General Requirements

At the transmitting end a scanning disk is employed containing 50 holes spirally spaced around the periphery of the disk rotating at a speed of 1060 r.p.m.[2] It is desired to rotate a similar scanning disk at the receiving end so that the hole through which the observer is looking at a neon lamp will be in a position corresponding to the hole which is transmitting light at the same instant at the transmitting end. Since there are 50 holes in each disk, the holes will be spaced apart 7.2 degrees, thus 7.2 degrees of arc correspond at the receiving end to the width of the picture. Since the horizontal resolving power is approximately the same as the vertical (0.02 of the picture dimension), the arc occupied by a picture element is 0.02×7.2 or 0.144 degree. In order not to appreciably impair the quality of the picture,

[1] Presented at the Summer Convention of the A. I. E. E., Detroit, Mich., June 20–24, 1927.

[2] This speed was determined by transmission considerations and is discussed in the companion paper by Messrs. Gannett and Green.

it is necessary to hold the synchronization within approximately ½ of the width of one element. This gives 0.144 degree divided by 2 or 0.07 degree as the requirement within which synchronization should be held. By way of comparison it might be mentioned that the angular twist in a length of 6 ft. of 1-in. steel shafting operated at rated load is of about the same order of magnitude.

An ordinary four-pole synchronous motor when operating at full load, unity power factor, has an angular phase displacement of about 20 electrical degrees between the impressed and back e.m.f. This corresponds to 10 mechanical degrees since the motor has two pairs of poles. If this motor is operated at constant load and the line voltage is varied, the phase angle will decrease with increasing voltage, or when the voltage is held constant and the load is varied the phase angle will increase with increasing load. It is at once apparent therefore that the ordinary type of synchronous motor will not even approach the degree of precision required for the reason that any minute change in line voltage or load will cause variations in its phase angle of lag with respect to the impressed frequency of a far greater amount than 0.07 degree. Consider, however, a motor having 120 pairs of poles. Allowing 20 electrical degrees as the normal full load phase displacement, this would be equivalent to 20 divided by 120 or 1/6 degree mechanical phase displacement. Even this amount is over twice the required permissible displacement of 0.07 degree. Since the variation of the phase displacement is the important factor and not the absolute amount of displacement, it is evident that if the line voltage and load are held reasonably constant a synchronous motor with 120 pairs of poles should be sufficiently precise.

Another requirement in addition to close phase synchronization is regulation of the acceleration or deceleration of the generator at the transmitting end. Such regulation is required due to the fact that an appreciable time is taken for the transmission of the synchronizing current a distance of 220 miles (circuit length) between New York and Washington. The velocity of propagation over the cable was approximately 19,000 miles per second while that of the picture on the open wire of 285 miles circuit length was about 175,000 miles per second, the corresponding times of transmission being .0116 second and .0016 second, leaving a difference of .01 second approximately. Since the total permissible error in synchronization is .07 degree, it is reasonable to allow .02 degree as error due to acceleration regulation. Let a be the acceleration in degrees per second per second. Substituting in the formula $s = \frac{1}{2}at^2$ gives $.02 = \frac{1}{2}a(.01)^2$ or $a = 400$ degrees

per second per second or a little over one revolution per second per second. For comparison consider a ¼-h.p. unregulated shunt motor. If the line voltage increases 10 per cent, it will cause an increase in speed from 4 per cent to 8 per cent depending on the magnetic saturation in its field circuit. This increase in speed will take place in a half second or more depending upon the moment of inertia of the load. Thus the acceleration in the case of a 1060-r.p.m. speed would be much greater than one revolution per second per second.

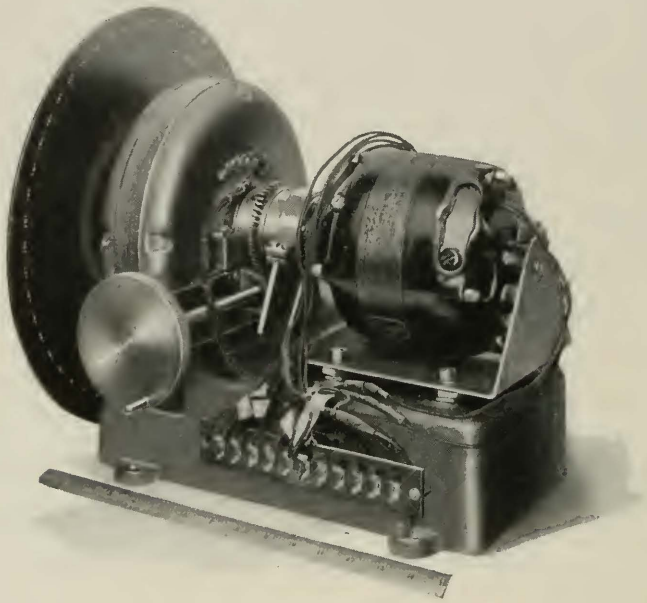


Fig. 1—Assembled motor

Since this problem of speed regulation is a separate one from that of the synchronization, the description of the regulating circuit is taken up later on.

Motor Design

In accordance with the phase displacement requirement as explained previously it was decided to build the synchronous motors with 120 pairs of poles, thus giving a frequency of 2125 cycles at 1062.5 r.p.m. which was the exact speed finally employed. For the sake of mechanical simplicity these machines were made of the variable reluctance type which gives one cycle per rotor tooth, thus requiring 120 teeth. The variable reluctance construction also simplifies the coil arrangement, the machine having only eight armature coils

instead of a separate coil for each tooth. Fig. 1 shows a photograph of the assembled motor and Fig. 2 an inside view of the stator and rotor.

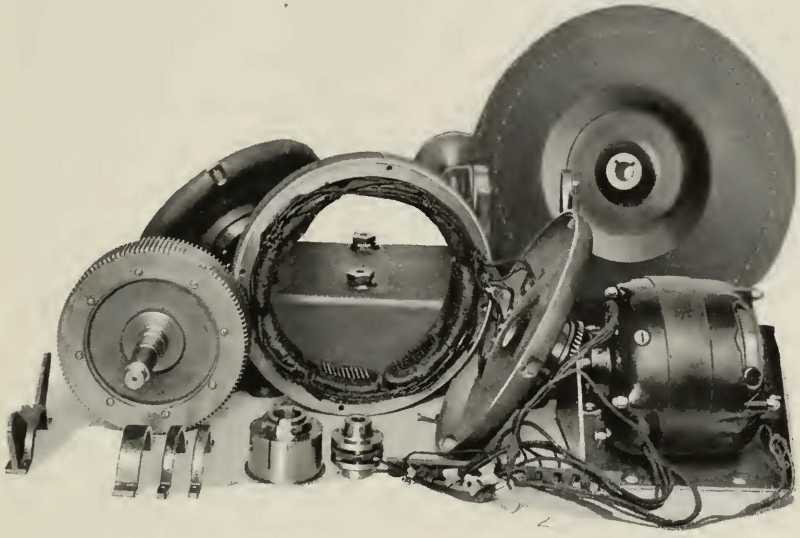


Fig. 2—Motor disassembled

In the preliminary experimental work two of these machines were directly connected (Fig. 3), permitting either machine to act as a

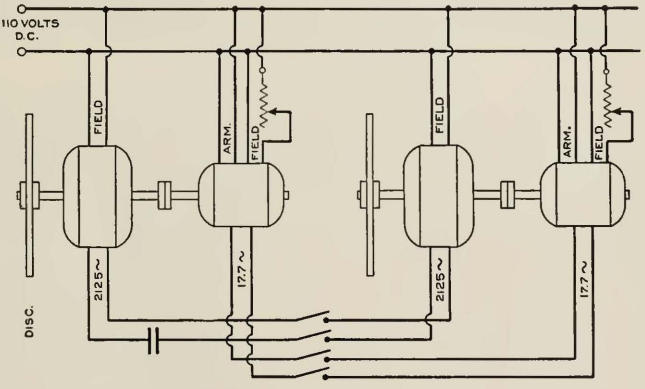


Fig. 3—Synchronization system over short wire line

synchronous motor loading down the other machine. Each machine was driven by a shunt d-c. motor having inherently poor regulation, the d-c. motors furnishing the power and the a-c. machines transferring

the variations from one d-c. machine to the other to hold synchronism in a completely two-way system. As was to be expected, it was found that the motors hunted badly at a frequency of about four cycles per second. In other words, instead of holding within a fixed electrical phase angle of 20 degrees the receiving motor oscillated throughout a phase angle of about $\pm$ 20 electrical degrees. This, of course, made the picture wabble back and forth across the aperture and was therefore unsatisfactory.

The ordinary method of preventing hunting by means of copper bars embedded in the pole faces was not practical on account of the large number of poles and limited space. The hunting trouble was cured by employing a series condenser between the motors using a value of capacity somewhat less than that required to tune the circuit. A rigid analytical treatment of this anti-hunting circuit is beyond the scope of this paper but its operation depends in general upon the curvature of the tuning curve due to the variation of the inductance of the machine with phase displacement. Since the condenser operates on the total inductance of the circuit, it is desirable to make the natural periods of oscillation of the two motors different. Otherwise a decrease in the inductance of one machine may be accompanied by a simultaneous and equal increase in the inductance of the other, thus leaving the total inductance unchanged and preventing the condenser from functioning. This was done by making one disk substantially heavier than the other.

The series condenser also neutralizes the greater part of the internal reactance of the motors, thereby increasing the steady state torque.

Framing of Picture

There was still one unsatisfactory feature in this system in that the motor at the receiving end could interlock in any one of 120 different angular positions whereas in order to get proper framing of the picture it must be synchronized at a particular angular position. For example, if the disk at the receiving end is exactly 180 degrees out with respect to the disk at the transmitting end, the observer will see the lower half of the picture on top; a dark space representing the dividing line between pictures and the upper half of the picture at the bottom. Similarly, if the disk is 90 degrees out at the receiving end, the lower quarter of the picture will appear on the top and the upper three quarters of the picture on the bottom. The disk at the receiving end may be brought into correct angular position by providing means for turning the entire motor through the necessary angle. It was found, however, that the rate at which the motor can be turned was limited

by the fact that if it were rapidly turned it would throw the motor out of step.

As an aid to framing, therefore, a second two-pole low frequency interlock was added to the system by providing the d-c. motors on each end with a pair of slip rings tapped to two opposite commutator bars. The d-c. shunt motors thus acted as converters furnishing 17.7 cycles at 1062.5 r.p.m. With this added feature on both the transmitting and receiving motors the process of synchronization was

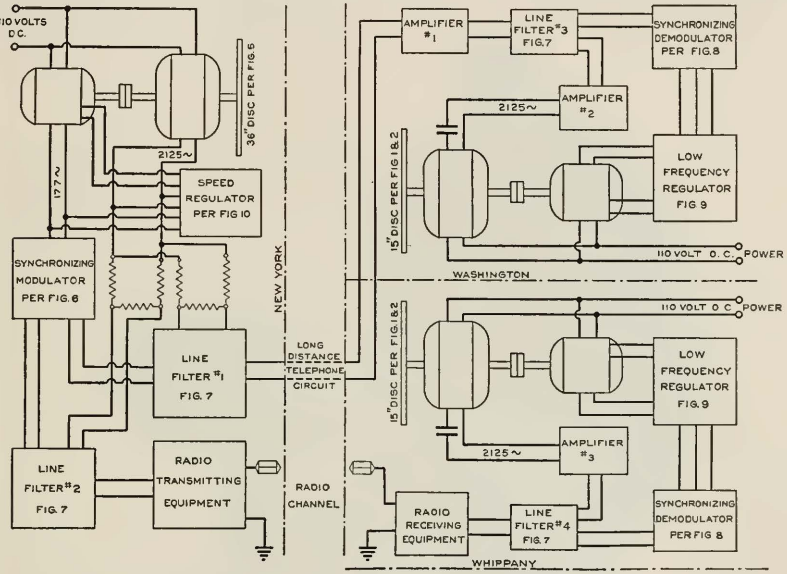


Fig. 4—Complete circuit of synchronizing system

first to close the 17.7-cycle circuit and adjust the field rheostat of the receiving motor until it came into step. Since this was a two-pole circuit there was only one angular position at which synchronization could occur. The high frequency synchronous machines were then connected together, thereby limiting the phase displacement to within .07 degree, as previously described. The high frequency motors in this system take the variation in load while the low frequency motor takes care of the steady constant component of load. Incidentally the addition of the low frequency synchronous motors greatly facilitated the synchronization of the high frequency motors inasmuch as it insured the proper initial speed. When the high frequency switch was closed there was merely a slight shift in phase angle to bring the receiving motor into step. The schematic circuit of the system thus far described is shown in Fig. 3.

Synchronization Over Long Lines

The above description explains the action of the synchronization system over lines of negligible impedance. In order, however, to secure similar results over a long distance telephone line or radio channel it is necessary to first attenuate the high and low frequencies to a power which can be safely applied to the transmitting end of the line and then amplify the power at the receiving end to restore it to the proper level. Fig. 4 shows the complete system employed.

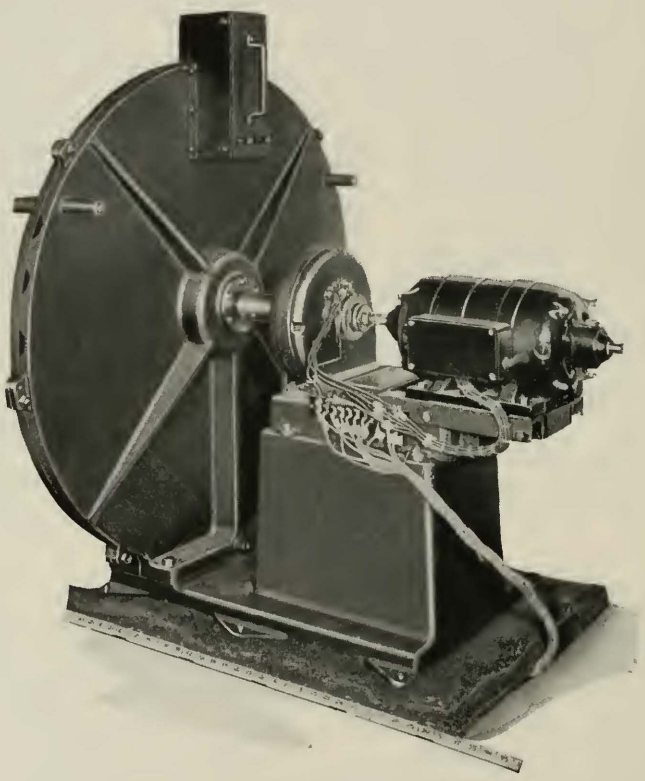


Fig. 5—Large scanning disc motor

While the high and low frequency machines on the transmitting end could have been designed so as to produce exactly the right power level, it was desirable, for the sake of interchangeability, to build the transmitting and receiving motor equipment of the same size. The output from the transmitting high frequency generator (shown in Fig. 2) when untuned was approximately 17 volts at 2125 cycles.

By means of a network this output was cut down to a level of 1 milliampere into 600 ohms impedance, the output impedance also being 600 ohms. This is a satisfactory level at which to transmit the high frequency, without inducing noise in adjacent wires in the telephone cables.

In the case of the low frequency interlock it was undesirable to attempt to transmit 17.7 cycles over a long distance line. The 17.7 cycles was therefore used to operate a polarized relay, the contacts of which modulated the output of a 760-cycle electro-mechanical oscillator [2] as shown in Fig. 6. In other words, the relay short-circuited

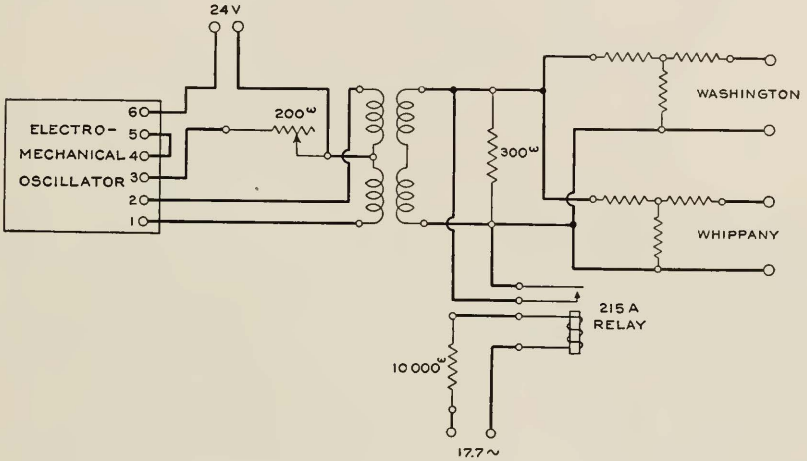


Fig. 6—Synchronizing modulator

the output of the oscillator alternate half cycles before application to the telephone line. Instead of using separate telephone pairs for the 2125-cycle and the modulated 760-cycle current, the two were combined by passing them through the line filter (shown in Fig. 7), thereby requiring only one pair for transmission of both frequencies. An identical network was employed for the radio channel. The problem of transmission of the synchronizing current is covered in the paper by Messrs. Gannett and Green and in the case of radio transmission in the paper by Mr. Nelson.

Receiving and Amplifying Circuits

Passing over this part of the problem, therefore, assume that the synchronizing currents have been obtained at the receiving end of the line. This power was delivered at a very low level, being about

[2] Described in the Bell Laboratories *Record*, March, 1927.

.3 of a milliampere into 600 ohms impedance, or 50 microwatts. It was then given a preliminary stage of amplification (amplifier No. 1, Fig. 4), passed through the line filter No. 3 (Fig. 7) and separated into 2125 cycles and 760 cycles modulated at 17.7 cycles. The 2125-cycle component was then amplified by two stages of amplification (amplifier No. 2) ending in push-pull 50-watt tubes and applied to

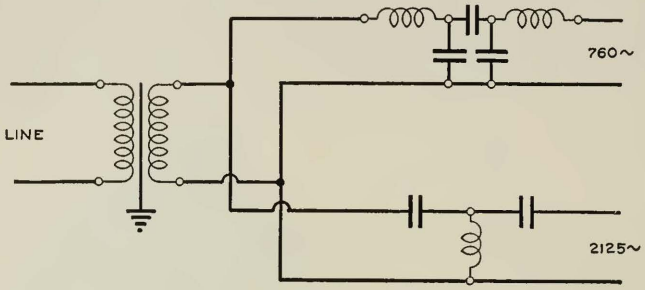


Fig. 7—Line filters for synchronizing frequencies

the high frequency motor. These amplifiers being of the standard type are not described. The terminal voltage on the output coil of the amplifier was made greater than that of the high frequency motor so that the power flow was normally from the amplifier to the motor.

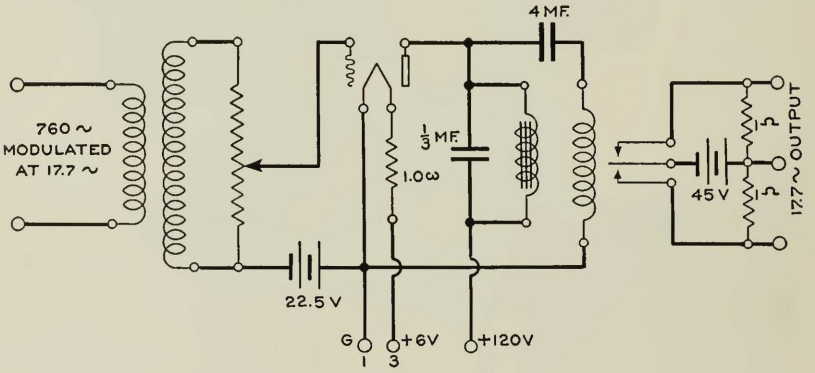


Fig. 8—Synchronizing demodulator

The anti-hunting condenser was retained between the amplifier and the motor.

In the case of the low frequency circuit the output from line filter No. 3 was received in the form of 760 cycles modulated at 17.7 cycles. This was passed through the demodulator (Fig. 8) which operated a polarized relay whose armature opened and closed its contacts at 17.7 cycles per second. The contacts of the relay provided square-wave low frequency current by interrupting power from a local battery

source. On account of the limited power output which the vibrating contacts could safely handle without sparking, it became necessary to amplify this low frequency output. While this would have been possible by the use of ordinary amplifier circuits, it was found preferable from the standpoint of economy of apparatus to apply the low frequency regulation through a field circuit of the receiving motor. Referring to Fig. 9 it will be noted that the plate circuit of the reg-

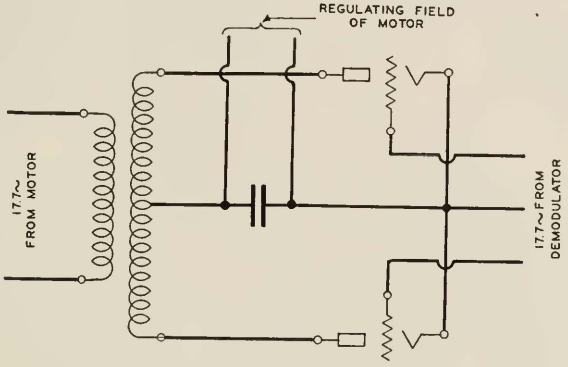


Fig. 9—Low frequency regulator

ulating tubes is supplied from the secondary of the transformer which is connected to the slip rings of the motor, while the grid circuit of these tubes is supplied with low frequency, low power 17.7 cycles from the contacts of the relay. As the motor is started up from rest the shunt field is weakened until the motor falls in step. At this point the frequency of the plate supply to the regulator tubes is identical with that supplied to the grids. If the phase relationship is such that the plates go positive at the same time that the grids are positive, then the space current of the tubes is increased and the regulating field (which is an aiding auxiliary field) is strengthened, thereby preventing a further rise in the speed of the motor. In other words, for each combination of load and line voltage there is an equilibrium phase position between the plate and grid voltages at which the corresponding regulating field current maintains the speed at the desired value.

Motor Operation

In actual operation the procedure was to first synchronize on the low frequency, and then on the high frequency circuit. The precise framing of the picture was then adjusted by rotating the motor by means of worm gearing through the necessary angle to center the image properly in the aperture. The high frequency current was of the order of 1.5 amperes at 2125 cycles with a terminal voltage of

100 volts at the high frequency motor. The power taken by the d-c. motor was approximately .8 ampere at 110 volts. The current through the regulating field controlled by the 17.7-cycle circuit was of the order of 20 to 40 milliamperes at 100 volts depending upon the phase position at which interlock occurred. It was found preferable to cut off the low frequency interlock feature after synchronization and framing had been obtained in order that irregularities in the time of contact closure of the relay might not produce changes in field strength of the d-c. motor which in turn would cause irregularities in power output. Such irregularities would give rise to phase shifts in the high frequency machine, thereby producing unsteadiness of the picture.

Operation on Radio Channel

In the case of transmission of the synchronizing current by radio instead of by wire the same apparatus is employed except that it was found necessary to use a much higher value of high frequency current in order to hold the high frequency motor in step, the current being approximately 4 amperes as compared to 1.5 amperes in the case of the other motors. This greater current was found to be necessary in order to hold the motor in step within the necessary phase angle of displacement, in spite of various types of interference picked up by the radio receiver, and associated circuits. This was mainly inductive interference from the picture and speech transmission sets arising from the fact that the synchronizing current was transmitted from New York to Whippany and picked up on a receiving set there, whereas the picture and voice current was transmitted from Whippany to New York. A certain amount of interference was also encountered from ship spark sets and static.

Speed Regulation of Transmitting Motor-Generator

As previously explained under "General Requirements" the essential requirement of the speed regulator at the transmitting end is to limit the acceleration to about one revolution per second per second, over intervals as small as .01 second. The ordinary type of centrifugally operated vibrating contact regulator keeps the motor continually accelerating and decelerating between an upper and lower speed limit and while such a system could theoretically be employed if the flywheel were made large enough, it was obviously preferable to employ a type of regulator in which the speed was inherently held constant without such acceleration and deceleration.

The regulating circuit employed is shown in Fig. 10. The complete theory of this regulating circuit is to be covered in another paper to be presented before the Institute. Briefly, the principle consists in

employing a sharply tuned circuit as the primary speed-controlling element resonating at a frequency slightly less than the frequency at which the machine is operated. A voltage from the high frequency generator is applied to this tuned circuit and thence to a detector tube which in turn operates on the grids of a pair of push-pull regulator tubes; these tubes controlling an auxiliary regulating field winding on the motor. The circuit also contains anti-hunting means, the

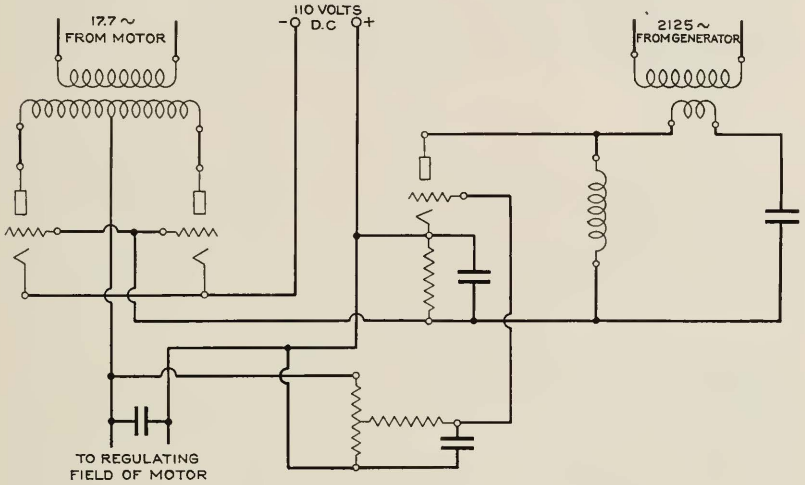


Fig. 10—Speed regulator

theory of which will be given in the later paper. Instead of applying this regulating circuit to the small 15-in. scanning disk motor shown in Fig. 3, it was decided on account of its greater flywheel effect to use the large 36-in. disk shown in Fig. 5 which was used for receiving the picture at New York. It therefore became the transmitter from the synchronizing standpoint for all of the other units although from the picture standpoint the big disk acted as a receiver.

Local Stations

In addition to the stations at Washington and Whippany there were three local stations in New York employing similar high and low frequency synchronous motors with 15-in. disks. These were controlled in the same manner except that first stage of amplification and the line filters were omitted. One station was employed for monitoring purposes, another operated a local transmitter, while the third operated the big grid receiver seen by the entire audience.

Wire Transmission System for Television[1]

By D. K. GANNETT and E. I. GREEN

SYNOPSIS: This paper deals with the transmission problems which were met and solved in connection with providing wire circuits from Washington to New York for the television demonstrations which took place on April 7, 1927, and following. For transmission of the television images a single transmission channel was set up combining the frequency ranges usually assigned to telegraph, telephone and certain carrier channels. The special line requirements were met so successfully that the television images transmitted from Washington were indistinguishable from those transmitted locally.

INTRODUCTION

A SYSTEM of television, to be worthy of the real meaning of the name, must be capable of operation over a considerable distance. Spanning this distance, there must be a connecting medium suitable for faithfully transmitting the television currents. This paper describes how the connecting medium was provided between Washington and New York for the recent television demonstrations,[2] by adapting to this purpose existing wire facilities of the Bell System.

Fortunately, wire facilities of the type which were available between Washington and New York had been utilized for some time to transmit simultaneously many telephone and telegraph messages, involving a frequency range more than ample for the television requirements, so that the transmission characteristics of the lines throughout the necessary range of frequencies were well known. The matter of providing a suitable channel to carry the television currents consisted, therefore, in throwing together the frequency ranges which had heretofore been utilized for providing a number of separate telephone and telegraph channels. In addition to providing this very wide band communication channel it was necessary to apply special distortion-correcting networks so that the overall channel would possess proper characteristics and also to take care to avoid introducing disturbances due to such things as line irregularities, noise, etc.

Due to the perfection of the transmission methods which were utilized, it was found that when the circuit was first established, in accordance with the requirements which had been deduced, the television images transmitted from Washington were indistinguishable in quality from those transmitted locally, this result being secured

[1] Presented at the Summer Convention of the A. I. E. E., Detroit, Mich., June 20–24, 1927.

[2] "Television," H. E. Ives; "The Production and Utilization of Television Signals," F. Gray, J. W. Horton and R. C. Mathes; "Synchronization in Television," H. M. Stoller and E. R. Morton.

without any deviation from the adjustments which had been worked out in the original design.

Requirements

General. The ideal requirement for a transmission line for television, or for that matter any other purpose, is, of course, that it introduce no distortion whatsoever, in which case there could be no question but that the television images obtained in the receiving apparatus after transmission over the long distance line would be identical with the image obtained with the transmission only over a distance of a few feet. Practical transmission lines, however, tend to introduce a certain amount of distortion and the less the allowable distortion which is specified the greater will be the cost of providing a proper line. Before going ahead with the matter of engineering the line required to transmit the television currents from Washington to New York it was, therefore, first necessary that the requirements be set. The requirements were made more severe than strictly necessary in cases where they were easy to meet.

Frequency Range. In any system for the electrical transmission of intelligence, the required frequency range is, in general, proportional to the speed of transmission. In the case of picture transmission or television, the speed of transmission may be expressed in terms of the number of picture elements which must be transmitted per second, where a picture element is the smallest unit area which it is intended to be able to distinguish in the received picture from its neighboring unit areas.

When the picture currents are transmitted in the most efficient manner, the frequency range necessary is approximately equal to half the number of picture elements which must be transmitted per second. A simple way of seeing this is to realize that as the picture elements are transmitted in sequence, the greatest possible rate of variation of detail is obtained when alternate picture elements are black and white. A complete cycle corresponds in this case, therefore, to the time interval required to transmit two picture elements.

According to this relationship this particular television system in which about 40,000 picture elements per second are transmitted should require a frequency range of approximately 20,000 cycles. As a matter of fact it was found by a laboratory test that due to certain characteristics of the apparatus a frequency range as great as this was ample, just detectable distortion being introduced in the reproduction of the human face when the range was narrowed to about 14,000 cycles. In providing the line circuit, however, extending the

frequency range to 20,000 cycles involved so little difficulty that it was decided to provide this very liberal frequency range.

In the particular television system which has been described the very low frequencies (below about 10 cycles) are suppressed. It was, therefore, not necessary that the line transmit these very low frequencies. The frequency range which the line should transmit was accordingly set as 10 cycles to 20,000 cycles.

Attenuation. Referring to still picture transmission, it has been found that variations of attenuation with frequency of several transmission units do not appreciably impair the quality of the picture. Since no great difficulty was anticipated in meeting closer limits, however, it was decided to set the limits for the variation of attenuation with frequency at $\pm$ 2 T U within the frequency range of 10 to 20,000 cycles.

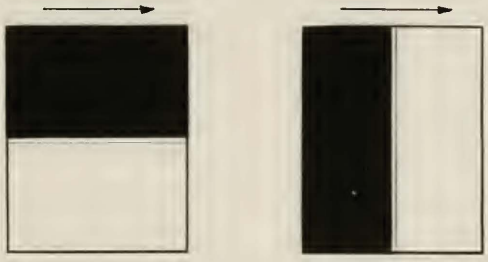
Phase Characteristics. A characteristic of wire lines, whose importance has been increasingly realized in recent years, is their phase characteristic. In speech transmission, transients due to unequal velocity of the different frequency components have been found to be an important consideration on some types of lines. In picture transmission and television, also, it is important that this phase distortion be controlled, as otherwise the image might be blurred due to the arrival of the various frequency components at different times. The type of transient which has been found to impair the quality of pictures is the type which is relatively rapid and the aim has been to make the phase characteristics such that those transients would be small.

The requirement with respect to phase for distortionless transmission is that β/ω be a constant where β is the phase change in radians for the entire circuit, and ω is equal to 2π times the frequency. β/ω is known as the "phase delay" or the steady-state time of transmission. $d\beta/d\omega$ is the time required for the transmission of the envelope of a wave whose components center closely about the frequency $\omega/2\pi$ and it will be referred to as the "envelope delay." Since it is more convenient to measure the envelope delay, the requirements were set up in terms of this quantity. When β/ω is constant, it is evident that $d\beta/d\omega$ is also constant. While the converse of this is not in general true, the conditions as actually encountered were such as to permit its use as a measure of the small variations involved.

The envelope delay characteristics of a number of circuits, which have been found to give varying degrees of transient on still pictures, have been measured. Also data were available from tests of picture transmission through filters and other networks whose delay charac-

teristics were known. From these various data, the permissible deviations of the delay characteristic for still picture transmission were determined, and dividing these figures by 50, the ratio of the rate of transmission in picture elements per second in the two cases, the limits for the television circuits were obtained. In this way it was decided to attempt to keep within ± 10 microseconds, if possible, with outside limits of ± 20 microseconds. Check tests of these limits were made with the television apparatus in the laboratory by transmitting the currents through various known networks, and noting the effect on the received image.

Unlike the attenuation requirements, the delay requirements for television are not the same over the entire frequency range, but are much more lenient in the lower frequency range, as was shown by experiments in the laboratory. A physical picture of the reason for this may be obtained by reference to Fig. 1.



Arrows show direction of motion of light spot

A B

Fig. 1

Fig. 1A shows a picture placed in position before the sending machine, consisting of a piece of cardboard the same size as the image-area which can be transmitted, the upper half of the cardboard being colored black, while the lower half is white. As has been explained in the paper by Messrs. Gray, Horton and Mathes, the picture is scanned by a spot of light which moves from left to right in successive lines, tracing 50 horizontal lines across the picture in one sixteenth of a second. The first 25 of the lines lie on the black and the remaining 25 on the light part of the picture. The process is repeated 16 times per second, each repetition of 50 lines giving one complete cycle of black and white. The frequency components in this case are multiples of 16 cycles. A transient which blurs the picture outline over a given number, n, of picture elements (downwards) corresponds to a time interval equal to the time of tracing n lines, *i.e.*, $n/800$ second.

Now consider Fig. 1B. Here the picture has been rotated 90 degrees.

In this position, a complete cycle of black and white is obtained with each line instead of with each 50 lines. The frequency components in this case are multiples of 800 cycles and bear the same relations to 800 cycles as the components spoken of above bear to 16 cycles. A transient which blurs the picture outline n picture elements (horizontally, this time) corresponds to a time interval of n forty thousandths of a second. Evidently the delay requirements are 50 times more lenient in the former case than in the latter so that the delay requirement at the highest frequencies, which determine the fine detail in the direction of scanning, is 50 times as severe as at low frequencies, which determine the fine detail in a direction perpendicular to the direction of scanning.

In the still pictures referred to, the transients extended in the direction of travel of the light spot and there were no transients analogous to those discussed here in connection with Fig. 1A. For this reason the delay limits determined from still picture transmission are the ones which apply to the higher frequencies. For the lower frequencies the requirements are obtained by multiplying the high-frequency requirements by 50. For these reasons, together with the result of a Fourier analysis of the picture current, the limits were set at $\pm$ 10 or $\pm$ 20 microseconds from 400 to 20,000 cycles. Below 400 cycles, the departures from the constant delay were permitted to be $\pm$ 500 or $\pm$ 1000 microseconds.

Noise. Another important requirement is that relating to the ratio of the picture currents to the extraneous interfering currents which may arise in the line from power induction and other sources. Early experience with the television apparatus showed that considerably more noise was permissible in the case of television than in the case of still picture transmission so that in this case comparison with the still picture transmission would result in an unduly severe requirement. This is thought to be explained by the fact that in the case of television the pictures are flashed before the eye 16 times per second and the effects of the extraneous currents occur on successive flashes in different positions, so that defects of one flash are corrected on the next.

A set of experiments was performed from which it was determined that if the ratio of average picture currents to average noise currents exceeded about 10 the results were satisfactory. In order to assure considerable margin above this figure, it was decided to make the average television current to be transmitted into the line 4 milliamperes.

Echoes. If two paths exist by which the currents may travel from the sending point to the receiving point, the length of the two paths

being different, a double image will be produced on the received picture, forming what may be termed visual echo. In the case of telephone lines, the echoes may exist on account of reflections between impedance irregularities in the circuit so that the currents arrive at the receiving point both by way of the direct transmission path and by way of a transmission path which includes an extra loop between two irregularities. If the echo is not greatly attenuated with respect to the main transmission, the result may be quite disturbing on the received picture. It has been found by experiment that the echo is too weak to be seen if it is more than 25 T U weaker than the main current and, accordingly, care was taken in setting up the New York-Washington circuit to avoid introducing echo paths of lower equivalent than this.

General Choice of Method

Two general methods are possible for transmitting the currents over the line circuits. One method is to transmit the currents directly without change of frequency. This method involves the transmission of the currents of the frequency range determined upon above, namely, from about 10 cycles to about 20,000 cycles per second.

The other general method is the carrier method, in which the television currents modulate a carrier current of suitable frequency and are thereby moved to another portion of the frequency spectrum prior to transmission over the line. At the receiving end of the line the carrier currents are then restored to the original frequencies of the television currents.

Several different schemes of carrier transmission are possible. The simplest is to modulate a carrier with the television currents and to transmit both side bands. This has the disadvantage of requiring the transmission of twice as wide a frequency range as that occupied by the original television currents. Another scheme is to transmit a single side band. A third possible scheme is to transmit both side bands for the lower frequencies and only one side band for the higher frequencies.

One advantage to be secured by the carrier method is that it lessens the severity of some of the line problems through avoiding the transmission of very low frequencies over the line circuit. At these frequencies the amount of noise found on lines is usually considerably greater than at the higher frequencies.

After weighing the relative merits of the carrier and direct transmission methods it was decided to make use of the latter because of its simplicity. An important factor in this decision was the successful development, for use in connecting the apparatus to the lines, of

transformers providing adequate transmission of the entire frequency range from 10 cycles to 20,000 cycles.

Arrangements for Television Circuits

Line Layout between New York and Washington. The layout of the wires between New York and Washington is shown in Fig. 2. The circuit over which the waves actually carrying the pictures were transmitted (marked Picture Circuit) consisted principally of a pair of copper wires 165 mils in diameter. At a number of places on the route the circuits were carried in cable as indicated in the figure. The total length of the television circuits was about 285 miles, of which 8 miles consisted of cables and the remainder of open wire.

Transpositions. As the circuits employed were originally designed for voice-frequency operation only, except for a section at the New York end, it was necessary to add transpositions to them to prevent interaction with adjacent circuits at the high frequencies involved in the television transmission. The high-frequency currents were thus prevented from passing over into the adjacent circuits which would have resulted in irregularities in the attenuation, line impedance and phase shift characteristics of the circuit.

Incidental Cables—Loading. Any appreciable length of non-loaded cable included in an open-wire television circuit has certain very objectionable effects. The impedance irregularities introduced by the cable destroy the uniformity of the line attenuation, impedance and phase shift characteristics as a function of frequency, and tend to produce echoes as described above. Types of loading developed for use on incidental cables occurring in circuits employed for carrier telephone and carrier telegraphy operation [2] were employed to reduce these effects to a minimum. This carrier loading is designed so that when used on No. 13 A. W. G. cable circuits it provides an impedance which approximates very closely that of the open wire. With a spacing of about 930 feet between loading coils, this loading has a nominal cut-off of about 45,000 cycles, which corresponds to an effective transmission range extending up to about 36,000 cycles. In order to obtain a close match between the impedances of the open-wire and the cable pairs, thereby avoiding impedance irregularities, 13-gage pairs were selected for the television circuits in all of the cables.

The length of the submarine cable under the Hackensack River (about 1100 feet) was too great to permit the use of regular carrier

[2] "Development and Application of Loading for Telephone Circuits," T. Shaw and W. Fondiller, *Journal A. I. E. E.*, Vol. XLV, pages 253–263, March, 1926.

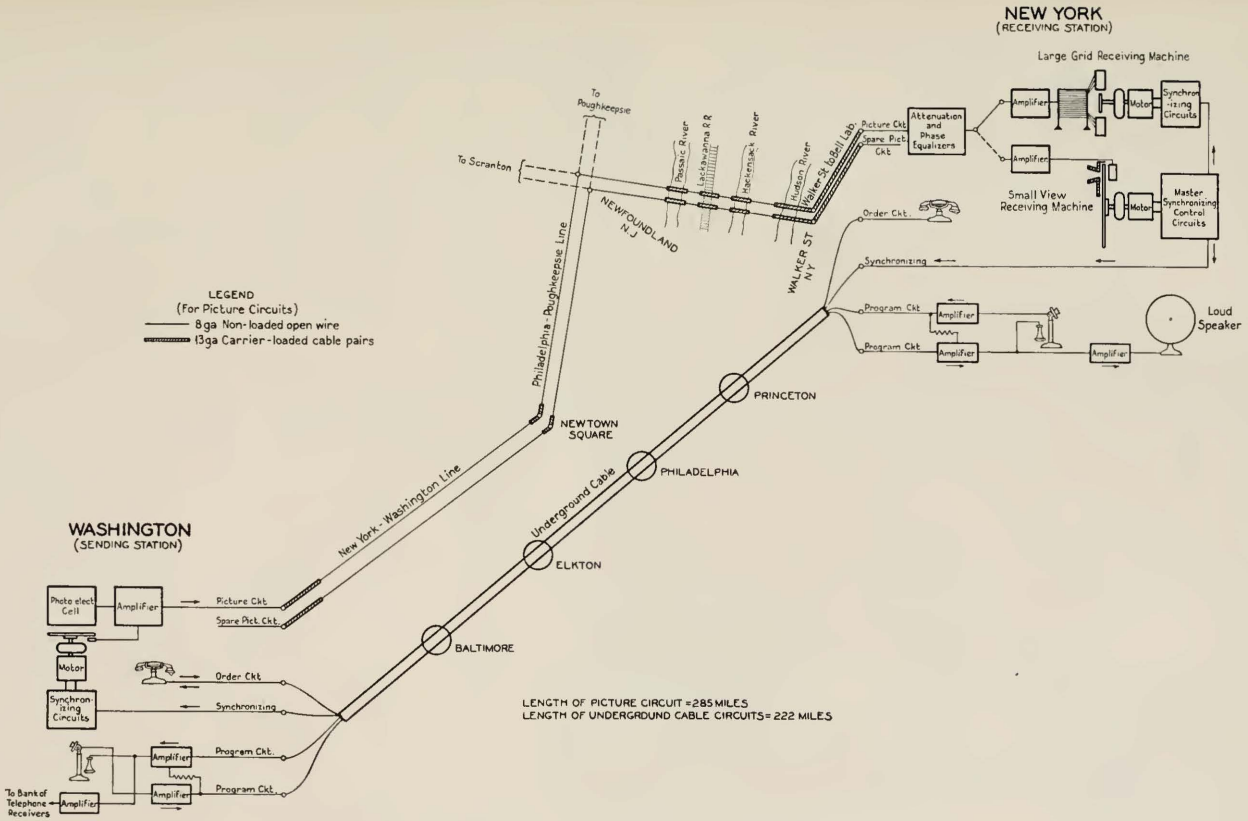


Fig. 2—Schematic diagram of circuits for television demonstration

loading, and a special loading arrangement having a slightly lower cut-off was, therefore, designed for this cable.

Equalization

Requirements. The requirements for the lines were stated earlier. In order to meet these overall requirements it was necessary to apply special forms of distortion-correcting networks.

Weather Changes. The above requirements applied, of course, to

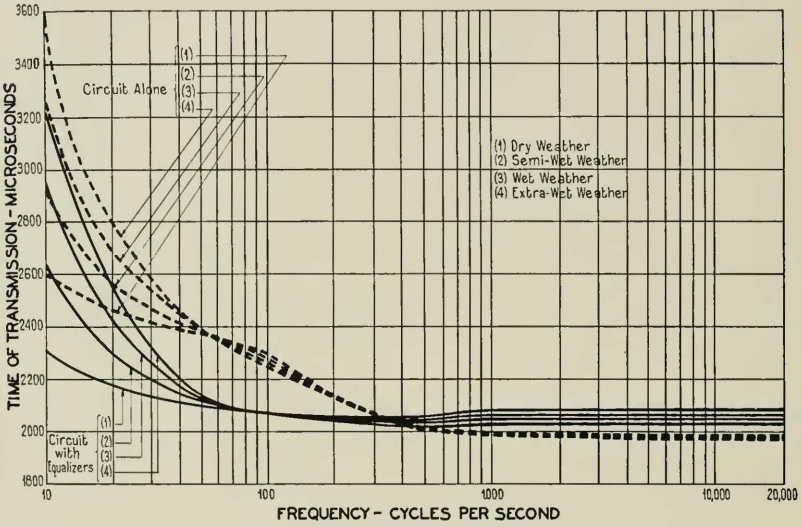


Fig. 3—Computed phase delay (β/ω) of television circuit with and without equalizers

all of the various weather conditions to which an open-wire circuit is subject. Due to the changes in the leakage conductance occurring at the insulators, the attenuation of an open-wire circuit varies with changing weather conditions. This change is particularly important at the higher frequencies. At 20,000 cycles, for example, the attenuation of a 165-mil open-wire pair may vary as much as 40 per cent for a change from dry weather to extra wet weather. For the circuit between Washington and New York this represents a possible attenuation change of about 10 T U, or a change of 10 to 1 in the magnitude of the received power. At 1000 cycles, the effect of wet weather is comparatively small, so that the net effect of the weather variations is to change the requirements for the attenuation equalizers. The phase shift introduced by an open-wire pair likewise varies to some extent with changes of weather, although the percentage variation is much smaller than in the case of the attenuation. In view of

these variations in the line characteristics it was decided to provide basic networks which would equalize for dry weather conditions, and to make available, in addition, several steps of equalization which would compensate for changes in the direction of wet weather.

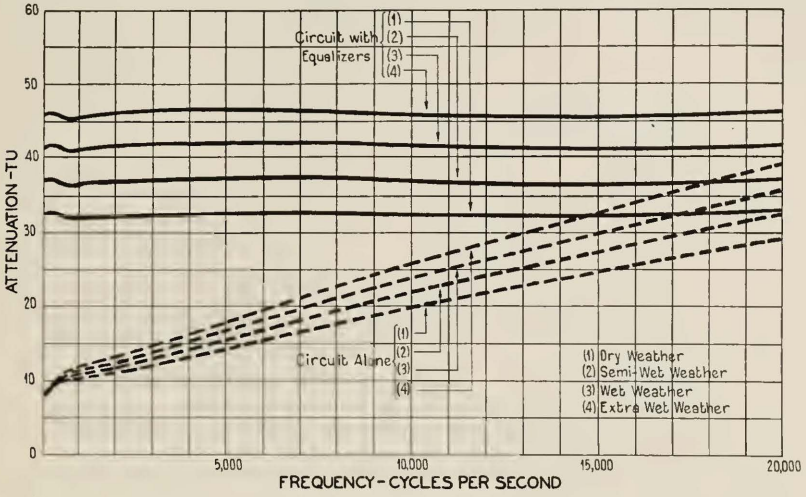
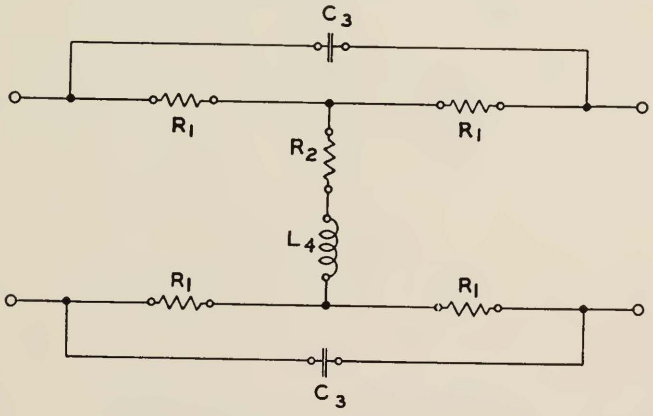


Fig. 4—Computed attenuation characteristics of television circuit with and without equalizers



$R_1 = 64.35$ OHM $\quad R_2 = 1334$ OHM
$C_3 = 6.112$ MF. $\quad L_4 = 1.100$ H.

Fig. 5—Low-frequency equalizing network (dry and wet weather)

Low-Frequency Network. Computed curves of attenuation and phase delay for the overall Washington-New York circuit without correcting networks are shown in Figs. 3 and 4, respectively. The

form of the dry weather attenuation curve suggested the use of two correcting networks, one for low frequencies, the other for high frequencies. The network which was designed to equalize the at-

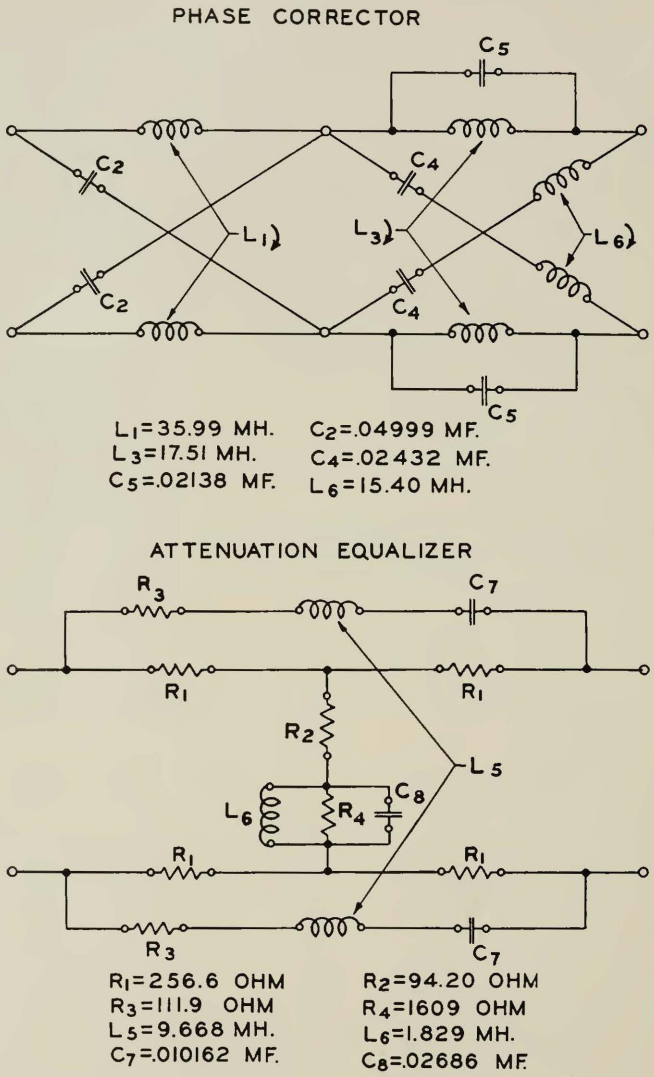


Fig. 6—High-frequency equalizing networks (dry weather)

tenuation at the lower frequencies is illustrated in Fig. 5. This network, in addition to equalizing the low-frequency attenuation, was made to provide sufficient correction for the low-frequency phase characteristic. It also proved satisfactory for all weather conditions.

High-Frequency Network for Dry Weather. The complete network for the correction at high frequencies under dry weather conditions was designed in two parts, an attenuation equalizer and a phaes corrector. These two structures are illustrated in Fig. 6. The

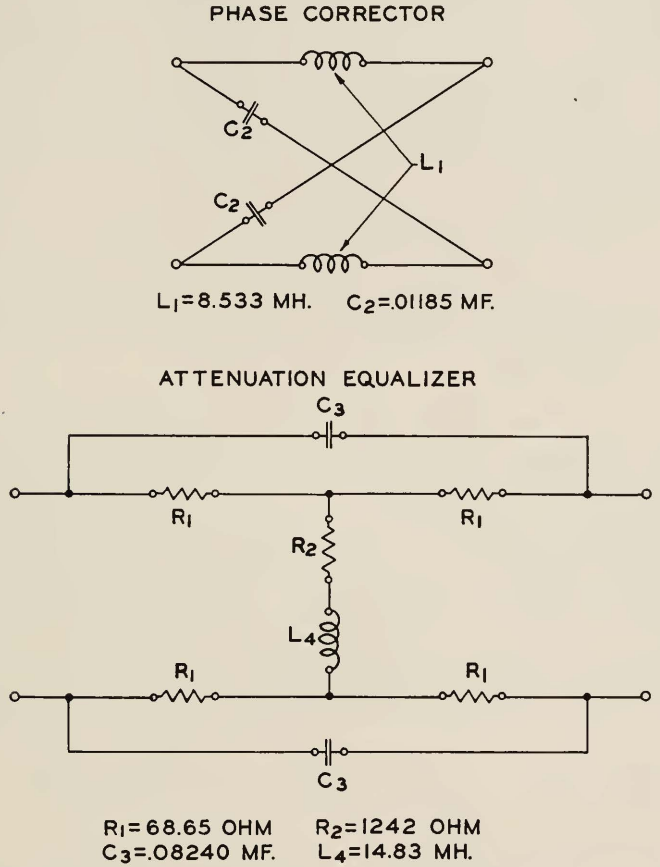


Fig. 7—Weather change equalizing networks

computed dry weather attenuation and phase delay resulting with the use of the combined low-frequency and high-frequency networks are illustrated in the curves of Figs. 3 and 4. It will be noted that the corrected attenuation curve is constant to within approximately $\pm$ 0.3 T U, while the corrected time of transmission falls well within the prescribed limits.

Weather Change Networks. Correction for the additional distortion introduced by changes from dry to wet weather was provided by three

additional networks which were, for convenience, of identical design. The results obtained by using one, two or three of these networks were made to correspond, respectively, to three assumed weather conditions which may be designated semi-wet, wet, and extra-wet. These three conditions were determined upon the basis of the range of leakage conditions which exist on open-wire lines under different weather conditions.

The attenuation equalizing and phase correcting networks for one of these steps are illustrated in Fig. 7, while the computed attenuation and phase delay obtained by the use of the three different steps of weather correction are shown in Figs. 3 and 4.

The networks described above are of the "constant-resistance" type, whose characteristic impedance is a pure resistance at all frequencies.[3] These networks are designed to be connected in series. The methods used in the design of the networks involve a large amount of mathematical theory, a discussion of which is not necessary for the purposes of this paper.

Synchronizing and Voice Circuits

So far the discussion has dealt only with the problem of transmitting the television currents. In addition to this, there is required the transmission of voice currents and of synchronizing currents. It is entirely feasible to transmit these currents together with the television currents over a single circuit. However, for the purpose of simplification, separate facilities were employed in the television experiments for picture, voice and synchronizing currents.

The diagram in Fig. 2 shows the circuits which were actually provided for the demonstrations. It will be seen that in addition to the two picture or television circuits, there were provided a synchronizing circuit, a four-wire "program" circuit, and an order circuit.

The method of synchronizing the sending and the receiving machines has already been described in the paper by Mr. Stoller. It requires two currents, one having a frequency of about 18 cycles and the other about 2125 cycles. In order that an ordinary telephone circuit might be used for this purpose, the lower frequency was made to modulate by means of a telegraph relay, a carrier current having a frequency of about 750 cycles per second. An amplifier-detector at the receiving end of the synchronizing system demodulated the 750-cycle current, delivering 18 cycles to the television apparatus.

The requirements for the synchronizing circuit were that it must

[3] Partially described in U. S. Patent No. 1,603,305 to O. J. Zobel.

transmit a narrow range near 750 cycles, and the single frequency of 2125 cycles. These synchronizing frequencies are determined by the speed of the motors, which was chosen so that the frequencies would be suitable for transmission over two channels of a voice-frequency carrier telegraph system,[4] but later it was found more convenient to use a separate telephone circuit.

The circuits labeled "program" provided telephonic communication between the observer at New York and the person being viewed at Washington. A loud speaker was also connected to this circuit at New York to transmit the voice to the audience when the large grid receiving arrangement was employed. A special by-passing connection was provided between the amplifiers at the terminals of the circuit so that speech from the local microphone could be heard as well as speech from the distant city.

The order circuit was for the purpose of providing communication between the engineers operating the television apparatus.

Line Measurements

In order to determine that the circuits set up as outlined above were satisfactory, their overall characteristics were measured. Certain matters of interest in this work are noted below.

Measurements of Envelope Delay. In order to measure the envelope delay to an accuracy comparable to the requirements for the lines, it was necessary to develop special apparatus. Fig. 8 shows in schematic form the circuits of the apparatus designed for this purpose. The apparatus measures not the absolute envelope delay of a circuit, but the relative delay of one circuit at any frequency from about 600 cycles to 20,000 cycles or more with respect to the delay on the other circuit at a fixed frequency.

The functioning of the apparatus may be briefly described as follows: Simultaneously into each line there was transmitted a carrier current, each carrier being modulated by 250-cycle current from the same oscillator. The modulation was accomplished in push-pull vacuum tube circuits so that the undesired products of modulation were eliminated by balance. The carrier on the line under measurement was adjusted to the frequency at which a measurement was desired, and the carrier on the other circuit, used for reference, was kept at a fixed frequency of 5100 cycles.

At the receiving point identical circuits were provided for amplifying

[4] "Voice-Frequency Carrier Telegraph System for Cables," B. P. Hamilton, H. Nyquist, M. B. Long and W. A. Phelps, *Journal A. I. E. E.*, Vol. XLIV, pages 213–218, March, 1925.

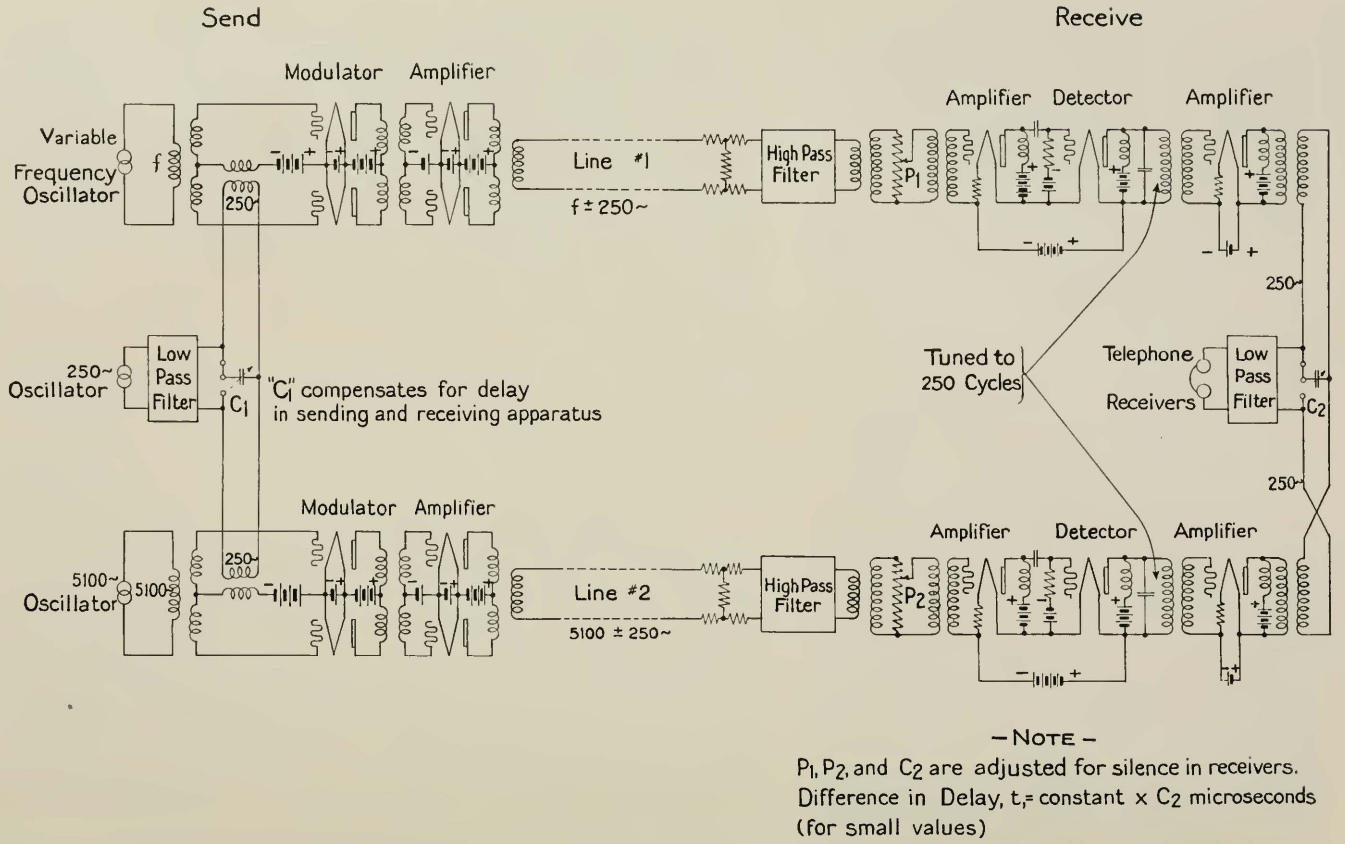


Fig. 8—Arrangements for measuring envelope delay of television circuits

and demodulating the received currents from the two circuits. The 250-cycle outputs from the two sets of receiving apparatus were connected in opposition to a pair of telephone receivers through a low-pass filter. Potentiometers P_1 and P_2 were provided for adjusting the relative intensities of the two 250-cycle output voltages and a condenser C_2 was arranged so that it could be used to change the phase of either of the 250-cycle voltages. It is evident, then, that

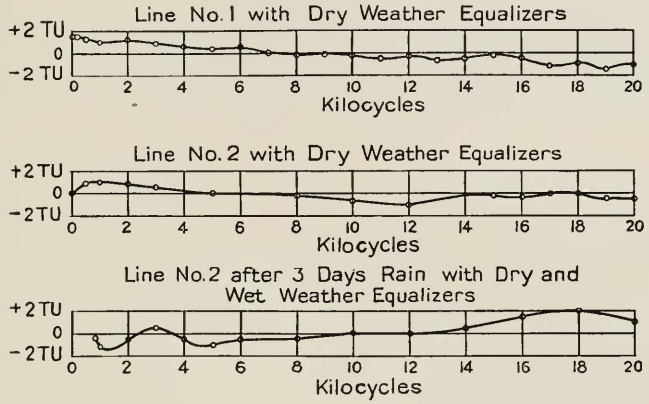


Fig. 9—Measured attenuation characteristics of line circuits plus equalizers

by making suitable adjustments the two voltages could be adjusted to exactly the same intensity and opposite phase so that no sound is heard in the telephone receivers. As long as the value of C_2 is small, the envelope delay of one line at the carrier frequency with respect to the delay of the other line at 5100 cycles is proportional to the value of C_2.

The condenser C_1 shown at the sending station is for the purpose of introducing a phase shift in the 250-cycle current of either channel relative to the other in order to compensate for the differences in delay of the apparatus itself at the two frequencies. The value of C_1 was determined by experiment before moving the sending apparatus to Washington and was adjusted to its calibrated value for each frequency when the oscillator frequency was adjusted.

The measurement of the phase shift of the 250-cycle current, which is transmitted by means of a carrier over a circuit as described above, is actually a measurement of the difference between the phases of the two received side-band currents situated 250 cycles either side of the carrier. The envelope delay is equal to $\Delta\beta/\Delta\omega$ where $\Delta\omega$ equals 2π times 500, and $\Delta\beta$ equals the measured difference in phase of the two side bands in radians.

Measurements and Performance. How well the requirements which were set up earlier were met by the lines and the distortion-correcting networks is shown in Figs. 9 and 10. The attenuation characteristics

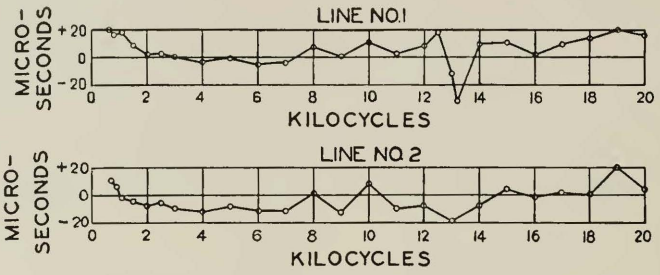


Fig. 10—Measured envelope delay of line circuits plus equalizers

are well within the established limits, and the phase characteristics show only a single slight departure for one circuit in a very narrow range of frequency. It is of interest, in view of the fact that the distortion-correcting networks were designed and built before any measurements were made on the lines they were to fit, that no changes or adjustments were found to be necessary in the networks, in order to obtain these characteristics.

Comparison of the television images obtained from transmission over the line with those obtained from transmission from one side of the room to the other, showed that no difference in quality could be observed.

Radio Transmission System for Television[1]

By EDWARD L. NELSON

Synopsis: Starting from the general requirements imposed on the transmitting medium, this paper discusses the engineering of a radio system for television purposes and describes the radio facilities actually employed for the recent Bell System demonstration. The tests to which the system was submitted to determine its suitability are outlined and the measured frequency-response characteristics are shown. An interesting phenomenon due to multi-path transmission, the production of positive and negative secondary images, is reported. A brief series of experiments concerned with the transmission of both voice and image "on a single wavelength" is also described.

IN other papers of this symposium, the general nature of the television problem has been discussed, the scope of the recent Bell System demonstration has been outlined, terminal apparatus for television has been described, and the general requirements to be met by the transmitting agency have been formulated. This paper is concerned with the problem of engineering a suitable radio system for television purposes and with a description of the radio facilities actually employed for the demonstration.

Requirements Imposed on the Radio System

The radio experiments were conducted from the Bell Telephone Laboratories' Experimental Station 3XN at Whippany, New Jersey. Between this point and the main Laboratories Building at 463 West Street, New York City, some 22 miles distant, three separate communication channels were required—one for the picture, a second for synchronizing, and a third for speech and music. The demonstration being of a three-cornered nature involving New York, Washington and Whippany, it was deemed to be highly advantageous to transmit the necessary synchronizing currents for both the wire and radio systems from a master generating set located in the auditorium of the West Street Building. Hence the synchronizing channel was required to operate from New York to Whippany, while the picture and speech channels necessarily transmitted in the reverse direction.

From the radio standpoint, the problem presented for solution may be described as follows:

1. There is given television transmitting and receiving apparatus designed to work into and out of specified impedances at stated signal

[1] Presented at the Summer Convention of the A. I. E. E., Detroit, Mich., June 20–24, 1927.

energy levels. Signal components ranging in frequency from 10 to 20,000 cycles must be transmitted with as little discrimination with respect to either amplitude or phase as reasonable design practices will permit. It is required that a suitable radio system be designed to afford satisfactory transmission between terminals when operated under prevailing conditions with respect to static, other radio traffic, and local electrical disturbances. The maximum allowable "noise" level is probably somewhat arbitrary but it has been found that if the ratio of signal to interference current is 10:1 the results are satisfactory. The variation of amplitude with frequency should probably not exceed $\pm$ 2 TU at any point in the required signal band. The equivalent of the circuit must be substantially constant; in other words, no fading effects can be tolerated. In this respect a variation of perhaps 3 TU is the maximum allowable.[2]

2. For synchronizing purposes, a second channel must be provided to transmit 17.7 and 2125 cycles, the impedances and the signal energy levels at both ends of the circuit being known. The grade of transmission required in this case is probably considerably lower than that needed for the picture circuit but stable operation must be assured.

3. Arrangements must also be made for a high quality telephone channel to transmit speech and music for loud speaker reproduction.

4. All of these channels must, of course, be capable of operating simultaneously without mutual interference and without effect on established radio services.

Preliminary Survey

In the vicinity of New York, an assignment of this type is surrounded with unusual difficulty due to the serious congestion which exists in the ether. Operations were started, therefore, by undertaking a survey of available frequency bands at periods of the day during which transmission might be required.

The pioneering nature of the project and the character of the apparatus available led to an early decision to base the system on the transmission of the carrier and both sidebands. Since the upper limit for the signal was specified as 20,000 cycles, an interference-free band somewhat greater than 40,000 cycles in width was, therefore, required. The unusual width of this band indicated the desirability of fixing upon a relatively high carrier frequency. No readily available

[2] Definitely agreed on limits were essential to proper coordination of the various development activities and figures of the order mentioned were assumed for design purposes.

substitute for the ordinary type of tuned circuit was at hand and such circuits discriminate seriously against side frequencies differing by more than a few per cent from the frequency to which they are adjusted.

The results of the survey disclosed two bands somewhat wider than that required centering approximately about 1575 and 1750 kilocycles. It was also conclusively demonstrated that the operation of the synchronizing channel at a frequency above the broadcasting



Fig. 1—General view of the Whippany Station, 3XN

band was entirely out of the question. With two broadcasting stations located in the immediate neighborhood, one producing a field strength of perhaps 50 millivolts per meter and the other several volts per meter, the operation of a third transmitter on an adjoining frequency with the maximum obtainable separation between antennae, resulted in an almost continuous interference spectrum. It was decided, therefore, to transfer the synchronizing channel to a frequency of the order of 185 kilocycles, which would be sufficiently remote to remove interference from this source, and to make further studies in the regions about 1575 and 1750 kilocycles based on transmission from Whippany. No difficulty was anticipated in making suitable arrangements for the speech channel on account of the narrower band required and the well-known nature of the problem.

The Whippany Station, 3XN

A general view of the station site at Whippany is shown in Fig. 1. The property consists of some 47 acres. The main laboratory building, which is located near its center, is a two-story structure affording approximately 18,000 square feet of floor space. The principal antenna system involves two 250-foot steel towers with a suitable buried ground system, which is placed some 500 feet out in front of the building in order that the latter may be clear of the denser portion



Fig. 2—Operating room at 3XN. Transmitter for television channel on the right. Power supply unit and radio transmitter for the speech channel in the center and on the left, respectively.

of the electric field. This antenna was assigned to the picture channel. For the voice channel, a separate structure located 500 feet in the rear of the laboratory building or approximately 1000 feet from the other was employed. The original supports in this second case were 60-foot wooden masts but subsequently metal topmasts were added, bringing the total height to 100 feet. Both antennae were energized by means of radio-frequency transmission lines. The antenna tuning and coupling apparatus was housed in small buildings placed under the center of each antenna, that for the larger structure having a copper roof which was securely connected to the ground network.

This type of installation is thought to afford a number of advantages. By separating the building and the antenna it becomes a much simpler matter to control the electrical factors which enter into the design of the latter. Removing the building from the field tends toward

reduced dielectric and eddy current losses and consequently toward higher antenna efficiency. The resulting improvement may be expected to more than compensate for the slight loss in the line, which should not exceed 3 per cent. Removing the field from the building is equally advantageous in that it simplifies the precautions which normally have to be taken to prevent the radio-frequency energy from affecting the performance of audio amplifiers and other supple-



Fig. 3—Television transmitting apparatus in the studio at Whippany

mentary vacuum tube apparatus. The most serious disadvantages arise from the fact that the antenna must be tuned and the current in it measured at a point remote from the transmitting apparatus proper.

In spite of the fact that the station building was not directly under either antenna, some difficulty was anticipated from radio-frequency fields produced within the transmitting equipment due to the relatively high amplification employed with the photoelectric cells. In order to minimize trouble of this nature a special shielded studio was con-

structed in one of the wings of the building to house the television terminal apparatus. Walls, ceiling and floor were completely covered with No. 24 gage sheet copper lapped about one inch and carefully soldered. The windows were covered with fine copper gauze. The door was covered with sheet copper which was carried around the edges so that in closing it made a firm wiping contact with the surrounding frame. Circuits for lighting and miscellaneous power service were led in through two specially constructed transformers fitted with grounded copper shields between the primary and secondary windings. The picture circuits leading to the radio transmitter, the microphone circuits, and the necessary studio signal and control circuits were run in lead cable and in most cases were brought into the room through suitable radio-frequency filters enclosed in metal boxes attached to the copper sheathing. In order to avoid the possibility of the heavy current leads to the arc bringing in radio-frequency energy, and to eliminate the noise and heat from the arc, provision was made for mounting the latter in its metal cabinet outside of the room. The circular opening through which the light beam was projected into the room was protected by the lamp cabinet which was also grounded to the sheathing. Satisfactory acoustic conditions within the studio were obtained by applying celotex wall board over the copper and by the use of suitable floor coverings.

Transmitting and Receiving Apparatus

For the television channel, arrangements were made to install a standard Western Electric 5-B Radio Broadcasting Transmitter and to modify it for the purpose. This transmitter is a 5-kilowatt unit (carrier output without modulation) designed for high quality telephone transmission in the 500–1500-kilocycle band. It will transmit signal components ranging from 50 to 5000 cycles without noteworthy discrimination. At 30 cycles and at 10,000 cycles there is some loss in efficiency and beyond these points the characteristic curve falls rapidly. The necessary changes, therefore, involved both the radio and audio circuits, the latter phase of the problem being perhaps the more difficult.

The schematic circuit of the modified transmitter is shown in Fig. 4. The revised radio frequency circuits were very similar to the standard arrangement, the changes mainly affecting the magnitudes of various coils and condensers. The output circuits were, of course, redesigned to meet the conditions imposed by the transmission line. The circuit was of the master oscillator—modulating amplifier—power amplifier type. The master oscillator employed a 50-watt tube operating in a

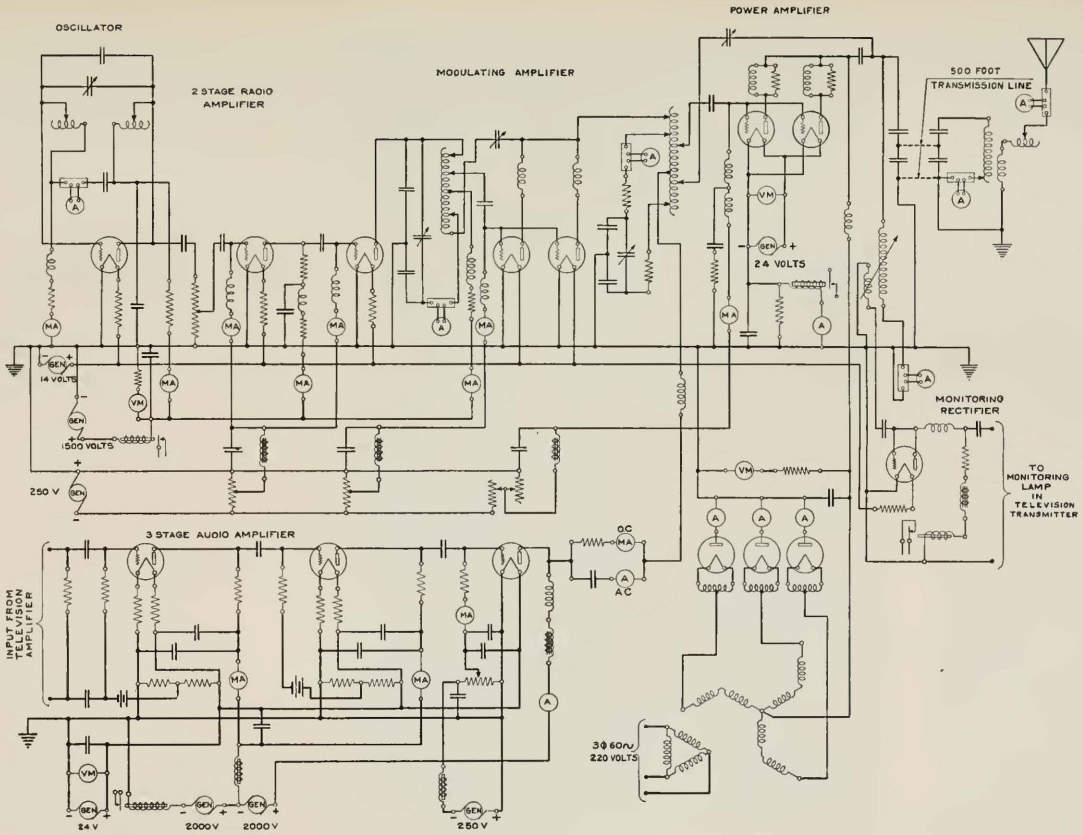


Fig. 4—Schematic of radio transmitter for television channel

circuit designed to afford a high degree of stability. This was connected to the input of the modulating amplifier through two radio-frequency stages, also employing 50-watt tubes. These two stages precluded the possibility of the oscillator frequency being appreciably altered by effects due to modulation. The modulating amplifier employed two 250-watt tubes in parallel and operated on the Heising system. In the standard equipment, the audio stages involve one 50-watt tube and two 250-watt tubes in parallel. To meet the more rigorous requirements of television with an ample factor of safety, this portion of the transmitter was removed from service and a specially constructed three-stage amplifier was substituted. As shown in the drawing, the latter consisted of two 50-watt resistance-coupled stages and a final power stage based on a 5-kilowatt water-cooled tube which raised the signal currents to a power level of approximately one half kilowatt.

In order that it might be possible to check the performance of the radio transmitter under all operating conditions, a suitable monitoring rectifier was constructed and coupled to the output circuit of the radio-frequency power amplifier. A circuit was run back to suitable switches on the television control panel so that either the output of the photo-electric cell amplifiers or the rectified output of the radio transmitter could be impressed on the pilot lamp of the television transmitter. By comparing the two images, it thus became a relatively simple matter to detect any serious maladjustment in the radio apparatus.

The problem of providing a suitable transmitter for the speech channel was rendered quite simple by the fact that at the time there was in process of development at Whippany a 50-kilowatt equipment intended for broadcasting applications. The detailed description of this transmitter is beyond the scope of the present paper. It may be said, however, that it consists of a piezo-electrically controlled master oscillator employing a 50-watt tube directly followed by a 50-watt modulating amplifier. Modulation is by the Heising system, employing one 50-watt and one 250-watt tube in the audio stages. The output of the modulating amplifier is amplified by three push-pull, neutralized, radio-frequency stages the last of which employs six water-cooled tubes at approximately 17,000 volts. This set is capable of delivering 50 kilowatts (unmodulated carrier) to the antenna and during modulation instantaneous peaks approaching 200 kilowatts are attained.

The radio receiver employed at Whippany for the reception of the synchronizing signals at 185 kilocycles presents no features of unusual interest. A double-tuned input circuit was used followed by three

stages of radio-frequency amplification, a detector, and two audio stages of conventional design employing transformer coupling. No serious difficulty was encountered in obtaining ample selectivity to insure satisfactory operation in the face of the strong local signals but care was necessary in locating the receiver and in laying out the antenna in order to avoid the inductive type of interference which is almost always experienced in the immediate vicinity of a large radio station. The receiving antenna was located approximately 700 feet from the two transmitting radiating systems.

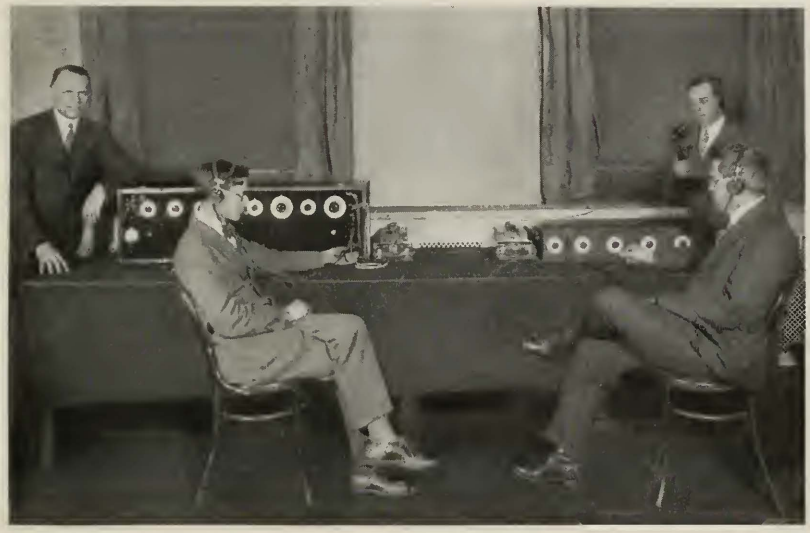


Fig. 5—Radio receiving equipment for the television and speech channels in the auditorium of Bell Telephone Laboratories, New York

The receiver employed at the New York terminus of the television channel presented a somewhat knotty problem on account of the relatively wide frequency band which it was required to pass while providing the maximum discrimination against interference. The width of the required band pointed very definitely toward the superheterodyne. This type of circuit is also very stable, permits of all the amplification that may be needed or that may be employed under ordinary noise conditions, and is very selective against interference immediately adjacent to the desired band. It is quite susceptible, however, to interference from components differing from the desired carrier frequency by an amount approximately equal to the intermediate frequency. If the interfering component lies in the neighborhood of the frequency of the oscillator, beats will be produced which

may or may not pass the intermediate-frequency amplifier and the associated filters depending on their design. If the interfering component lies on the opposite side of the wanted carrier from the oscillator and differs from the former by the intermediate frequency, it will be passed by the receiver, subject only to the attenuation due to the radio-frequency circuits (the input circuits tuned to the wanted

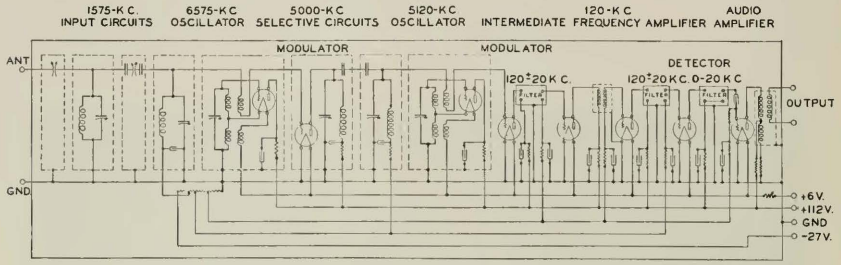


Fig. 6—Schematic of triple detection receiver for television channel

carrier). This characteristic must be given careful consideration in the design of selective receivers of the superheterodyne type and has led to the introduction of carefully designed, loosely coupled, input circuits or an initial tuned radio-frequency stage for this purpose. Neither of these expedients were possibilities in the television receiver, however, because of the extraordinary width of the required transmission band. Recourse was had, therefore, to a triple detection arrangement. Speaking somewhat in the vernacular, the desired signal was "beat up" to 5000 kilocycles where it was passed through sharply tuned coupled circuits, then "beat down" to 120 kilocycles, amplified, filtered and rectified, finally passing through a suitable low pass filter, audio amplifier and output transformer to the television reproducing apparatus.

The circuit arrangement is shown schematically in Fig. 6. Two tuned circuits with capacity coupling were connected to the input of the first detector or modulator. A relatively tight coupling was employed to produce the well-known double-peaked resonance curve capable of affording the required band width. The antenna was not tuned but was loosely coupled to the selective circuits by means of an adjustable capacity. The incoming radio signal was impressed upon the grid of the modulator tube along with a suitable voltage from an oscillator operating at 6575 kilocycles. The 5000-kilocycle components which resulted were selected by means of two carefully designed tuned circuits also capacity coupled. The purpose of this stage in the process will be evident if it is appreciated that at 1575

kilocycles, ± 20 kilocycles represents a 2.6 per cent band while at 5000 kilocycles the same side frequencies represent only a 0.8 per cent band. In the latter case, therefore, it is possible to employ materially sharper circuits without discriminating against the higher signal components. The 5000-kilocycle circuits connected to the grid of a second detector or modulator tube upon which suitable voltages from a 5120-kilocycle oscillator were impressed. The 120-kilocycle components in the output of this modulator were selected by means of a band-pass filter which worked into a two-stage intermediate-frequency amplifier. A second band-pass filter led to the third or final detector. A 20-kilocycle low-pass filter was employed in the plate circuit of the latter. This filter was designed for a low input impedance at 120 kilocycles in order to meet the necessary condition for efficient rectifier action and it also served as a coupling element for the audio stage which followed. A special output transformer with a permalloy core was provided to step down to the relatively low impedance of the line leading to the television apparatus proper.

A superheterodyne receiver of more conventional design was employed for the speech receiver. The circuit arrangement involved a double-tuned input circuit, one tuned radio-frequency stage, oscillator and modulator, two intermediate-frequency stages, detector and one audio stage. It was highly selective and afforded substantially distortionless transmission for signal frequencies ranging from 50 to 5500 cycles.

The transmitting equipment for the synchronizing channel consisted of a Western Electric 6-A Radio Broadcasting Transmitter modified to operate at 185 kilocycles. In order to avoid the necessity of transmitting directly the 17.7-cycle component required for synchronizing purposes, a 760-cycle carrier was modulated at 17.7 cycles by means of a relay and impressed upon the input of the radio transmitter together with the steady 2125-cycle component. At the receiving end, the 2125- and modulated 760-cycle components were separated by means of suitable filters, and the latter rectified to produce the desired 17.7-cycle current.

Tests of the System

As soon as the various apparatus units could be made ready for service, a comprehensive series of transmission tests was undertaken. In order to determine the relative suitability of the 1575- and 1750-kilocycle bands disclosed by the preliminary survey, transmissions from Whippany at intervals throughout the day were arranged. Field strength measurements were taken at the receiving point

employing apparatus of the type described by Englund and Friis[3] and observations on the relative strength of the received signals were made by inserting a sensitive microammeter in the plate circuit of the third detector of the television receiver. These data indicated that the lower frequency band suffered considerably less attenuation and also afforded much more stable transmission. In spite of the comparatively short distance (approximately 22 miles), marked fading was experienced beginning with the sunset period and increasing in amplitude as the night advanced. The high frequency band proved to be particularly disadvantageous in this respect. It was decided, therefore, to fix upon the lower frequency band and to confine the demonstration to the afternoon when reasonably stable transmission conditions prevailed.

Following the choice of a definite operating frequency, a number of modifications were made in the transmitting antenna to improve its efficiency and increase the field strength at the receiver. This work finally resulted in a measured field strength of approximately 2500 microvolts per meter for an antenna input of 5 kilowatts.

Further consideration of the available data on transmission and traffic conditions and the performance characteristics of the apparatus units involved lead to a choice of 1450 kilocycles for the speech channel. In spite of an antenna input of approximately 30 kilowatts, the initial tests at this frequency were very unsatisfactory due to inadequate field strength at the receiver which necessarily resulted in an abnormally high noise level. The height of the antenna was, therefore, increased from 60 to 100 feet by installing iron pipe topmasts. This change brought the field strength at the receiver to approximately the same value as that obtained for the television channel (2500 microvolts per meter) which was considered to be satisfactory for the purpose.

In order to insure that the reproduction of the picture might not suffer from serious discrimination against essential frequencies at some point in the radio system, very careful tests were made on the individual units and on the system as a whole.

The frequency characteristic of the transmitter was determined by connecting a vacuum tube oscillator producing a relatively pure wave to its input terminals through a suitable network involving a thermal milliammeter and an adjustable artificial line. A rectifier of known characteristics and a second thermal meter protected against radio-frequency currents by means of a low-pass filter were coupled to the

[3] "Methods for Measurements of Radio Field Strengths," C. R. Englund and H. T. Friis. Presented to the Spring Convention A. I. E. E. at Pittsfield, Mass., May 25, 1927.

output circuit of the water-cooled tubes. Employing a frequency of 1000 cycles, the input was adjusted to produce normal modulation and the readings of the input and output meters noted. The oscillator frequency was then changed by a convenient amount while holding the input reading constant and the artificial line readjusted, if necessary, to produce constant output current. Under these conditions, any change in the setting of the artificial line indicates an equal variation in the transmission efficiency of the transmitter which is evaluated by this method directly in TU.

The characteristic of the receiver was determined in a similar manner. A low power transmitter of known characteristics was connected to it through a suitable attenuating network which, in so far as the receiver was concerned, simulated the receiving antenna. The radio-frequency input to the receiver was adjusted to approximately the normal value and a series of measurements taken with variable audio-frequency inputs as indicated above.

The overall measurements were also based on a similar procedure impressing a constant input on the 600-ohm input terminals of the transmitter through a suitable artificial line and adjusting the latter to give a constant current into a 600-ohm load at the output of the receiver, taking necessary precautions, of course, to preclude overloading at any point in the system.

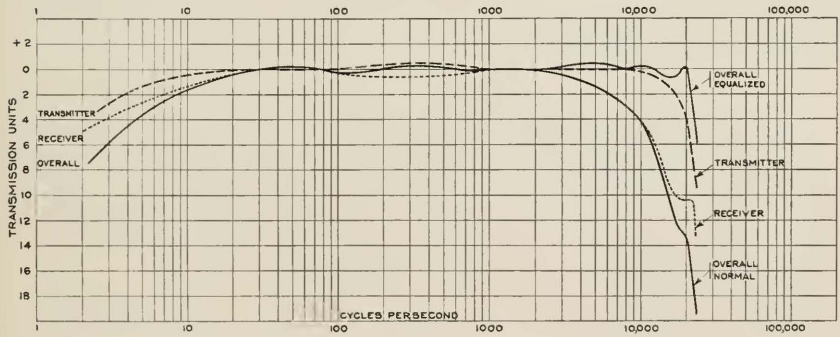


Fig. 7—Measured characteristics of television channel

The experimental characteristic curves thus obtained are shown in Fig. 7, where the abscissæ represent cycles per second and the ordinates departure from the 1000-cycle value in TU. As will be noted, at the lower frequencies exceptionally good performance was obtained, the overall characteristic being only 2 TU down (or deficient) at 10 cycles and only 6 TU down at 3 cycles. The results for the higher frequencies, however, were not so satisfactory, a loss of approximately

13 TU being observed at 20,000 cycles, probably due to the tuned circuits in the receiver. Since modification of these circuits to obtain a flatter characteristic would have been difficult and would have occasioned a noteworthy sacrifice in selectivity, a compensation network was designed for use in the 600-ohm output circuit of the receiver which introduced a negligible loss at 20,000 cycles, a substantially constant loss of 13 TU at frequencies below 2000 cycles, and for intermediate frequencies losses represented by the height of the "normal overall" curve above the horizontal line representing − 13 TU. With this network connected between the receiver and the television equipment, the average level throughout the band was, therefore, reduced some 13 TU but the resulting characteristic as measured beyond the network was that which has been designated "overall equalized." Above 20,000 cycles the characteristics all fell very rapidly which is an indication of the degree of selectivity attained. This was contributed to by the radio-frequency tuned circuits, the band-pass filters in the intermediate-frequency amplifier and the 20,000-cycle low-pass filter between the final detector and audio amplifier. The individual characteristics of the various filters were designed to be 60 TU down 20 kilocycles from the specified cut-off frequency.

Similar measurements were made upon the speech channel but a less thorough study was deemed sufficient in that case due to the existing background of experience.

Effects of Fading

With the system as outlined above, very satisfactory performance was obtained during the afternoon and early evening hours when reasonably stable transmission conditions were prevalent. Later at night, however, when marked fading became evident, some rather unexpected but easily explainable phenomena were observed which may be of sufficient interest to warrant brief mention.

When marked fading occurred, the normally clear reproduction was accompanied by "ghosts" or additional images which faded in and out in an erratic manner, sometimes appearing as positives and sometimes as negatives. The effect was most clearly observed when using one of the various types of test screens employed, a white card bearing a black diamond-shaped outline, approximately a square with its diagonals vertical and horizontal. With this simple type of pattern, it became evident that the secondary images were additional reproductions which were "out of frame" by a greater or less amount. In other words, each of these additional images consisted of a portion

of two diamonds placed side by side with the corners just touching. Images "out of frame" along the vertical axis are frequently seen on the motion picture screen.

The explanation is fairly obvious. The present more or less generally accepted view of fading is that it is a manifestation of transmission along two or more paths, at least one of which is variable, producing a continually changing phase relationship between the components and a corresponding waxing and waning of the resultant signal. In the present case, the major image was probably produced by the so-called "ground wave." The secondary images probably resulted from components which were transmitted upward at a relatively sharp angle and turned back to the receiving station from the Heaviside layer, the difference in framing being due to the longer time of transmission.

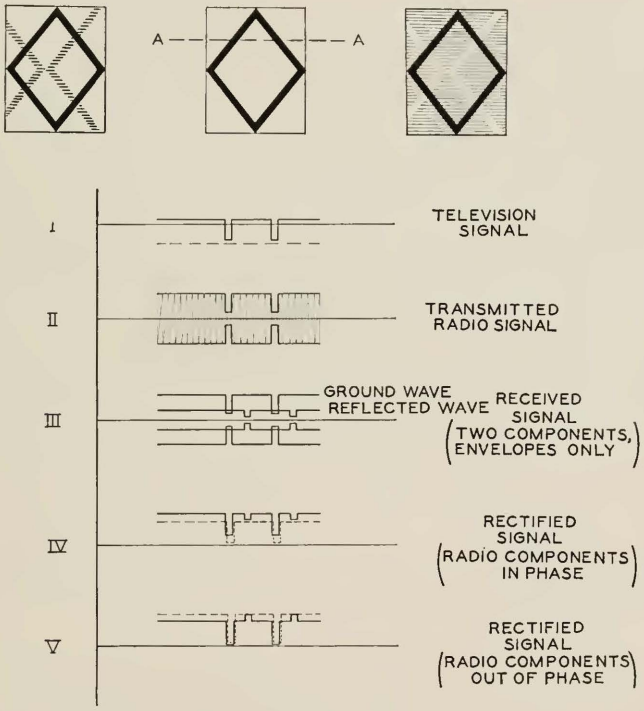


Fig. 8—Production of positive and negative secondary images due to multi-path transmission

The production of negative secondary images is a most interesting phase of the phenomena. This effect may be explained quite easily by means of a series of signal diagrams such as is shown in Fig. 8.

If attention is confined to the interval during which scanning takes place along the line AA, it is evident that the television signal will have the form shown. Amplitudes above the dotted line indicate the current through the photoelectric cell. Since transformer-coupled amplifiers are employed in the television apparatus, however, the direct component is eliminated and the zero axis for the input to the radio transmitter is the solid line. Sketch II shows the modulated output of the radio transmitter. The received signal, shown in III, is assumed to consist of two components, the larger due to the "ground wave," and the smaller due to reflected energy from the Heaviside layer. The latter lags somewhat because of the greater length of the transmission path. The resultant of these two components will necessarily depend on the relative phase of the two carriers at the receiving point. Two cases are considered: when the components are exactly in phase, and when they are exactly out of phase. The effect at intermediate positions may be readily evaluated from these examples. With the components in phase, the detector output is proportional to their sum which is shown in IV. It is evident that this will result in a major image and a secondary positive image. If the components are out of phase, the rectified signal shown in V results. It is simply a matter of subtracting amplitudes. This resultant consists of the desired signal with the amplitude somewhat reduced which will produce a gray background. The secondary image will be formed by the two small peaks shown and will be lighter than the background, in other words a negative.

A pattern frequently observed was the diamond with a cross through its center due to a secondary image. This represents a change in framing of approximately one half line. With 17.7 pictures per second and 50 lines per picture, this corresponds to a difference in transmission time of $1/17.7 \times 1/50 \times 1/2$ or 5.65×10^{-4} seconds. A rough computation of the height of the reflecting layer based on this figure and a distance of 22 miles between transmitting and receiving stations gives 100 kilometers, which is substantially in agreement with determinations made by other methods.

Transmission of Voice and Image with a Common Carrier Frequency

Following the demonstration, a brief series of supplementary tests was arranged to obtain some appreciation on experimental grounds of the problems involved in transmitting both voice and image with a single radio transmitter. The system employed may be considered as the extension of carrier current technique to radio, but has been

described in various other terms: "multiplex radio," "double modulation," "the Hammond system," etc. The output of a 30,000-cycle oscillator was modulated with the speech signal. The resulting carrier and sidebands were selected by means of a suitable filter passing frequency components ranging between 25,000 and 35,000 cycles and impressed on the input terminals of the radio transmitter along with the 10 to 20,000-cycle signal from the television apparatus. A suitable low-pass filter was employed in the line to the latter in order to preclude "crosstalk" due to 25,000–35,000-cycle energy working back into the final amplifier stages. The input to the radio transmitter thus consisted of a band extending from 10 to 20,000 cycles together with a 25,000 to 35,000 band, with a particularly strong component at 30,000 cycles representing the low-frequency carrier.

In order that it might be capable of handling this wider band without discrimination, further modifications in the radio transmitter were required. In the case of some of the radio-frequency circuits, which were required to pass a 70,000-cycle band, it was found to be necessary to insert resistance to reduce the sharpness of resonance. On account of lack of time, it was not possible to obtain a complete series of characteristic curves for the transmitter under these conditions. Isolated measurements with a single-frequency input of 35,000 cycles indicated, however, that components of this order could be transmitted without serious loss and the subsequent performance of the system as a whole confirmed this conclusion.

It is well known that if a sinusoidal alternating current $i = I_0 \sin \omega t$ is modulated with a signal of frequency $f = \Phi/2\pi$, the resulting modulated current may be represented by the expression:

$$i = I_0 \sin \omega t + \frac{kI_0}{2} \sin (\omega + \Phi)t + \frac{kI_0}{2} \sin (\omega - \Phi)t,$$

where k is a fraction indicative of the degree of modulation. In other words, a modulated current, or wave, may be resolved into three components: (1) a steady component, known as the "carrier," which has the amplitude and frequency of the original unmodulated current, (2) an "upper sideband" which is equivalent to the signal spectrum with each individual frequency increased by an amount equal to the carrier frequency, and (3) a "lower sideband" which is an inverted reproduction of the signal spectrum, that is, each individual signal component is laid off in the downward direction from the carrier frequency, or subtracted from it. Hence, assuming a carrier frequency of 1575 kilocycles and a signal input to the radio transmitter

of the type described above, the antenna current, or the transmitted wave, may be represented diagrammatically as shown in Fig. 9.

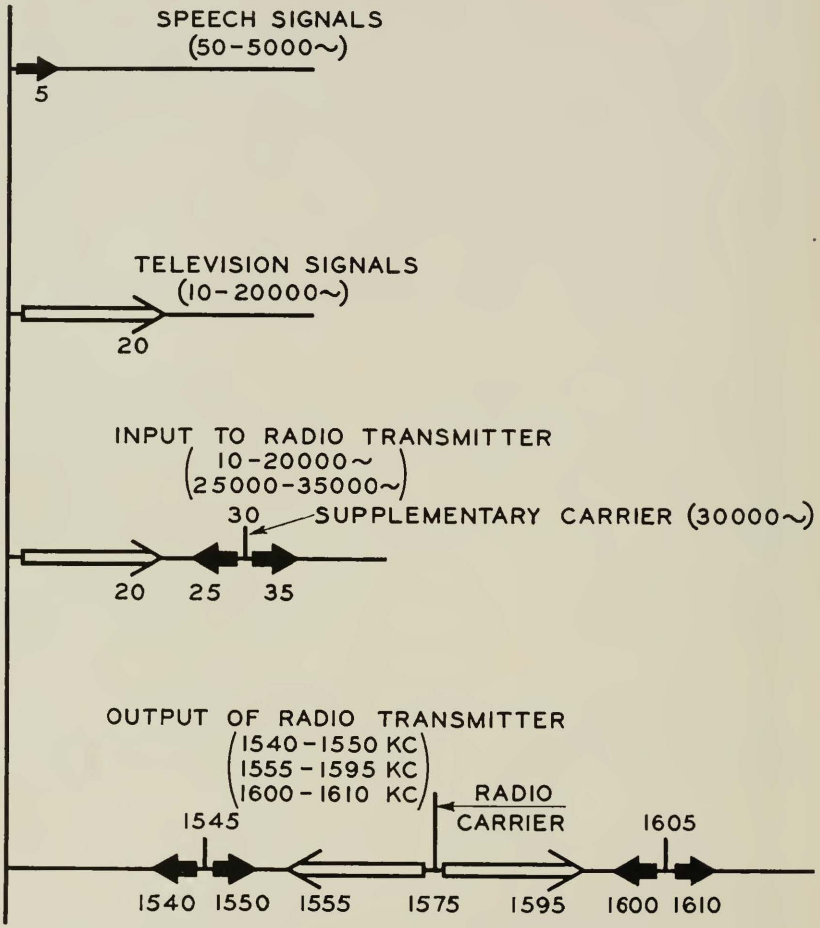


Fig. 9—Diagrammatical representation of frequency conversions in multiplex radio system

It is evident that this type of radio signal can be received by employing an arrangement which will accept the entire band and subject it to rectification in the usual manner. If this is done, the television signal and the 30,000-cycle supplementary carrier modulated with speech will appear at the output of the detector. Branch circuits with suitable filters will enable these two components to be separated and the television signal passed on to the reproducing apparatus. The

other component must be rectified to derive the original speech signal, which may then be impressed on the loud speaker amplifiers.

The reception scheme actually employed during the experiments was somewhat different. The television signal was received separately by means of the triple detection set employed for the demonstration. The speech signal was received in a similar manner employing the set utilized for the speech channel during the demonstration. This latter receiver was tuned to 1545 kilocycles. That reception in this manner is feasible, is evident from the diagram. The 1540–1550-kilocycle zone contains two speech sidebands and a carrier of 1575 − 30 or 1545 kilocycles. It is quite possible, therefore, to demodulate in one step, instead of "beating" the various components against the main carrier (1575 kilocycles) to produce a 30-kilocycle supplementary carrier which must be rectified a second time to derive the speech signal. The 1600–1610-kilocycle band was ignored. The receivers were sufficiently selective that, with the 5-kilocycle interval which existed between the two bands, no noteworthy crosstalk was experienced.

The results obtained in this manner were not as satisfactory as those to be had with the other system described. This can be attributed to two factors, both concerned with the transmitting apparatus: (1) In order to transmit both signals with the same transmitter, that is, the same vacuum tubes, the individual current amplitudes had to be reduced to at least one half, resulting in too weak a radio signal to clear the prevailing noise levels in New York, (2) In spite of the reduced amplitudes, a certain amount of inter-modulation was experienced in the transmitter which resulted in "crosstalk" between the channels. Notwithstanding these deficiencies, however, it was possible to recognize the speaker and to understand his remarks; but a short time ago, the performance would have been considered a very noteworthy achievement.

Experiments of this nature, although not new, are of particular interest where television is concerned, since, as Dr. Ives has indicated, the logical trend of development is toward a finer picture structure involving the transmission of much wider frequency bands, or what is more likely, the use of parallel scanning schemes and multi-channel transmission. The work, while necessarily somewhat cursory, may, therefore, be of value in affording an indication of the significance of multi-channel radio transmission in this connection. From a popular standpoint, these tests have been described as the transmission of both voice and image "on a single wave-length." To what extent this statement falls short of actually representing the facts in the case is

obvious from Fig. 9. It will be seen that a wider frequency band is actually employed with this system than was required for two separate channels. Furthermore, this wider band is much less effectively utilized. Two bands are required for the voice channel in place of one. At the receiver, one of these bands was disregarded. To have received both would have required apparatus accepting twice the band width and the gain in signal would have been offset by the corresponding increase in noise level. For all useful purposes, therefore, the energy radiated in the form of the second band is wasted.

To proceed further with a discussion of multi-channel radio transmission is beyond the scope of the present paper. Whatever the system employed, however, one conclusion illustrated by these experiments may be pointed to with confidence: television by radio requires a discrete and fairly wide frequency band. Hence the frequently predicted introduction of television as an adjunct to radio broadcasting without extensive changes in existing channel arrangements is extremely unlikely.

Contemporary Advances in Physics. XIV. Introduction to Wave-Mechanics

By KARL K. DARROW

IN a period when a limited domain of physical phenomena is exciting wide fervent interest and commanding intensive study, and continues for years to monopolize the attention of many brilliant theorists, sometimes it is the fortune of an ingenious mind to express or interpret or picture the already-discovered laws in a new way which makes so greatly favourable an impression, that in a moment it sweeps its rivals from the field. The new theory may not lead to more or better agreements with experience than did its predecessors; it need not make predictions which they were incapable of making; its mathematical processes may be identical with theirs, the old symbols reappearing with new names in the old equations. Contrariwise it may be born well endowed with these advantages which normally decide the contest between old theories and new, yet owe its victory not to them at all. It triumphs because it seems natural or sensible or reasonable or elegant or beautiful—words said of a theory which fulfils some deep-seated demand or evades some deep-rooted prejudice in the minds of its judges. Later its vogue may pass, not through the disclosure of any intrinsic defect, but because the physicists of the rising generation do not share the prejudices and the predilections of those who first applauded it. The kinetic theory of gases was welcomed by a generation which wished to believe in atoms; the electromagnetic theory by people prejudiced against the notion of action at a distance; the quantum-theory has always had to do battle against those who yearn for continuity in their images of Nature, and the theory to which these pages are devoted has captivated the world of physics in a few brief months because it seems to promise a fulfilment of that long-baffled and insuppressible desire.

Wave-mechanics being a new way of interpreting a vast field of well-known phenomena, it is unnecessary as indeed it would be impossible for me to recite in this place everything which the new theory is meant to explain. A few years hence, indeed, we may recognize in certain phenomena only newly or not yet discovered the securest basis for the new conceptions; but for the present, any adequate description of the facts on which Bohr's atom-model is based is nearly sufficient. I will recall only that the cardinal and dominant facts of the field which is the hunting-ground of the present generation

of theorists are these: atoms exist in Stationary States—they emit or absorb radiation in passing from one of these States to another—the frequency of the radiation is proportional to the energy-difference between the two States from one to the other of which the atom passes. Moreover, for certain kinds of atoms and molecules there are empirical formulæ which express known interrelations among the energy-values of the various Stationary States. These in brief are the major facts to be explained.

Bohr proved that the energy-values of the Stationary States of the hydrogen atom could be reproduced by affirming, *first*, that the atom consists of an electron and a nucleus of known masses and equal and opposite known charges; *second*, that these revolve around their common centre of mass according to the classical laws of mechanics and without radiating energy; *third*, that among all the conceivable orbits which such particles might describe there are certain ellipses, distinguished by certain especial and peculiar features, which alone the particles are permitted to choose—to each "permitted" ellipse there corresponds a Stationary State, and each Stationary State may be visualized as a permitted ellipse.

The first of these assumptions has never since departed from the physicists' world-pictures. In wave-mechanics it is still implicit, though easily overlooked. The second and third have not so firm a foothold. As I have elsewhere remarked, they are and always will be as good as they ever really were. If we make the first two of Bohr's assumptions, then it follows as a matter of course that whichever Stationary State of the hydrogen atom we may wish to consider or may hereafter discover, we shall always be able to find an elliptic orbit with the proper energy-value to serve as its picture. Yet this alone is not an important fact; the serious question is, whether the family of all permitted elliptic orbits is set apart from the vast multitude of forbidden ones by some simple and striking distinction which they all share and none of the rest possesses, whether they rejoice in some intrinsic patent of aristocracy. At first it seemed so; now, however, it turns out that the distinctive feature which originally was supposed to ennoble just the orbits required to account for the Stationary States, and no others, is not perfectly suited to every case. This weakened the prestige of the elliptic orbits; and though the introduction of the Spinning Electron has done much to save the situation, it has not done enough to preserve them from the crescent disparagement of those who never really liked them.

With other atoms and with molecules, the situation is much the same. Bohr and his successors visualized atoms as groups of electrons

surrounding nuclei; diatomic molecules, as paired nuclei surrounded by their jointly shared electron-family, capable of revolving like a dumbbell around their centre of mass and of vibrating like the two ends of a spring along their line of centres. These pictures persist in wave-mechanics; but the permitted vibration-amplitudes, the permitted rotation-speeds, and the permitted electron-orbits adduced to symbolize the Stationary States languish for the moment in the same discredit as the permitted elliptic orbits of the hydrogen atom.

Meanwhile, the humiliation of the electron-orbits accentuates the grave defect of the original atom-model of Bohr. That model offered nothing to interpret the fact that when an atom passes between two Stationary States of energy-values (let me say) E_i and E_j, it emits (or absorbs) radiation of the precise frequency $(E_i - E_j)/h$, the quotient of the energy-difference by the notorious constant of Planck. Neither in the initial State nor in the final State are the constituent parts of the atom-model vibrating with this frequency (except in occasional untypical cases). The frequencies of the waves streaming out from the atom do not agree with the frequencies of the motions assumed to exist inside the atom—a very uncomfortable idea, altogether discordant with all our experience of sound and electrical circuits.

If it should be found possible to incorporate into the atom-model something vibratory, having for its vibration-frequency the quotient of the energy-value of the then-existing Stationary State by Planck's constant: then in the foregoing case this "something" would be vibrating initially with frequency E_i/h and finally with frequency E_j/h, and the frequency of the emitted radiation would be the heterodyne or beat-frequency of these two. This is an agreeable idea; and wave-mechanics offers it. If then it should be found possible to arrive at the energy-values of the Stationary States by imposing conditions upon this vibrating entity instead of the electron-orbits, we should achieve as much as the electron-orbits enable us to achieve, and have the foregoing advantage also, and perhaps others as well. This is what wave-mechanics promises.

To this introduction I wish to join two warnings before plunging into the exposition. In the first place, wave-mechanics has several aspects, and may be approached from several directions; the one which I have chosen for this article is not the one which de Broglie elected nor the one which Schroedinger prefers.[1] In the second place, wave-

[1] I suspect that the method of exposition which I shall follow is the one which Schroedinger meant when he wrote "I had originally the intention of establishing the new formulation of the quantum-conditions in this more visualizable (*anschaulich*) way, but preferred a neutral mathematical form, because it makes the essence clearer." Schroedinger himself stresses the formal likeness between ordinary me-

mechanics is yet incomplete. It has been applied with success to many problems, but there are situations—those involving the Spinning Electron, for instance—in which the way to apply it is not yet clear, and many theorists are groping. The new theory is still plastic; many minds, perhaps the hands of many experimenters, have yet to work upon it before it is molded into its final shape.

Classical Mechanics and Wave-Mechanics

The underlying principles of "classical" or "Newtonian" mechanics may be stated in several alternative ways, each of which is especially well adapted to certain particular classes of problems. The most familiar of the statements is Newton's own. Unfortunately, it is another and less current which is the most expedient for the problems with which we have to deal. This formulation I will derive from Newton's, by imagining a particular extremely simple mechanical system and using Cartesian coordinates.

Conceive then a particle of mass m and charge e, moving in an electrostatic field of which the potential is a function $U(x, y, z)$ of the coordinates.[2] Its momentum is a vector of which the components are $m\dot{x}$, $m\dot{y}$, $m\dot{z}$. These are called the *momenta with respect to the coordinates* x, y, z, and are designated by p_x, p_y, p_z. The force upon the particle is the negative of the product of e into the gradient of the potential, a vector of which the components are dU/dx, dU/dy, dU/dz.

Newton's way of stating the underlying principles of mechanics then gives:

$$dp_x/dt = \dot{p}_x = -e\,dU/dx; \qquad \dot{p}_y = -e\,dU/dy, \qquad \dot{p}_z = -e\,dU/dz. \quad (1)$$

Multiplying the members of these three equations by $\dot{x}$, $\dot{y}$ and $\dot{z}$ respectively, and adding, we find:

$$\frac{d}{dt}\frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) = -e\left(\frac{dU}{dx}\frac{dx}{dt} + \frac{dU}{dy}\frac{dy}{dt} + \frac{dU}{dz}\frac{dz}{dt}\right). \quad (2)$$

On the left we have the rate of change of the kinetic energy T of the particle as it travels along its path. To interpret the right-hand

chanics and geometrical optics on the one hand and wave-mechanics and diffraction-theory on the other. I have not yet found this comparison helpful, and therefore cannot present it in a convincing manner.

I wish to acknowledge the valuable assistance of my colleague Mr. L. A. MacColl in preparing the mathematical portions of this paper.

[2] The reader will doubtless recognize that I am leading up to the case of the electron traveling in the field of a nucleus; I must therefore recall that in the case of the electron the charge e is intrinsically negative, and that according to the classical electromagnetic theory equation (1) should contain a term describing the reaction of the emitted radiation upon the electron—a term which is omitted in all contemporary atomic theories.

member, introduce a symbol V to designate the value of U at the locality where at any moment the particle actually is, multiplied by $+e$; this is the potential-energy-function of the particle, and the right-hand member of (2) is its rate of change. Therefore:

$$\frac{d}{dt}(T + V) = 0,$$

$$T + V = \text{constant} = E. \tag{3}$$

The constant E is (by definition) the energy. As the behavior of the particle depends upon the field, the ensemble of particle and field should be considered as one entity, the *system*, of which kinetic energy T and potential-energy-function V and total energy E are properties.

To bring out the next feature, I take the still more specific case of a particle of charge e and mass m moving in the inverse-square central field of a "nucleus," an immobile point-charge equal in magnitude and opposite in sign to the electron-charge. Using Cartesian coordinates with the origin at the nucleus, we have $V = -e^2/\sqrt{x^2 + y^2 + z^2}$; using polar coordinates,[3] we have $V = -e^2/r$. It is obvious that polar coordinates permit a much simpler expression for V than do Cartesians; on the other hand, they entail a distinctly more complicated expression for T. The proper choice of coordinates is often a vital question. For a few paragraphs I will carry along the reasoning in both coordinate-systems. The underlying equation (3) becomes, in the one and in the other:

$$\tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - e^2/\sqrt{x^2 + y^2 + z^2} = E, \tag{4a}$$

$$\tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + r^2 \sin^2\theta \cdot \dot{\varphi}^2) - e^2/r = E. \tag{4b}$$

In these equations, we have the potential-energy-function expressed as a function of the *coordinates* (x, y, z or r, θ, φ) and the kinetic energy expressed in terms of the coordinates and the *velocities* ($\dot{x}$, $\dot{y}$, $\dot{z}$ or $\dot{r}$, $\dot{\theta}$, $\dot{\varphi}$). It is desirable to express the kinetic energy in terms of the coordinates and the *momenta*. We have already met the momenta in Cartesian coordinates, the quantities $m\dot{x}$, $m\dot{y}$, $m\dot{z}$. It is obvious that they are the derivatives of the expression for the kinetic energy with respect to the velocities, always in Cartesian coordinates:

$$p_x = dT/d\dot{x}; \qquad p_y = dT/d\dot{y}; \qquad p_z = dT/d\dot{z}. \tag{5}$$

The momenta in any other coordinate-system are defined in the same

[3] The equations of transformation are: $x = r \sin\theta \cos\phi$, $y = r \sin\theta \sin\phi$, $z = r \cos\theta$.

way; first the kinetic energy is expressed as a function of the velocities, then differentiated with respect to these. In polar coordinates

$$p_r = dT/d\dot{r} = m\dot{r}; \qquad p_\theta = dT/d\dot{\theta} = mr^2\dot{\theta}; \qquad p_\varphi = dT/d\dot{\varphi} = mr^2 \sin^2\theta \cdot \dot{\varphi}. \tag{6}$$

Expressing in the equations (4a) and (4b) the kinetic energy in terms of the coordinates and momenta, we have

$$\frac{1}{2m}(p_x^2 + p_y^2 + p_z^2) - e^2/\sqrt{x^2 + y^2 + z^2} = E, \tag{7a}$$

$$\frac{1}{2m}\left(p_r^2 + \frac{1}{r^2}p_\theta^2 + \frac{1}{r^2 \sin^2\theta}p_\varphi^2\right) - e^2/r = E. \tag{7b}$$

Whenever in any problem the kinetic energy and the potential energy of the system are given as functions of coordinates and momenta, the problem is prepared for treatment by the methods of classical mechanics.

To make the next step, we consider the function $L = T - V$, the difference between the kinetic energy and the potential-energy-function of the particle, a function of the particle as it travels along its path in the force-field:

$$L = T - V = 2T - E \tag{8}$$

and the time-integral of this function

$$W = \int L dt = \int 2T dt - Et. \tag{9}$$

Into the expression for W, insert explicitly the expression for kinetic energy in Cartesian or in polar (or in any other) coordinates:

$$W = m\int(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)dt - Et = m\int(\dot{x}dx + \dot{y}dy + \dot{z}dz) - Et, \tag{10a}$$

$$W = m\int(\dot{r}^2 + r^2\dot{\theta}^2 + r^2 \sin^2\theta \cdot \dot{\varphi}^2)dt - Et = m\int(\dot{r}dr + r^2\dot{\theta}d\theta + r^2 \sin^2\theta \cdot \dot{\varphi}d\varphi) - Et. \tag{10b}$$

From all of this it follows that

$$p_x = dW/dx, \qquad p_y = dW/dy, \qquad p_z = dW/dz, \tag{11a}$$

$$p_r = dW/dr, \qquad p_\theta = dW/d\theta, \qquad p_\varphi = dW/d\varphi, \tag{11b}$$

and in general, *the momenta belonging to any coordinate-system are the derivatives of the function W with respect to the coordinates.*

Into the fundamental equation (7a) substitute these expressions for the momenta, and obtain:

$$\frac{1}{2m}\left[\left(\frac{\partial W}{\partial x}\right)^2 + \left(\frac{\partial W}{\partial y}\right)^2 + \left(\frac{\partial W}{\partial z}\right)^2\right] + V(x, y, z) = E \tag{12}$$

or, seeing that the quantity $\sqrt{(\partial W/\partial x)^2 + (\partial W/\partial y)^2 + (\partial W/\partial z)^2}$ is the magnitude $|\nabla W|$ of the *gradient* of the function W, the gradient of a function being a vector well known in vector-analysis and denoted by prefixing the sign ∇ or the abbreviation *grad* to the symbol of the function:

$$|\nabla W|^2 = 2m(E - V). \tag{13}$$

This equation governs the space-derivatives of the function W; it is complemented by the equation derived from (9) which governs the time-derivative of W:

$$\partial W/\partial t = -E. \tag{13a}$$

At this point the procedure of classical mechanics and the procedure of wave-mechanics diverge from one another.

Were we to follow the classical procedure, we should perform certain integrations and other processes, and arrive in the end at equations describing trajectories or orbits—in the particular case of an inverse-square central force-field, at equations describing elliptical orbits. The particular elliptical orbit to which the reasoning would conduct us would be determined by the value which had originally been assigned to the energy E, and the values which we attributed to the various constants of integration supervening in the course of the working-out. The function W, having served its purpose, would have vanished from the scene, leaving with us the electron swinging in its orbit within the atom or the planet in its orbit across the heavens.

The procedure of wave-mechanics, however, is based upon the observation that the equations (13) and (13a) together are *the description of a family of wave-fronts, traveling with the speed* $E/\sqrt{2m(E - V)}$ *through space.*

To display this aspect of the equation, let it be supposed at some prescribed time-instant t_0 the function W has a certain prescribed constant value W_0 at every point of a surface S_0; for instance, that at time $t_0 = 1$ it is equal to unity all over the sphere of unit radius centered at the origin. It is to be shown that at a slightly later instant $t_0 + dt$ there is again a surface everywhere over which the value of W is W_0, this not however being the same surface S_0, but another—a surface S_1 so placed that from any point P_0 on S_0 the shortest line to S_1 is perpendicular to S_0 and its length is $(E/\sqrt{2m(E - V)}dt)$.

This is easily shown. Imagine a vehicle [4] which at the instant t_0

[4] I use this word instead of "particle" lest this entity be confused with the moving electron to which the foregoing equations relate. The electron does travel along a curve normal to the surfaces of constant W, not however with the speed u about to be defined, but with a different speed related to u in a curious and significant way (cf. the allusion on p. 695).

is traveling through P_0, along the line normal to S_0, with a speed to be designated by u. At the instant $t_0 + dt$ it occupies a locality where the value of W is given by the formula:

$$W_0 + dW = W_0 + |\nabla W| ds + \left(\frac{\partial W}{\partial t}\right) dt$$
$$= W_0 + u|\nabla W| dt - E dt, \tag{14}$$

for in the time-interval dt it travels over a distance $ds = u dt$ along the normal to the surface S_0, and along this normal the slope of the function W is equal to $|\nabla W|$, and meanwhile at each point of space W is varying directly with time by virtue of the term $-Et$ occurring in the equation (9) which defines it. Now if the imaginary vehicle happens to be moving with just the speed defined by the equation

$$u = E/\nabla W = E/\sqrt{2m(E - V)}, \tag{15}$$

the coefficient of dt in equation (14) vanishes; that is, the vehicle as it moves outward keeps up with the prescribed value of W; but this is the same thing as saying that the value given for u in (15) is the speed of the wave-front.

At this (if not an earlier) stage of the argument, one begins to wonder what W "really is"; one turns back to seek the original definition of this artfully constructed function, so suddenly advanced from an auxiliary to the central rôle of the theory; one tries to grasp it, to form an image of it. I can do little to satisfy this very human craving. I can point out that W is that quantity "action" with which the Principle of Least Action has to do; this feature scarcely makes it more conceivable, but at least enhances its prestige. I can point out that since no one has ever seen what moves or is inside an atom, the conception of waves in an intangible medium curling and flowing around a centre is no more far-fetched than the conception of intangible particles sailing in ellipses around a nucleus. (To this one can reply that the planets in their courses supply a visible analogue for the notion of revolving electrons, but no one has seen in the sky the wave-fronts of the function W.) I can point out that for some important purposes, notably the prediction of the Stationary States, it makes no difference what the function W "really is"—no more difference than it makes to the solver of a quadratic equation whether the variable be called x or t, whether in the mind of the propounder of the equation it stood for distance or for time. One might in fact begin with the forthcoming equation (20) as foundation, laying it down without introduction or apology; yet there must be deep-lying

interconnections between the classical mechanics and the new, which such a procedure might mask. I can refer the reader to Schroedinger's own attempts to interpret W, some of which will figure in the last section of this article; or I can invite him to grow his own conception of W. This last in fact is what I will do.

Now if it is proposed to regard the fundamental dynamical equation (13) as the description of a family of wave-fronts perpetually wandering through space with the speed $E/\sqrt{2m(E-V)}$—and this is precisely what is proposed—then the description is obviously incomplete; for it omits to state the wave-length of these waves or the frequency of whatever be the vibrating thing which manifests itself by the waves, and indeed if the frequency were separately stated there would be no place for it in such an equation as (13). That equation, in fact, may be compared with the bare statement that the ripples traveling over the water of a pond from the place where a stone fell in are circles expanding at a given speed, or that the sound-waves proceeding through air from a distant source are plane waves traveling about 340 metres per second. To describe the ripples or the sound-waves completely it is essential to discover some ampler equation; a like extension is necessary here.

In treating familiar vibrating mechanical systems, stretched strings and tensed membranes and the like, it is customary to employ the general Wave-Equation

$$u^2\left(\frac{d^2\Psi}{dx^2}+\frac{d^2\Psi}{dy^2}+\frac{d^2\Psi}{dz^2}\right)\equiv u^2\nabla^2\Psi=\frac{d^2\Psi}{dt^2} \tag{16}$$

in which ∇^2 stands for the Laplacian differential operator (page 671); Ψ stands for the sidewise displacement of the string or distortion of the membrane or whatever it is that is transmitted as a wave; and u for the speed of propagation of the wave. It is furthermore customary to supplement this by the equation

$$d^2\Psi/dt^2 = -4\pi^2\nu^2\Psi, \tag{17}$$

in which ν stands for the frequency of the vibration; combining which with (16), one obtains

$$\nabla^2\Psi+k^2\Psi\equiv\nabla^2\Psi+\frac{4\pi^2\nu^2}{u^2}\Psi\equiv\nabla^2\Psi+\frac{4\pi^2}{\lambda^2}\Psi=0, \tag{18}$$

in which $\lambda = u/\nu$ stands for the wave-length of the wave-motion.

All of these matters will be developed at length in the following section. At this point it is necessary only to return to the description

of the wave-motion partially but only partially described by (13), and complete it by the assertion—not an inevitable nor a self-evident assumption, but an original and daring hypothesis—that it is indeed a wave-motion endowed with a frequency, and this the frequency

$$\nu = E/h. \tag{19}$$

This manner of introducing into every mechanical system a vibration-frequency linked with its energy by the vital quantum-relation (19) was the invention of Louis de Broglie.

The wave-equation to which this hypothesis leads us then is:

$$\nabla^2\Psi + \frac{8\pi^2 m}{h^2}(E - V)\Psi = 0. \tag{20}$$

This is a particular form of the wave-equation of de Broglie and Schroedinger. It is the form which I will use throughout this article, for it is adequate to the first steps in the processes of atom-design—adequate, for instance, to supply a theory of the major features of the spectrum of atomic hydrogen, though not of its fine-structure; adequate also to interpret the data of the experiment of Davisson and Germer, and sufficient for an introduction to the ways of thinking which constitute wave-mechanics. Nevertheless it is certainly not the general wave-equation, for it is subject to at least two limitations.

The first of these is, that equation (20) is based upon Newtonian, not upon relativistic mechanics. We should therefore expect it to be valid only for slow-moving particles, to be the limiting form of a relativistic wave-equation appropriate to all velocities. Such an equation, indeed, was the first propounded by de Broglie. The past history of atomic theory suggests that we should need it when embarking upon the enterprise of explaining the fine-structure of the hydrogen spectrum. The latest developments in that history, however, indicate that the mere replacement of equation (20) by its relativistic analogue would not suffice for that enterprise; due allowance must be made in addition for the "spin" of the electron.[5] Wave-mechanics being yet too young to have furnished an answer to this twofold problem, the relativistic equation still wants what may in the end turn out to be its main experimental support. Yet it can scarcely be doubted that relativity must figure in the general wave-equation.

The second limitation upon equation (20) is due to its origin in

[5] For the application of the relativistic equation to the hydrogen atom without allowance for the spinning electron, see V. Fock, *Zs. f. Phys.*, **38**, pp. 262–269 (1926). See also the first footnote on p. 688.

equation (13), and to a peculiar feature of that equation—to the fact that in it the magnitude of the gradient of W stands equated to a function of the coordinates. This indeed is the feature which rendered it possible to imagine flowing waves. Now this feature occurs because the system to which equation (13) relates—the particle voyaging in a force-field—has a *kinetic-energy-function which is the sum of the squares of the momenta* (multiplied by a constant). Had we presupposed a system possessing a kinetic-energy-function not capable of being so expressed—two particles of different masses voyaging in a force-field, or a rigid rotating body of irregular shape, for example—the equation which we should have obtained in lieu of (13) would not have had the peculiar feature aforesaid; the wave-picture would not have offered itself, much less the equation (20) which was superposed upon the wave-picture. It is precisely at this obstacle that the mode of thought known as *non-Euclidean geometry* proves itself useful. It proposes equations of a general type, which can be written down for every system of which the kinetic-energy-function is preassigned, and which for the single particle floating in a force-field become the equation (13) and (20). In the language of non-Euclidean geometry, even the words and the symbols for *wave* and *wave-speed* and *gradient* and *Laplacian* are preserved; but whether they are advantageous to anyone not already versed in this subject may well be doubted. Suffice it to say, that non-Euclidean geometry provides a general equation [6] of which (20) is a special case, and that the general equation has already justified its existence by its successes in dealing with certain atom-models and molecule-models such as the rigid rotator used in the study of band-spectra. But the question as to what the waves "really are" becomes in these cases all the darker and more perplexing.

One further step, and we attain to the idea on which the calculation of the energy-values of the Stationary States reposes.

It is very well known that a medium capable of transmitting waves, and *bounded* in certain ways, may develop what are variously known as standing waves—stationary wave-patterns—the phenomena of resonance. Air enclosed in a box, a string pinched at the ends, a membrane clamped around its circumference, the mobile electricity in

[6] Let the kinetic-energy-function of the system, expressed in terms of the coordinates and velocities, be written

$$T = \sum_i \sum_j Q_{ij}\dot{q}_i\dot{q}_j$$

and let Δ stand for the Laplacian operator in the non-Euclidean configuration-space of which the metric is $ds^2 = \Sigma\Sigma Q_{ij}dq_idq_j$; then the general wave-equation of de Broglie and Schroedinger is:

$$h^2\Delta\psi + 8\pi^2(E - V)\psi = 0.$$

a tuned circuit—each of these vibrates in a wave-pattern of "nodes" and "loops" if the frequency of vibration imposed upon it conforms to one of its own "natural frequencies" or "resonance frequencies." To each of these natural frequencies corresponds a particular pattern of loops and nodes; when one of them is impressed upon the medium, its corresponding wave-pattern springs into existence, and would continue forever were it not for friction internal or external. When any frequency not agreeing with one of the resonances is imposed upon the bounded medium, the resulting motion is very much more complicated. The calculation of these natural frequencies, the mapping of these vibration-patterns, is performed by using the methods of one of the great divisions of mathematical physics—the methods underlying the Theory of Acoustics.

May the Stationary States, then, of a natural atomic system be visualized as stationary wave-patterns such as these, and their energy-values as the products of the natural frequencies by the constant of Planck? Are the problems of atomic theory to be solved by devising atom-models imitated after familiar resonant bodies or tuned circuits, and applying to these "acoustic models" the mathematical technique of the Theory of Acoustics? This idea was developed by E. Schroedinger.[7]

Familiar Examples of Stationary Wave-Patterns

To display the laws governing wave-patterns, I will develop three examples: the stretched string, the tensed membrane, the ball of fluid confined in a spherical shell. The first of these is the simplest and most familiar of all instances; excursions into the theory of vibrating systems commence always at the wire of the piano and the string of the violin. Physically, this is a case of one dimension (distance, measured along the length of the string); mathematically, it is a case of two variables (that distance, and the time). The example of the tensed membrane is not unfamiliar in the practice of telephony, though many of the diaphragms of actual instruments are too thick to be considered such; for a membrane is, by definition, infinitely thin. It is a case of two dimensions and three variables. It will reveal to us the desirability of choosing for each specific problem its appropriate set of coordinates; and we shall observe what happens when one of the chosen coordinates is cyclic, being an angle which for all practical purposes returns to its original value when increased by 2π; and we

[7] Since the present article is based henceforth chiefly on Schroedinger's publications, I wish to make particular reference here to works embodying de Broglie's contributions: his own *Ondes et mouvements* (Paris, Gauthier-Villars, 1926) and article in *Jour. de Phys.* (6), **7**, pp. 321–337 (1926); L. Brillouin, *ibid.*, pp. 353–368.

shall encounter functions not so widely known as the simple sine and cosine which suffice for the case of the stretched string. The little-known example of the ball of fluid, with its three dimensions and four variables, will repeat these lessons, and will serve as the final stepping-stone to the wave-motions imagined by de Broglie and by Schroedinger. To proceed to these, it will suffice to imagine strings and fluids not uniform like those of the simple theory of vibrating systems and sound, but varying from point to point in a curious and artificial way.

Example of the Stretched String

Imagine a stretched string, infinitely long, extended along the x-axis of a system of coordinates. Designate the tension in the string by T, the (linear) density of the string by ρ. To derive the differential equation governing the motion, conceive the string as a succession of short straight segments (Figure 1). Each segment exerts upon its neighbors a force, which is the tension in the string. When the string lies straight along the axis of x, each segment lies in equilibrium between the equal and opposite forces which its neighbors exert upon it. When however the string is drawn sidewise (remaining, we shall suppose, in the xy-plane) the neighbors of each segment are oblique to it and to one another, the forces which they exert upon it have components along the y-direction. These components are in general unequal, and their algebraic sum is a force urging the segment along the y-direction. Denote by dx the length of such a segment, by y its lateral displacement, by θ the angle between it and the axis of x; so that $dy/dx = \tan\theta$, and ρdx stands for the mass of the segment. The resultant force upon the segment is given by:

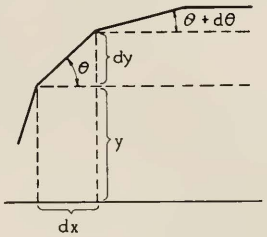


Fig. 1

$$F = T[\sin(\theta + d\theta) - \sin\theta] = T[\tan(\theta + d\theta) - \tan\theta]$$
$$= T \cdot d(\tan\theta/dx)dx = T(d^2y/dx^2)dx \qquad (101)$$

to the degree of approximation to which the difference between $\sin\theta$ and $\tan\theta$ may be neglected.[8]

Equating this to the product of mass by acceleration, we obtain:

$$\rho d^2y/dt^2 = T(d^2y/dx^2) \qquad (102)$$

[8] This is the degree of approximation all but universal in the theory of vibrating systems and sound. The conclusions from this theory are therefore strictly valid only in the limit of infinitesimal displacements or distortions.

or using dots to symbolize differentiation with respect to time, and dashes to represent differentiation with respect to space:

$$\ddot{y} = \sqrt{T/\rho} \cdot y''. \tag{103}$$

This equation, a linear combination of a second derivative with respect to time and a second derivative with respect to space, is the first and simplest of our wave-equations.

It is called a wave-equation, because it may represent—it does not necessarily represent, but it may—a shape or a figure or a distortion of the string (whichever one may choose to call it) which travels continually and indefinitely along the string with a constant speed.

To illustrate this possibility, let us suppose that at the time $t = 0$ the string is distorted into a sinusoidal curve described by the equation:

$$y = A \sin mx \quad \text{at} \quad t = 0 \tag{104}$$

and that its points are moving parallel to the y-axis with speeds described by the equation:

$$\dot{y} = nA \cos mx \quad \text{at} \quad t = 0. \tag{105}$$

At any other moment t, the configuration of the string is described by the equations:

$$y = A \sin (nt + mx), \qquad \dot{y} = nA \cos (nt + mx), \tag{106}$$

for these satisfy the differential equation which underlies the whole theory, and they satisfy also the "initial conditions" specified by (104) and (105). They satisfy these equations, that is to say, provided that a certain relation is fulfilled among the constants n and m, and the quantities T and ρ which describe the physical nature of the stretched string; this relation being:

$$n/m = \sqrt{T/\rho}. \tag{107}$$

If this relation is fulfilled, the condition of the string throughout all time is described by the equations (106).

Examining these equations, we perceive that they signify that the values of displacement and speed, which at the time $t = 0$ existed at any point x_0 on the string, are at any other time t to be found at the point $x_1 = x_0 - (n/m)t$. These values are moving steadily along the string; the whole configuration of the string, its sinusoidal shape and its transverse velocities, is slipping steadily lengthwise in the direction of decreasing x—the shape of the string is being transmitted as a

wave, with the ratio of the constants n/m for its speed of propagation u:

$$u = n/m = \sqrt{T/\rho}. \tag{108}$$

This result justifies the title *wave-equation* for the differential equation (103), and the meaning *speed of propagation* for its coefficient $\sqrt{T/\rho}$.

The reader will scarcely have failed to notice, however, that the result was obtained only by prescribing very sharply defined physical conditions. The string was supposed infinitely long; it was supposed distorted into the form of a sine-wave; the transverse speeds of its successive particles at the instant $t = 0$ were preassigned as rigorously as their positions. Were we to alter this last specification, we should arrive at very different results. If for instance we should make the assumption that at $t = 0$ the string is distorted into a sine-wave *and is stationary*, the equations (106) would not be adequate to describe what happens. We should then be forced to have recourse to a more general solution of the differential equation:

$$y = C \sin (nt + mx) + D \sin (nt - mx) \tag{109}$$

and to adjust the constants C and D so as to conform to the newly prescribed initial conditions, which are:

$$y = A \sin mx, \qquad \dot{y} = 0 \quad \text{at} \quad t = 0. \tag{110}$$

The adjustment is attained by making $C = D = \frac{1}{2}A$, whereupon we get:

$$y = A \sin nt \cos mx, \tag{111}$$

an equation which describes not a wave advancing perpetually along the string, but a stationary oscillation with nodes and loops of vibration, like those which a violin-string properly bowed exhibits, those in the air-column of Kundt's tube which the hillocks of dust reveal. One would hardly detect by instinct in this stationary wave-pattern the superposition of two oppositely gliding wave-trains each traveling with the speed $u = n/m = \sqrt{T/\rho}$. Yet the one is always equivalent to the other, and in the equation (111), the coefficients n and m are linked to one another through the wave-speed characterizing the string, and the equation may be written

$$y = A \sin umt \cos mx, \qquad u = \sqrt{T/\rho}. \tag{112}$$

Although the tension and the density of the string thus determine n when m is preassigned (or vice versa), nothing so far brought upon the scene compels any limitations upon the coefficient m. The infi-

nitely long wire can sustain vibrations of any wave-length, or vibrations of two or any number of wave-lengths simultaneously, with any interrelation whatever among their several amplitudes and phases. On this fact rests our freedom to impose any initial conditions whatsoever on such a wire (subject to the usual restrictions of continuity and finiteness). For, if it be demanded that at $t = 0$ the displacement y shall vary along the wire according to any totally arbitrary function $f(x)$, and the transverse speed y according to any totally arbitrary function $g(x)$, then we have only to expand these functions f and g into Fourier series, or if need be, Fourier integrals; and each term in such an expansion corresponds to such a solution as (109), with a specific value of m and such specific values of C and D as the initial conditions require; and the configuration of the wire forever before and after is described by the sum of all these solutions. In such a case we should not see an unchanging distortion of the wire slipping steadily along its length with a constant speed, nor a stationary pattern of nodes and loops. All the obvious features of wave-motions would be blotted out; and yet the infinitely complicated and variable figure of the string would be equivalent, in the last analysis, to a multitude of sinusoidal wave-trains perpetually gliding to and fro with the same uniform speed.

As soon, however, as we impose *boundary-conditions*, the vibrations which the string can execute are severely restricted.

As a simple and familiar example of boundary-conditions, I will assume that the string is clamped at the points $x = 0$ and $x = L$, and concern myself only with the finite length of string, L, comprised between these two fixed extremities.

As a preparation for future developments, it is advisable to restate the underlying differential equation, and solve it *ab initio*. We have:

$$\ddot{y} = u^2 y'', \tag{113}$$

in which u stands for the speed of propagation of a sine-wave along an infinite wire. We essay a tentative solution, in the form of a product of a function of t only by a function of x only:

$$y = g(t) \cdot f(x). \tag{114}$$

The differential equation subjects the functions g and f to the condition:

$$f''/f = \ddot{g}/u^2 g = -m^2, \tag{115}$$

for, since the first member of this triple equation does not depend on t, and the second does not depend on x, each of the two must be inde-

pendent of both t and x, and equal to a constant which (for the sake of consistency with prior notation) I denote by $-m^2$. Solutions of these differential equations into which the underlying one was broken up are these:

$$f = A \cos mx + B \sin mx, \qquad g = C \cos mut + D \sin mut. \quad (116)$$

So far, there is no limitation upon m.

Now come the boundary-conditions, formulated thus:

$$f(0) = f(L) = 0. \quad (117)$$

We have now encountered, in its simplest example, the peculiar and characteristic problem of the Theory of Acoustics, which is also the peculiar and characteristic problem of the type of Atomic Theory which is inherent in wave-mechanics. This is not the question which we meet in the theory of moving particles, where we are asked what path a particle will follow through all future time if its position and velocity at a single moment are given. A similar question will indeed presently be asked and answered; but this peculiar problem intrudes itself at the beginning.

To adjust the function $f(x)$ to the boundary conditions, it is evident that we must set $A = 0$ and $\sin mL = 0$; therefore we must assume that m has one of the values:

$$m = k\pi/L \qquad k = 1, 2, 3, 4 \cdots. \quad (118)$$

The boundary-conditions have compelled the coefficient m to choose among a rigidly defined series of values. The wave-lengths, and consequently the frequencies, of the permitted vibrations are strictly determined.

The permitted values of m are known in German as the *Eigenwerte* of the differential equation for the boundary-conditions in question. The English term would be "characteristic values"; but it is long and has many meanings, and I think it preferable to borrow the German word as a foreshadowing of the application which Schroedinger has made peculiarly his own. To each *Eigenwert* of m there corresponds a value of the vibration-frequency $mu/2\pi$, which in German is called an *Eigenfrequenz;* but here we may as well keep to the English term *natural frequency.*

To each *Eigenwert* there corresponds a solution of the differential equation, an *Eigenfunktion.* In the present instance the *Eigenfunktion* corresponding to the *Eigenwert* $m = k\pi/L$ is:

$$y_k = \sin \frac{k\pi}{L} x \left(C_k \cos \frac{k\pi u}{L} t + D_k \sin \frac{k\pi u}{L} t \right). \quad (119)$$

It represents a sinusoidal stationary oscillation of the wire, with nodes at the ends and at $(k - 1)$ points spaced evenly between the ends—a case not difficult to realize with a violin-string, if k be not too great. The constants C and D specify the amplitude of the oscillation, and its phase at any given instant.

It is of course not necessary that the motion of the wire should conform to a single *Eigenfunktion*. Any number of *Eigenfunktionen*, corresponding to different permitted values of m—different integer values of k—might coexist simultaneously, each with its particular values of C_k and D_k; the actual distortion of the wire would be the superposition of all. It would in fact be necessary to adjust the initial distortion of the wire and the initial velocities of its points with infinite accuracy, to cause its future motion to conform exactly to a single *Eigenfunktion*. On the other hand, any choice whatever of initial distortion and initial velocities would entail a future motion compounded out of the various *Eigenfunktionen* with suitable values of C_k and D_k, which could be computed. This process corresponds to that of determining the future orbit of a particle of which the position and the velocity at a given instant are preassigned.[9] Both in acoustics and in wave-mechanics it is, as a rule, much more laborious than the determination of natural frequencies; and happily it is often less important, though not always to be neglected.

Example of the Tensed Membrane

The differential equation of the tensed membrane is:

$$\nabla^2 z = \frac{d^2z}{dx^2} + \frac{d^2z}{dy^2} = \frac{1}{u^2}\frac{d^2z}{dt^2}. \tag{120}$$

The coordinate-axes of x and y lie in the equilibrium plane of the membrane, and z stands for the displacement of any point of the membrane normally from this plane. The symbol u stands for the speed of a sine-wave traveling in an infinite membrane of the same tension T and surface-density ρ as the actual one, and is determined by the equation:

$$u^2 = T/\rho, \tag{121}$$

which is derived by an obvious extension of the method employed in deriving the like equation for a stretched string. In an actual bounded membrane the motion may be tremendously complicated, but it can

[9] Inversely, the imposition of quantum-conditions upon orbits corresponds to the determination of natural frequencies; here is the bridge between the atom-models with electron-orbits and the atom-models of wave-mechanics.

be analyzed into a multitude of wave-trains traveling to and fro with the speed u.

The symbol ∇^2 (to be read *del* or *nabla-squared*) stands for the Laplacian operator which in rectangular coordinates is d^2/dx^2, or $(d^2/dx^2 + d^2/dy^2)$, or $(d^2/dx^2 + d^2/dy^2 + d^2/dz^2)$, according as we are dealing with one, two or three dimensions. In other coordinates than rectangular, it naturally assumes other forms. Now in these problems of two and three dimensions, the choice of coordinate-system and the imposition of boundary-conditions are two decisions which cannot be separated from one another. Were we to decree that the membrane should be square or rectangular with its edges clamped, the suitable coordinate-system would be the rectangular. The problem would then be extremely simple (the reader can easily solve it for himself by using the method adopted for the stretched string, and will arrive at very similar results) but not so instructive to us as the problem of the circular membrane with clamped edge. For this we must adopt polar coordinates (with the origin at the centre of the membrane, naturally). In these, the Laplacian operator assumes the form:

$$\nabla^2 = \frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} + \frac{1}{r^2}\frac{d^2}{d\theta^2}. \tag{122}$$

We restate the fundamental differential equation (120) in this fashion; we essay a tentative solution in the form of a product of a function $f(r)$ of r exclusively, a function $F(\theta)$ of θ exclusively, and a function $g(t)$ of t exclusively; and we discover as before that each of these functions is subjected to a differential equation of its own. The procedure is like that already used in the case of the stretched string clamped at its ends. First we have

$$\frac{1}{f}\frac{d^2f}{dr^2} + \frac{1}{rf}\frac{df}{dr} + \frac{1}{r^2F}\frac{d^2F}{d\theta^2} = \frac{1}{u^2g}\frac{d^2g}{dt^2} = -m^2, \tag{123}$$

for, since the first member of this triplet does not depend on t, the second not on r nor on θ, both must be independent of all three variables and equal to a constant which, as before, I denote by $-m^2$. The differential equation for the factor dependent on t has the solution:

$$g(t) = A\cos mut + B\sin mut. \tag{124}$$

Our experience with the stretched string suggests that m will be restricted to certain *Eigenwerte*, derived from the boundary-conditions; and this is true; but before arriving at these, we must attend to the differential equation governing the functions f and F. This assumes the form:

$$\frac{r^2}{f}\frac{d^2f}{dr^2} + \frac{r}{f}\frac{df}{dr} + m^2r^2 = -\frac{1}{F}\frac{d^2F}{d\theta^2} = \lambda^2, \tag{125}$$

both members of the equation being, by the familiar reasoning, equal to a constant which I denote by λ^2. It follows that the function $F(\theta)$ is of the form:

$$F(\theta) = C\cos\lambda\theta + D\sin\lambda\theta \tag{126}$$

and the coefficient λ thus far seems to be unrestricted. But it carries its own restrictions in itself; for the coordinate θ is a cyclic coordinate, like longitude on the earth; whenever it is altered by 2π, we are back at the same place. The function $F(\theta)$ must therefore repeat itself whenever θ is altered by 2π; but this will not occur, unless λ is an integer:

$$\lambda = 0, 1, 2, 3 \cdots. \tag{126a}$$

These are the *Eigenwerte*, and the functions (126) with one or another of these values assigned to λ are the *Eigenfunktionen*, of the equation (125). In this case we have obtained *Eigenwerte* for the parameter and *Eigenfunktionen* for the solutions of a differential equation, not out of boundary conditions but out of the simple fact that the independent variable is by its nature cyclic. Such cases will occur in the undulatory mechanics.

We arrive at the third and last step of the problem: the determination of the function $f(r)$. It is governed by the differential equations:

$$\frac{d^2f}{dr^2} + \frac{1}{r}\frac{df}{dr} + \left(m^2 - \frac{\lambda^2}{r^2}\right)f = 0, \tag{127}$$

a distinct equation for each of the permitted integer values of λ. As the solution of such an equation as (115) is a sine-function of the variable mx, so the solution of such an equation as (127) is a function of the variable mx; not however a sine-function, but a Bessel function. For the values 0, 1, 2, $\cdots$ of λ, the solutions of (127) are the Bessel functions of order 0, 1, 2, $\cdots$, denoted by $J_0(mr)$, $J_i(mr)$, $J_2(mr)$, and so forth.

Like the sine-function of mx, the Bessel functions of mr oscillate back and forth between negative and positive values as their variable increases from zero to infinity, and pass through zero at an infinite number of discrete values of mr. These do not lie at equal intervals, as do the values of mx at which sin mx vanishes. Their values may be found in the tables; I shall designate them as b^1, b^2, b^3, $\cdots$ in order of increasing magnitude, using the superscripts not as expo-

nents, but as ordinal numbers so that I may reserve the subscripts to distinguish the various Bessel functions from one another. The function

$$Z = J_\lambda(mr)(C \cos \lambda\theta + D \sin \lambda\theta)(A \cos mut + B \sin mut) \quad (129)$$

represents a stationary oscillation of an infinitely extended membrane, in which λ lines intersecting one another at the origin are nodal lines, and an infinity of concentric circles centred at the origin are nodal circles. These lines and circles are motionless while the sections of the membrane which they delimit vibrate with the frequency $mu/2\pi$. The λ lines are spaced uniformly in angle; the radii r_1, r_2, $\cdots$ of the infinity of circles are obtained by dividing m into the roots b_λ^1, b_λ^2, b_λ^3, $\cdots$ of the Bessel function of order λ, $J_\lambda(mr)$.

How then does the boundary-condition upon the finite membrane enter in? Obviously, if a membrane of radius R be clamped at its edge, and if it is vibrating in the manner described by (129), then the edge must coincide with one of the nodal circles; the radius R must be equal to one of the quantities b_λ^i/m. Or rather, since the nodal circles are to be adjusted to the size of the diaphragm and not the size of the diaphragm to the nodal circles, the coefficient m must conform to one of the equations:

$$m = b_\lambda^1/R, \quad \text{or} \quad b_\lambda^2/R, \quad \text{or} \quad b_\lambda^3/R, \quad \cdots. \quad (130)$$

These equations define Eigenwerte of the parameter m *in the differential equation of the tensed membrane.* There is a double infinity of these—an infinite series of them for each of the *Eigenwerte* of the parameter λ. To each corresponds a natural frequency of the membrane, and to each corresponds an *Eigenfunktion*, the one written down in (129) with the proper value of m taken from (130). The constants A, B, C, and D in the *Eigenfunktionen* specify the amplitude of the oscillation, the phase of the vibrations at any given instant, and the orientation of the nodal lines with respect to any given axis. Any number of *Eigenfunktionen* may coexist simultaneously; the actual distortion of the membrane will be the superposition of all. Any initial conditions imposed on z and $\dot{z}$ (and not involving discontinuities or infinities) could be satisfied by adjusting the constants.

Example of the Ball of Fluid

Among the familiar vibrating systems the ball of fluid presents the closest analogy to the atom-model for the hydrogen atom in wave-mechanics, the wave-patterns in the two cases being strikingly alike.

In three dimensions and in polar coordinates (those appropriate to the boundary-conditions which we shall impose) the wave-equation assumes the somewhat alarmingly intricate form:

$$\ddot{\Psi} = u^2\nabla^2\Psi = u^2\frac{\operatorname{cosec}\theta}{r^2}\left[\frac{d}{dr}\left(r^2\sin\theta\frac{d\Psi}{dr}\right) + \frac{d}{d\varphi}\left(\operatorname{cosec}\theta\frac{d\Psi}{d\varphi}\right) + \frac{d}{d\theta}\left(\sin\theta\frac{d\Psi}{d\theta}\right)\right]. \tag{131}$$

The argument Ψ can no longer be visualized as a displacement perpendicular to the equilibrium-position of the undistorted medium, since all three dimensions are already used up. The reader may visualize it, if he will, as a condensation or a rarefaction, after the fashion of sound-waves. Perhaps not to visualize it at all would be a better preparation for the study of wave-mechanics.

In the familiar way, we essay a solution in the form of a product of a function of time $g(t)$, a function of radius $f(r)$, a function $\Phi(\phi)$ of the longitude-angle ϕ and a function $\Theta(\theta)$ of the colatitude-angle θ. As before, we find that the time-function is of the form:

$$g(t) = A\cos mut + B\sin mut \tag{132}$$

and, as before, we shall find that the boundary-conditions confine the coefficient m and the frequency $mu/2\pi$ to certain "permitted" values. The angle-functions and the radius-function are governed by the differential equations:

$$\frac{1}{f}\left[\frac{d}{dr}\left(r^2\frac{df}{dr}\right) + m^2r^2f\right] = -\frac{1}{Y}\operatorname{cosec}\theta\left[\frac{d}{d\varphi}\left(\operatorname{cosec}\theta\frac{dY}{d\varphi}\right) + \frac{d}{d\theta}\left(\sin\theta\frac{dY}{d\theta}\right)\right] = \lambda, \tag{133}$$

in which Y stands for the product of Θ and Φ, and λ for a constant which seems to be arbitrary, but as a matter of fact is constrained by the same circumstance as arose in the case of the membrane; for, whenever φ is altered by 2π and θ by π, we are back at the same place as before, and the function Y must have the same value as before; and this will occur only if

$$\lambda = n(n+1), \qquad n = 0, 1, 2, 3, \cdots, \tag{134}$$

these being the *Eigenwerte* for the differential equation in (133) for the angle-function.[10] The corresponding *Eigenfunktionen* are spherical

[10] This and the following statements about the functions Y_n are proved by writing Y in the second of equations (133) as the product of a function of θ and a function of ϕ, and so dissolving the equation into two in the manner which I have already

harmonics. To each value of n belongs a "spherical harmonic of order n," which itself is a sum of $(2n + 1)$ terms, each multiplied by a constant which is at our disposal and can be adjusted to fit initial conditions or to emphasize particular modes of vibration. These terms are products of sine-functions of φ by peculiar functions, the *Legendrian functions* $P_{n,\,s}$, of the variable θ; so that the *Eigenfunktion* for a permitted value $n(n + 1)$ of the parameter λ has this for its most general form:

$$Y_n(\theta,\,\varphi) = a_{n,\,0}P_{n,\,0}(\cos\theta) + \sum_{s=1}^{n} a_{n,\,s}\cos(s\varphi)P_{n,\,s}(\cos\theta) + \sum_{s=1}^{n} b_{n,\,s}\sin(s\varphi)P_{n,\,s}(\cos\theta). \tag{135}$$

Each term by itself describes a particular mode of vibration of the fluid; the sum represents a superposition of divers modes of vibration. If we isolate one of these modes by giving to n some particular value n_1, and to s some particular value s_1, and causing all the constants a and b in (135) to vanish except $a_{n_1,\,s_1}$ and $b_{n_1,\,s_1}$; we then find that Y, and consequently Ψ, and consequently the motion altogether, vanishes at s_1 values of φ and at $n_1 - s_1$ values of θ. If we draw a sphere centred at the origin, we find that its surface bears s_1 nodal meridian-circles, and $n_1 - s_1$ nodal latitude-circles, along which there is perpetual rest. If we consider all the spheres at once—if, that is to say, we consider the entire volume of the fluid medium—we see that when the fluid is vibrating in the mode distinguished by the integers (I had almost said "quantum-numbers"!) n_1 and s_1, it is divided into compartments by s_1 nodal planes intersecting along the axis $\theta = 0°$, and $n_1 - s_1$ double-cones having that axis for their axis and the origin for this apex.

We have not yet considered the dependence of the wave-motion on the radius r; but the close analogy between this and the corresponding stage of the problem of the tensed membrane will make the task easy. The differential equation (133) for $f(r)$ resembles Bessel's equation (127), and has the somewhat similar solution

$$f(r) = \frac{1}{\sqrt{r}} J_{n+\frac{1}{2}}(mr). \tag{136}$$

used five or six times; the values of the constant s in equation (135) are the *Eigenwerte* of the latter of these two. I thought it desirable not to overload the exposition by carrying through all stages of the process of solution, especially as the splitting of $Y_n(\theta,\,\phi)$ into the two functions is of secondary importance in the atom-model to which all this leads up; nevertheless the reader may find it advantageous to supply the lack.

This function vanishes, entailing the vanishing of the wave-motion, at an infinity of discrete values of the variable mr:—the roots of the function, which I denote in order of increasing magnitude by B^1, B^2, B^3, $\cdots$. In an infinite medium we could assign any value whatever to r, and then there would be an infinity of nodal spheres, their radii given by B^1/m, B^2/m, B^3/m, $\cdots$. If the medium is bounded by a rigid spherical wall of radius R, the coefficient m must possess one of the values B^i/R, so that one of the nodal spheres may coincide with the wall. These are the *Eigenwerte* of the constant m, and the natural frequencies of the corresponding vibrations are given by $B^iu/2\pi R$. The *Eigenfunktionen* are given by the equation (136) with the various values B^i/R substituted for the parameter m.

The *Eigenfunktionen* of the fundamental differential equation for the fluid sphere are, therefore, each a product of a radius-function given by (136), with a "permitted" value for the constant m determined by the boundary-condition; an angle-function given by (135), with "permitted" values for the constants n and s, determined by the fact that the angles are cyclic variables; and a time-function given by (132), with a "permitted" vibration-frequency determined by the boundary-condition. Each *Eigenfunktion* with the indices m, n, s describes a mode of vibration, in which the fluid sphere is divided into compartments by s meridian planes, $(n - s)$ double-cones, and a certain number of spheres, upon each of which the fluid is perpetually at rest; within the compartments, it vibrates with a prescribed frequency.

Atom-Models in Wave-Mechanics

Case of a "String" for which the Wave-Speed is Variable, or even Imaginary

Thus far I have used the images of the stretched string, the tensed membrane, and the elastic fluid to illustrate the behavior of the differential equation

$$u^2\nabla^2\Psi = d^2\Psi/dt^2, \tag{151}$$

when the coefficient u^2 is a positive constant. In these examples u^2 is interpreted as the ratio of the intrinsically positive quantities "tension" (or "pressure") and "density," and turns out to be equal to the square of the speed of propagation of sine-waves in the string, membrane, or fluid. In certain problems of undulatory mechanics we encounter just such an equation. In some of the most important applications of Schroedinger's theory, however, one meets with differential equations of the type of (151), in which however the coefficient u^2 depends on the coordinates and even assumes negative

values! Such equations need not be more difficult to solve than the conventional wave-equation in which u^2 stands for a positive constant; but the image of the elastic medium becomes unsatisfying. In the one-dimensional case, so long as u^2 remains a positive function of x, one can visualize a string of which the density varies along its length; but when u^2 passes through zero and becomes negative, the wave-speed attains zero and is superseded by an imaginary quantity. One may speak, in such a case, of a "string" or a "fluid" characterized by an "imaginary wave-speed." So speaking, one comes perilously close to the verge of using words devoid of physical meaning; but otherwise, there is no verbal language with which to relieve the monotony of the procession of equations.

The differential equation of the type of (151), with a constant negative value of the coefficient u^2, is not a difficult one. Confining ourselves to one dimension, we find for one of the solutions of the equation for a "string with constant imaginary wave-speed" this expression:

$$\Psi = (A \cos mUt + B \sin mUt)(Ce^{mx} + De^{-mx}), \qquad (152)$$

in which U stands for the (real) square root of $-u^2$. This is a much less tractable function than the product of sine-functions which serves when u^2 is positive. One cannot, for instance, find *Eigenwerte* for the constant m whereby the function can be made to vanish at all times at two distinct points upon the "string"; or rather, one can find only the value $m = 0$, which fulfils this familiar boundary-condition by destroying the function. Similarly, one cannot force Ψ to remain finite everywhere except by annulling either m or else both A and B, again destroying the function. Vibrations which are sine-functions of time are, however, permitted by the differential equation.

Consider now the equation

$$d^2y/dx^2 = (a - bx^2)d^2y/dt^2, \qquad (153)$$

which may be regarded as the wave-equation of a string of which the wave-speed varies with x along its length as the function $(a - bx^2)^{-1/2}$, being therefore real over the central part from $x = -\sqrt{a/b}$ to $x = +\sqrt{a/b}$, and imaginary from each extremity of this central range outward to infinity. In the usual way, we derive the equations:

$$y = f(x)g(t), \qquad g = A \cos \nu t + B \sin \nu t,$$
$$d^2f/dx^2 + \nu^2(a - bx^2)f = d^2f/dx^2 + (C - x^2)f = 0, \qquad (154)$$

and it is incumbent upon us to solve the equation [11] for $f(x)$.

[11] The constant $\nu^2 b$ has been equated to unity, which entails no loss of generality.

Essay a solution in the form of a power-series, multiplied by $e^{-(1/2)x^2}$:

$$f(x) = e^{-(1/2)x^2} \sum_{n=0}^{\infty} a_n x^n. \tag{155}$$

Substitute this into the differential equation, and group all the terms involving the same power of x. For each such group, we have

$$a_{n+2}(n + 1)(n + 2)x^n - a_n(2n + 1 - C)x^n, \tag{156}$$

and equating each group separately to zero, we arrive at the relation

$$a_{n+2}/a_n = (2n + 1 - C)/(n + 1)(n + 2). \tag{157}$$

Put $a_0 = 0$, thus causing all the even-numbered coefficients to vanish; assign any arbitrary value to a_1, and calculate the odd-numbered coefficients a_3, a_5, a_7, and so onward. Or, put a_1 and all the odd-numbered coefficients equal to zero, assign any arbitrary value to a_0, and calculate the even-numbered coefficients a_2, a_4, a_6, and so onward. Either way we shall get a solution of (154), whatever the value of the parameter C; but there are certain specific values of C which admit a peculiar sort of solution. It is, in fact, evident from (156) that we shall arrive at two entirely distinct results, according as C is or is not equal to some value of $(2n + 1)$—according, that is to say, as C is or is not an odd integer. For, if C is equal to an odd integer $(2n + 1)$, the chain of coefficients will come to an abrupt end at the member having that particular value of n; it and all the succeeding members will be zero; the power-series in the tentative (and adequate) expression (155) for the unknown function $f(x)$ will consist of a finite number of terms. But, if C is not equal to an odd integer, the power-series will go on forever.

Here we have a new kind of *Eigenwert*. If the parameter C, in the differential equation for the curious kind of "string" which I have just defined, has for its value one of the numbers:

$$C = 2n + 1, \qquad n = 0, 1, 2, 3, 4, \cdots, \tag{158}$$

the equation enjoys a special sort of solution. If the parameter does not have one of these *Eigenwerte*, the solution of the differential equation is altogether different.

Let us see what difference these *Eigenwerte* make in the general solution (155) of the differential equation. If the parameter C has some other value than one of these, the series $a_n x^n$ goes on forever; and as x approaches infinity, the value of its summation increases at such a rate as to overwhelm the steadily declining factor $e^{-(1/2)x^2}$, so that the function $f(x)$ is infinite at both ends of the range $-\infty < x < \infty$.

If however C is equal to one of the *Eigenwerte*, the series $a_n x^n$ comes to an abrupt end; and as x approaches infinity, the decline of the factor $e^{-(1/2)x^2}$ overpowers the increase of the summation, and $f(x)$ remains finite at infinity. The values 1, 3, 5, $\cdots$ of the constant C are therefore the *Eigenwerte* which permit solutions *which remain finite all through the range of values of the independent variable from positive to negative infinity.* This condition replaces the boundary-conditions applied to the ordinary stretched string.

The *Eigenfunktionen* are:

$$f_m(x) = e^{-(1/2)x^2} H_m(x), \tag{159}$$

the symbol $H_m(x)$ standing for the finite series $\sum a_n x^n$ constructed according to the rules of the foregoing paragraphs, and terminating at the mth term. These are known as the *polynomials of Hermite.*[12]

Interpretation of the Simple-Harmonic Linear Oscillator by Wave-Mechanics

The foregoing section contains all that is necessary to Schroedinger's theory[13] of the linear simple-harmonic oscillator—an object, or a concept, famous in the history of the quantum-theory; for it was the linear oscillator which Planck first "quantized"—of which, that is to say, Planck first proposed that it be endowed with the power of receiving and retaining and disbursing energy only in fixed finite amounts; thereby arriving at an explanation of the black-body radiation-law, and founding the quantum theory.

Conceive a particle of mass m, constrained to move along the x-axis, attracted to the origin by a force $-k^2x$ proportional to its displacement, and consequently prone to oscillate to and fro across the origin with frequency $\nu_0 = k/2\pi\sqrt{m}$. Its potential energy is the following function of x:

$$V = \tfrac{1}{2}k^2x^2 = 2\pi^2 m\nu_0^2 x^2. \tag{160}$$

The wave-equation assumes the form

$$\frac{d^2\Psi}{dx^2} + \frac{8\pi^2 m}{h^2}(E - 2\pi^2 m\nu_0^2 x^2)\Psi = 0. \tag{161}$$

A simple change of variable ($q = x \cdot 2\pi\sqrt{m\nu_0/h}$) transforms this into the equation (154):

$$d^2\Psi/dq^2 + (C - q^2)\Psi = 0; \qquad C = 2E/h\nu_0. \tag{162}$$

[12] The first five are written down by Schroedinger, *Ann. d. Phys.*, **79**, p. 515 (1927). An arbitrary numerical multiplier remains at disposal.

[13] Schroedinger, *Ann. d. Phys.*, **79**, pp. 514–519 (1926); for the general case in which the restoring-force is not supposed to vary as the displacement, consult H. A. Kramers, *ZS. f. Phys.*, **39**, pp. 828–840 (1926).

According to Schroedinger *the Stationary States of the linear oscillator are distinguished by the energy-values which cause this equation to have a solution finite at all values of the variable, infinity included.*

These are the values of the constant C which cause the parameter C to take one of the *Eigenwerte* set down in (158).

The energy-values of the Stationary States should therefore be

$$\begin{aligned} E_n &= \frac{h\nu_0}{2}(2n+1) \\ &= \frac{1}{2}h\nu_0,\ \frac{3}{2}h\nu_0,\ \frac{5}{2}h\nu_0,\ \cdots. \end{aligned} \tag{163}$$

The successive permitted energy-values of the linear simple-harmonic oscillator of frequency ν_0, the energy-values of its consecutive Stationary States, are therefore specified by wave-mechanics as the products of the fundamental factor $h\nu_0$ by the consecutive "half-integers" 1/2, 3/2, 5/2, and so onward.

The linear simple-harmonic oscillator thus furnishes an instance of "half-quantum-numbers." In most of the earlier theories it was either assumed or inferred that this "Planck" oscillator displayed "whole quantum-numbers"—that its permitted energy-values were the products of $h\nu_0$ by the successive integers 1, 2, 3, 4, $\cdots$. However, in the interpretation of certain features of band-spectra by the assumption that the two atoms of a diatomic molecule vibrate as linear oscillators along their line of centres, the half-quantum-numbers sometimes led to better agreement with experience than did the whole-quantum-numbers.

The *Eigenfunktionen* corresponding to the consecutive Stationary States are these:

$$\Psi_n(x) = \text{const} \cdot e^{-2\pi^2 m\nu_0 x^2/h} H_n(2\pi x\sqrt{m\nu_0/h}). \tag{164}$$

The first five of these *Eigenfunktionen* are exhibited in Fig. 2. These curves may be regarded, if the reader so chooses, as the stationary-wave patterns of "loops" and "nodes," exhibited by five resonating strings along which the wave-speed varies according to the five laws obtained by assigning the first five values given by (163) to the constant E in the equation:

$$u = \frac{E}{\sqrt{2m(E - 2\pi^2 m\nu_0^2 x^2)}}. \tag{165}$$

The various Stationary States of a linear oscillator are therefore imaged not as the fundamental and the overtones of one and the same string, but as the fundamental (and exclusive) nodes of vibration of distinct

strings. It is important to realize this. Schroedinger's way of thinking provides not a single atom-model for each sort of atom, but as many distinct models as there are Stationary States.*

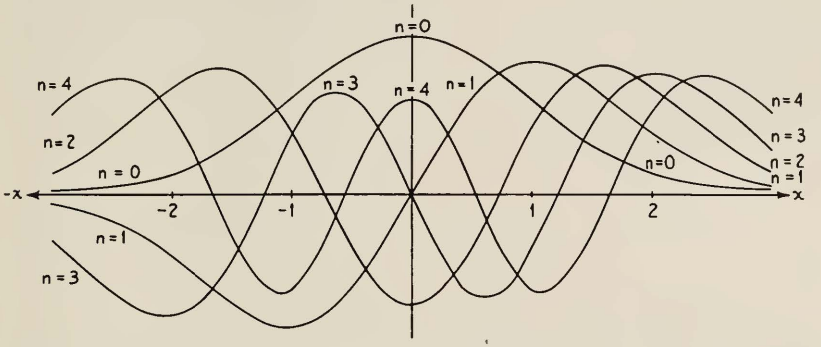


Fig. 2 (after Schroedinger).

Interpretation of the Hydrogen Atom by Wave-Mechanics

The hydrogen atom is conceived as a system endowed with the potential energy $V = -e^2/r$. This form for the potential energy, I recall, is obtained by imagining an electron and a nucleus, or more precisely two point-charges $+e$ and $-e$, separated by a distance denoted by r. The image of the electron and the nucleus does not come over explicitly into the new theory; but in spirit it does come over, for the potential-energy-function derived from that image is the basis for the new theory.

Polar coordinates for the wave-equation are imperiously suggested by a potential-energy-function of this form, and consequently it is thus expressed:

$$\frac{E^2}{2m(E + e^2/r)} \nabla^2\Psi = \frac{d^2\Psi}{dt^2}, \tag{171}$$

and putting E/h for the vibration-frequency, we attain

$$\nabla^2\Psi + \frac{8\pi^2 m}{h^2}\left(E + \frac{e^2}{r}\right)\Psi = 0. \tag{172}$$

The resemblance of these equations to those laid down for the ball of fluid is as unmistakable as the resemblance of the wave-equation for a linear oscillator to that of a stretched string. Here we have the case of a fluid in which the wave-speed varies from point to point, according to the law

$$u^2 = E^2/2m(E + e^2/r), \tag{173}$$

* Some may find satisfaction in conceiving, as my colleague Dr. T. C. Fry suggests, a "string" so constructed that the speed of propagation of waves along it is a function of their frequency.

and we meet the problem of finding modes of vibration and stationary wave-patterns.

If E is supposed positive, the wave-speed is everywhere real. Boundary-conditions of the usual sorts (*e.g.*, the prescription that the fluid shall be confined within a rigid spherical wall of given radius) might be imposed, and then *Eigenwerte* of the constant E could be calculated, and from these the wave-patterns and natural frequencies of the fluid. If no such boundary-conditions were prescribed, the equation (172) could be solved with any value of E.

If E is supposed negative, the whole state of affairs is changed. The wave-speed is now real within the sphere of radius $-e^2/E$, zero over this sphere and imaginary beyond it. This recalls the case of the "string" proposed as an analogy for the linear oscillator, for which the wave-speed was real along its central segment and imaginary from each end of its central segment onwards to infinity. There are important differences: in the present case, the variable r assumes positive values only, and the wave-speed at $r = 0$ is infinite though real.

In the case of the "string" with variable and at some points imaginary wave-speed, we found that the law of variation of wave-speed could be so chosen that the "string" enjoys a natural mode of vibration with a stationary wave-pattern and a natural resonance-frequency. This was done by selecting any of a series of *Eigenwerte* for a parameter of the differential equation. Here we shall do likewise.

Essaying for the function Ψ in (172) a solution in the form of a product of a function of θ and φ exclusively by a function of r exclusively, we arrive in the familiar way at differential equations:

$$\operatorname{cosec}\theta\left\{\frac{d}{d\theta}\left(\sin\theta\frac{dY}{d\theta}\right)+\frac{d}{d\varphi}\left(\operatorname{cosec}\theta\frac{dY}{d\varphi}\right)\right\}=-\lambda Y. \quad (175)$$

$$\frac{d}{dr}\left(r^2\frac{df}{dr}\right)+\frac{8\pi^2 m r^2}{h^2}\left(E+\frac{e^2}{r}\right)f=+\lambda f, \quad (174)$$

The equation (175) is the identical one which we encountered in the case of the ball of fluid. Here, as there, the fact that the variables θ and φ are cyclic requires *Eigenwerte* of the constant λ:

$$\lambda = l(l+1), \qquad l = 0, 1, 2, 3, 4, \cdots \quad (176)$$

Equation (174), however, is not the same as the corresponding equation (133) of the prior case; here we find the difference between the fluids of actual experience and the "imaginary fluid" which is to serve as material for the atom-model supplied by wave-mechanics for hydrogen.

If in that equation (174) one were to assign an arbitrarily chosen

negative value to the parameter E, one would in general not be able to find a solution which is finite both at the origin and at infinity. This is the same situation as occurred in the theory of the linear oscillator, where an arbitrary choice of a value for the parameter there called C would in general have led to a solution implying infinite amplitude at both ends of the "string."

Schroedinger however discovered [14] that there is a series of *Eigenwerte* for the parameter E, each of which (subject to a limitation to be introduced below) entails a solution which is single-valued, continuous and finite over the entire range of the variable r.

These *Eigenwerte* are the following:

$$E_n = -2\pi^2 me^2/h^2 n^2; \qquad n = 1, 2, 3, 4, \cdots. \tag{177}$$

The consecutive permitted energy-values of the system of potential-energy-function $-e^2/r$, *the Stationary States of the model for the hydrogen atom, are therefore specified by wave-mechanics as the quotients of the fundamental factor* $-2\pi^2 me^4/h^2$ *by the squares of the consecutive integers from unity onward.*

These agree with experiment. The formula (177) is in fact the renowned formula of Bohr, from which the whole contemporary theory of spectra sprang; a formula so successful that it is scarcely conceivable that any alternative theory should ever win acceptance unless by presenting the identical equation over again.

Schroedinger's models for the hydrogen atom in its various Stationary States thus are imaginary fluids each pervading the whole of space, and in each of which the wave-speed depends on the distance r from a centre, according to a peculiar law—the law obtained by inserting into the formula (173) the appropriate value for E, chosen from the sequence given in (177). If into (173) we were to put any value chosen at random for the energy-constant E, we should be inventing an imaginary fluid; but, in general, this fluid would not be capable of sustaining a continuous stationary wave-pattern of finite amplitude. Only when one of Bohr's sequence of energy-values is chosen do we get a fluid able to resonate as a ball of actual physical substance can.

The next task is to enquire into the wave-patterns in the imaginary fluids corresponding to these various permitted energy-values. This is much more difficult than the same problem for the imaginary strings corresponding to the various permitted energy-values of the linear oscillator, and the new complexities are not altogether due to the fact that we now have three dimensions to deal with instead of one; they

[14] Schroedinger, *Ann. d. Phys.*, **79**, pp. 361–376 (1926). For an alternative method of proof see A. S. Eddington, *Nature*, **120**, p. 117 (1927).

are due chiefly to the fact that the system is mathematically "degenerate." Owing to this circumstance there are more than one possible mode of vibration, more than one stationary wave-pattern, for each (except the first) of the permitted energy-values. To describe these it is necessary to consider both of the equations (174) and (175).

Since the equation (175) is identical with the corresponding equation derived for balls of actual physical fluids, the modes of vibration for Schroedinger's atom-model are identical with the modes of vibration of actual fluid spheres *insofar as the dependence on angle is concerned.* The imaginary fluid is divided into compartments by nodal planes, nodal double-cones and nodal spheres; and the division by planes and double-cones is identically such as we should find in the corresponding mode of vibration of an actual fluid ball; it is only the division by nodal spheres which differs.

To the first *Eigenwert*, E_1 ($n = 1$) there corresponds a single *Eigenfunktion* of equation (174); to the second, E_2, a pair; to the third, three; and so forth. This multiplicity is linked with the limitation upon the *Eigenwerte* which was foreshadowed above. In the expression for the function Ψ as a product of functions of the individual variables

$$\Psi(r, \theta, \varphi) = F(r)\, Y_l(\theta, \varphi), \tag{178}$$

if we assign an *Eigenwert* E_n to the parameter E in the first factor according to (174), we have still a choice of values to assign to the parameter l in the second factor according to (176). This choice however is limited. We must not take any value of l as great as or greater than the value adopted for n; otherwise the value of E_n would not be an *Eigenwert* in the sense adopted. Thus for $n = 1$ we are restricted to the choice $l = 0$; for $n = 2$ we have the alternative of $l = 0$ or $l = 1$; for $n = 3$ the option of $n = 0$, 1, or 2, and so forth. Each *Eigenwert* E_n thus admits $(n - 1)$ distinct spherical harmonics $Y_1(\theta, \varphi)$, $Y_2(\theta, \varphi) \cdots Y_{n-1}(\theta, \varphi)$ as solutions of equation (175); and to each of these there corresponds, with each of these there goes, a distinct *Eigenfunktion* $F_{n,\,l}(r)$ of the equation (174), which is expressed as follows in terms of a variable $\rho = \dfrac{2\pi\sqrt{-2mE_n}}{h}\, r = \dfrac{4\pi^2 me^2}{nh^2}\, r \equiv \dfrac{1}{na_0}\, r$ instead of r to make the function seem less intricate:[15]

$$X_{n,\,l}(\rho) = \text{const.}\ \rho^{+l} e^{-\rho} \sum_{k=0}^{n-l-1} \frac{(-2\rho)^k}{k!} \binom{n+l}{n-l-1-k}, \tag{180}$$

[15] The factor in parentheses in equation (180) stands for the "number of combinations of $(n + l)$ quantities taken $(n - l - 1 - k)$ at a time," which is the $(n - l - 1 - k)$th coefficient in the binomial expansion of $(a + b)^{l+n}$.

The function $X_{n,\,l}(\rho)$ has $(n - l - 1)$ roots, so that the corresponding mode of vibration has $(n - l - 1)$ nodal spheres. To each permitted energy-value E_n there consequently correspond n different solutions of the general equation (172), differing from one another in respect of the number of nodal spheres:

$$\Psi_{n,\,l}(r,\,\theta,\,\phi) = X_{n,\,l}(\rho)\,Y_1(\theta,\,\phi); \qquad l = 0,\,1,\,2\,\cdots\,(n-1). \quad (181)$$

Each of these describes a permitted class of modes of vibration, owing to the subdivision of the spherical harmonic Y_1 into terms according to (135).

Allowing for the subdivision of the spherical harmonics, there are $(1 + 2 + 3 + \cdots n) = n(n + 1)/2$ modes of vibration for the nth permitted energy-value E_n.

The equation (181) exhibits the various modes of vibration of which our imaginary "fluid," the model for the hydrogen atom, is capable. It would be possible to describe these with a wealth of verbal detail. I hesitate to do so; for vast amounts of industry and ink have been expended during the last twelve years in tracing and describing electron-orbits, which are now quite out of fashion; and who dares affirm that in another five years the vibrating imaginary fluid will not be *démodé*? Yet it is altogether probable that for some years to come, if not for all time, the image of the vibrating fluid will furnish the customary symbolism for expressing the data of experiment. Therefore let me point out some features of the vibrations corresponding to the first (or "lowest," or "deepest") three states of the hydrogen atom:

Normal State, $n = 1$. One *Eigenfunktion*, $X_{1,\,0}(\rho)$; an exponential function of r, decreasing steadily from the origin to infinity, with no nodal spheres. Corresponding spherical harmonic $Y_0(\theta,\,\varphi)$,—a constant. The vibration consequently is described by

$$\Psi(r) = \text{const. } e^{-r/a_0} \quad (a_0 = h^2/4\pi^2me^2) \qquad (182)$$

and is endowed with perfect spherical symmetry.

First Excited State, $n = 2$ (the state into which the atom relapses after emitting any line of the Balmer series). Two *Eigenfunktionen* $X_{2,\,0}$ and $X_{2,\,1}$; the first represents a vibration with a single nodal sphere, the second a vibration diminishing steadily in amplitude from the origin outward. The first is to be multiplied by $Y_0(\theta,\,\varphi)$ to obtain the complete description of the vibration; Y_0 being a constant, this mode is endowed with perfect spherical symmetry. The second is to be multiplied by Y_1, which is a combination of terms written out

in equation (135); the permissible mode, or rather modes, of vibration involve nodal planes and double-cones, which the reader may work out for himself with the aid of (135).

Second Excited State, $n = 3$ (the state from which the atom departs when it emits the line *H*-alpha). Three *Eigenfunktionen* $X_{3,0}$, $X_{3,1}$ and $X_{3,2}$. The first corresponds to a vibration with two nodal spheres, and perfect spherical symmetry. The second and third correspond to vibrations with one nodal sphere, and with a steady diminution of amplitude from the centre outward, respectively; but being multiplied with the spherical harmonics Y_1 and Y_2, they describe modes which are not endowed with spherical symmetry, and involve nodal double-cones and nodal planes.

Generally: the state distinguished by the numeral n enjoys n distinct *Eigenfunktionen*, describing vibrations having respectively 0, 1, 2, 3 $\cdots$ $(n-1)$ nodal spheres; to the *Eigenfunktion* with the maximum number of nodal spheres corresponds a single mode of vibration which is spherically symmetric, to the others various modes with varying members of nodal double-cones and planes.

If this is destined to be the "language of the future" for describing the data of experiment, it will be necessary to have dictionaries for translating it out of (or into) the "language of the present," the vocabulary of the Bohr-Sommerfeld atom-model in which Stationary States are represented by electron-orbits. They will contain definitions such as these: the numeral n is the total-quantum-number of the electron-orbits—the numeral l is one unit smaller than the azimuthal quantum-number k of the electron-orbit—the numeral $(n - l - 1)$, to which the number of nodal spheres is equal, is the radial quantum-number of the electron-orbit. To elucidate these "definitions" of the future dictionary, I recall that the Bohr-Sommerfeld atom-model provided, for the hydrogen atom in its state of energy-values E_n, a family of n distinct electron-orbits, of which one is circular while the other $(n-1)$ are ellipses of varying degrees of eccentricity.[16] These ellipses were selected by laying down the conditions, that the integral $\int p_\varphi d\varphi$ of the angular momentum p_φ around the orbit shall be equal to the product of h by some integer k equal to or less than the prescribed n; and the integral $\int p_r dr$ of the radial momentum $\int p_r dr$ shall be equal to the product of h by the integer $(n - k)$; so that the sum of the integrals $\int p_\varphi d\varphi$ and $\int p_r dr$ shall be equal to the product of h by n. The quantities n, k and $n - k$ were given the names *total*, *azimuthal*, *radial quantum-number*. "Defini-

[16] The introduction some twenty months ago of the "spinning electron" caused a modification of this picture; for those who accept the modification, it is the "language of antiquity" which is compared in this paragraph with the "language of the future."

tions" such as those above (which are not necessarily the only self-consistent nor the best ones) make it possible to translate orbits of the orbit-model into modes of vibration of the wave-model, and vice versa; and to devise definitions for these three kinds of quantum-numbers from the qualities of the vibrations themselves.

Perturbations

Inasmuch as the wave-mechanics indicates n different *Eigenfunktionen* with n different collections of nodal spheres (not to speak of the still more greatly varied possibilities of nodal planes and double-cones) for the Stationary State having the *Eigenwert* and energy-value E_n, one may well ask whether there is any chance of distinguishing which of these, or which linear combination of these (for the differential equation will permit any) is actually adopted by a hydrogen atom.

Translating into the language of the Bohr-Sommerfeld atom-model, we find the question in this form: is there any way of distinguishing which of the n permitted elliptical electron-orbits is actually adopted?

When the question was asked in this form, it was answered by pointing out that if the force exerted upon the electron were not the pure inverse-square force ascribed to the nucleus, but the sum of this and a *perturbing force*, the energy-values of the n permitted ellipses would cease to coincide exactly. If for instance the atom under examination were composed of a nucleus of charge $11e$, a group of ten electrons very close to it and an "outer" electron relatively far out (the conventional model for a sodium atom in certain states); then the group of ten inner electrons would act upon the outer one with a perturbing force, and the n permitted ellipses of the outer one would be endowed with distinct energy-values—the single Stationary State of the outer electron would be dissolved into n distinguishable states. Even in the hydrogen atom, the dependence of the mass of the electron upon its speed should separate the energy-values of the various ellipses which but for this fact would share a common energy E_n, and produce the fine-structure of the hydrogen lines.

The very same thing occurs in wave-mechanics; and from the effect of a perturbing force, allowance for which is made in the potential-energy-function introduced into the wave-equation, we may expect to be able to distinguish the different modes of vibration attributed to a single *Eigenwert* and a single Stationary State of the unperturbed hydrogen atom.[17]

[17] In the language of the mathematicians, the perturbing forces remove the degeneracy of the problem; some kinds of perturbation remove it completely, others in part.

The results are, in fact, just like those obtained with the Bohr-Sommerfeld atom-model; and this is somewhat embarrassing. For, in order to perfect the Bohr-Sommerfeld model and establish a complete analogy between (for instance) the sodium spectrum on the one hand and the fine-structure of the hydrogen lines on the other hand, it was necessary to introduce a new feature—the "spinning electron." Something of the sort must evidently be done again—the "spinning electron" must be imported into the undulatory mechanics; but the exact way to do it seems as yet to elude the virtuosi of mathematical physics.[18]

In one case—when the perturbing force is an impressed electric field—the results obtained by the method of Bohr and Sommerfeld and those obtained by the method of Schroedinger agree to first approximation with each other and with the data of experience, without the introduction of a "spinning electron." As this case of the "Stark Effect" furnishes a convenient transition to the last section of the article, I will quote the results.[19]

The Stark Effect

Imagine a hydrogen atom, upon which an electric field F parallel to some arbitrary direction which we call the z-direction is acting. Owing to this field, the electron at the point x, y, z and the nucleus at the origin (we are still using the concept of the nucleus and the electron!) possess a potential energy composed of the "intrinsic" term $-e^2/r$ and the "perturbation" $+eFz$. The wave-equation takes the form:

$$\nabla^2\Psi + \frac{8\pi^2 m}{h^2}\left(E + \frac{e^2}{r} - eFz\right) = 0. \qquad (183)$$

Paraboloidal coordinates are indicated for this problem. Instead of the planes, double-cones and spheres of the polar coordinate-system which we earlier used, it is desirable to employ planes and two families of paraboloids of rotation; the planes intersect one another along the line through the nucleus parallel to the field (hitherto called the z-axis), and the two families of paraboloids have their common foci at the nucleus and their noses pointing opposite ways along that axis. The transformation is made by the equations:

$$x = \sqrt{\xi\eta}\cos\varphi, \qquad y = \sqrt{\xi\eta}\sin\varphi, \qquad z = \tfrac{1}{2}(\xi - \eta) \qquad (184)$$

[18] Unless the problem has been solved by C. F. Richter (cf. preliminary note in *Proc. Nat. Acad. Sci.*, **13**, pp. 476–479; 1927).

[19] Schroedinger, *Ann. d. Phys.*, **80**, pp. 457–464 (1926); P. S. Epstein, *Phys. Rev.* (2), **28**, pp. 695–710 (1926).

and the wave-equation appears in this guise:

$$\frac{d}{d\xi}\left(\xi\frac{d\Psi}{d\xi}\right)+\frac{d}{d\eta}\left(\eta\frac{d\Psi}{d\eta}\right)+\frac{1}{4}\left(\frac{1}{\xi}+\frac{1}{\eta}\right)\frac{d^2\Psi}{d\varphi^2} + \frac{2\pi^2 m}{h^2}\left[E(\xi+\eta)+2e^2-\tfrac{1}{2}eF(\xi^2-\eta^2)\right]\Psi = 0. \tag{185}$$

Essaying as tentative solution a product of a function of φ by a function of ξ and a function of η, we obtain as usual three differential equations, involving E and two other parameters, to which specific *Eigenwerte* must be assigned either because the variable φ is cyclic, or because for values other than these *Eigenwerte* the solutions become infinite for certain values of the variable.

Suppose that we set $F = 0$, and ascertain these *Eigenwerte*, and insert them into the equations: we then find the imaginary fluid vibrating in a stationary wave-pattern, oscillating in compartments divided from one another by nodal planes and by nodal paraboloids pointing up or down the field. To each of the energy-values E_n there correspond $(1 + 2 + 3 + \cdots n)$ distinct wave-patterns, each having a distinctive number k_1 of nodal paraboloids of the one family, a distinctive number k_2 of nodal paraboloids of the other family, and a distinctive number s of nodal planes; the values of k_1 and k_2 and s are limited by the conditions that they must be integers, that they cannot be less than zero nor greater than n, and that their sum must be equal to $(n - 1)$: that is,

$$k_1 + k_2 + s + 1 = n. \tag{186}$$

(Translating into the language of the electron-orbits, we find that s becomes the equatorial quantum-number which represents the angular momentum of the electron around the direction of the field (in terms of the unit $h/2\pi$) and k_1 and k_2 become the parabolic quantum-numbers.)

Introducing now the impressed electric field F, we find that among the $(1 + 2 + 3 + \cdots n)$ modes of vibration which originally shared the energy-value E_n, those for which $k_1 = k_2$ retain this energy-value, while the rest are displaced by varying amounts given by the celebrated Epstein formula:

$$\Delta E = \frac{3Fh^2n}{8\pi^2me}(k_1 - k_2). \tag{187}$$

The Stationary State of energy-value E_n is thus "resolved" or "split" into several—not, however, into the full number $(1 + 2 + 3 + \cdots n)$ corresponding to the total number of modes of vibration, for some of these still share identical energy-values. The line resulting from the

transition between two States, E_3 and E_2 for instance, is thus resolved into a set of lines lying close together. These individual "Stark-effect components" testify to the individual existence of the several distinct modes of vibration which, when there is no impressed electric field, should share a common energy-value E_n and be indistinguishable from one another.[20]

In the closing section we shall consider another aspect of these Stark-effect components. At this point I wish only to allude to a quaint little paradox which may already have disconcerted the reader. I have just said that the imaginary "fluid" executes stationary vibrations in which it is divided into compartments by nodal planes and nodal paraboloids, even when the impressed field F is made equal to zero; but earlier I said that the "fluid" representing the unperturbed hydrogen atom executes vibrations in which it is divided into compartments by nodal planes, nodal double-cones and nodal spheres. There is no actual contradiction between these two assertions; for a mode of vibration of the one kind can be obtained by superposing two or more modes of vibration of the other kind, with a proper distribution of relative amplitudes. Take the specific case of the "first excited state" of the hydrogen atom, $n = 2$. By the earlier process, we find three wave-patterns: (a) with one nodal sphere, (b) with one nodal double-cone, (c) with one nodal plane. By the later process, we find three wave-patterns: (α) with one nodal paraboloid facing one way; (β) with one nodal paraboloid facing the other way; (γ) with one nodal plane. The wave-patterns (c) and (γ) are evidently the same, while either (α) or (β) can be reproduced by superposing (a), (b) and (c) with the proper relative amplitudes.[21] If the field F acting upon a hydrogen atom in the first excited state were to be gradually reduced to zero, it would leave the atom, or to speak more carefully the "imaginary fluid," vibrating in a manner which would be one of the modes (α), (β) or (γ), hence a cleverly adjusted superposition of the three modes (a), (b) and (c). Suppose however that a very small field F were to be applied to a hitherto unperturbed atom; why should it necessarily find ready for it a vibration with precisely the proper relative adjustment of the modes (a), (b) and (c)? or if it did not, if it should find the atom vibrating say in mode (α), how would it persuade the "fluid" to change over into the manner of vibration suitable for its own operations to begin?[22]

[20] A couple of "contour maps" of the wave-patterns for two of these paraboloidal modes of vibration are given by F. G. Slack (*Ann. d. Phys.*, **82**, pp. 576–584; 1927).

[21] I have not actually proved this, but believe that it must follow from Schroedinger's general theorem.

[22] This same curious thing occurs in a somewhat different guise when the electron-orbit theory is adopted.

Interpretation of the Rotator by Wave-Mechanics

The rotator or rotating body, the "spinning-top" as the Germans often call it, is a very important object in the workshop of the builder of atom-models. It is the accepted molecule-model used in theorizing about the polarization of gases by electric and magnetic fields, and the basis of the accepted molecule-model used in explaining the band-spectra of diatomic and polyatomic gases. Most models devised for the latter purpose combine the features of the rotator and the linear oscillator; but for the present purpose it is sufficient to view these separately, conceiving the rotator as a perfectly rigid whirling body.

The treatment of the rotator by wave-mechanics is in one respect admirably simple, but eventually we are led into the mazes of the General Equation with its non-Euclidean geometry. One can however avoid the complexity long enough to benefit by the intelligible feature, by considering first a rotator such as was invented more than fifty years ago to account for the specific heats of diatomic gases such as hydrogen—a dumbbell not permitted to spin around its own axis-of-figure or line-of-centres, but revolving around some axis passing through its center-of-mass perpendicular to its line-of-centres. The orientation of such a dumbbell is specified by the angles θ and ϕ which define, in a polar coordinate-system, the direction in which its axis-of-figure is pointing. The energy is exclusively kinetic, so that the term containing V vanishes from the wave-equation, a circumstance which is very helpful. Representing by A the moment of inertia of the dumbbell about the axis of revolution, we find the wave-equation in the form:[23]

$$\nabla^2\psi + \frac{8\pi^2 EA}{h^2}\psi = 0. \qquad (190)$$

In this equation the Laplacian operator is to be expressed in the polar coordinates θ and ϕ, as it was expressed in equation (131), but without the terms involving the third and missing coordinate r. We have before us, therefore, the second of equations (133), with a specific value for the constant there called λ:

$$-\operatorname{cosec}\theta\left[\frac{d}{d\phi}\left(\operatorname{cosec}\theta\frac{d\psi}{d\phi}\right)+\frac{d}{d\theta}\left(\sin\theta\frac{d\psi}{d\theta}\right)\right]=\frac{8\pi^2 EA}{h^2}\psi. \qquad (191)$$

Here, as there, the function ψ must repeat itself whenever θ is altered by any multiple of π and ϕ by any multiple of 2π; for then we are back at the same place, i.e. at the same orientation of the rotator.

[23] Schroedinger, *Ann. d. Phys.*, **79**, pp. 520–522 (1926).

Here, as there, this necessity imposes of itself a certain condition upon the coefficient of ψ in the right-hand member, which is tantamount to this condition imposed on E:

$$E = n(n+1)\frac{h^2}{8\pi^2 A} = (n+\tfrac{1}{2})^2\frac{h^2}{8\pi^2 A} + \text{const.}, \qquad n = 0, 1, 2, 3 \cdots. \quad (192)$$

The energy of the rotator is thus *by the mere fact that the variables are cyclic* limited to a single sequence of permitted values, furnishing incidentally another example of half-quantum-numbers.

The *Eigenwerte*, the permitted energy-values, are thus for the rotator determined by an exceptionally lucid condition; yet the complications of the General Equation already appear on the horizon. Equation (190) differs from the wave-equation which I have hitherto used by virtue of the substitution of moment-of-inertia A for mass m. This replacement seems sensible enough; one might rely on intuition in this particular case; but strictly it is caused by the form preassumed for the General Wave-Equation and by the specific form of the kinetic-energy-function for this specially restricted kind of rotator. If now we remove the restriction, and permit the rotator to spin about its axis-of-figure at the same time as it whirls about some axis normal to that—if we pose the general problem of the rigid rotator unrestricted save by the conditions which the wave-equation imposes, it is necessary to invoke the General Equation with the non-Euclidean geometry. The problem is soluble, and has been solved;[24] the utility of the results for the interpretation of band-spectra gives valuable support to the form selected by de Broglie and Schroedinger for the General Equation.

The polarization of a gas by an electric (or magnetic) field may be treated by supposing that each atom is an electric (or magnetic) doublet. The treatment is simplest if we may assume that the electric (or magnetic) axis of the doublet coincides with the axis-of-figure of a dumbbell-molecule, not allowed to spin around its axis-of-figure. Let M stand for the moment of such a doublet, and suppose the field H to be parallel to the direction from which the angle θ of the foregoing paragraphs is measured. The field supplies the potential-energy term to be added to the left-hand member of equation (190); this new term is:

$$-V\psi = (MH\cos\theta)\psi. \quad (193)$$

It is easy to see that the wave-equation has *Eigenwerte*, so that the atoms are in effect limited to certain "permitted" orientations in the

[24] F. Reiche, *ZS. f. Phys.*, **39**, pp. 444–464 (1926); R. de L. Kronig, I. I. Rabi, *Phys. Rev.* (2), **29**, pp. 262–269 (1927).

field—a conclusion from the earlier atomic theory, which for magnetic fields has become a fact of experience through the experiments of Gerlach and Stern and others. To calculate the polarization of a gas, it is necessary to make a further assumption concerning the relative probabilities of these various orientations in a gas in thermal equilibrium; having done so, one obtains a formula for the polarization, or the dielectric constant, or the susceptibility of the gas as function of applied field and temperature. The customary assumption leads to a formula which, in the limiting case of high temperature and low field, agrees with the celebrated equation of Langevin for the polarization of a paramagnetic gas by a magnetic field:[25]

$$\text{Susceptibility} = I/H = NM^2/3kT. \tag{194}$$

Interpretation of the Free Electron in Wave-Mechanics

We now depart from the calculation of *Eigenwerte* and Stationary States, and return to the original ideas of de Broglie.

For a free electron moving in a field-free region—or any particle moving in a region where no force acts upon it—with a constant speed v along a direction which I will take as the x-direction, the (non-relativistic) wave-equation assumes the form:

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 mE}{h^2}\psi = 0 \qquad (E = \tfrac{1}{2}mv^2). \tag{195}$$

This equation admits a sine-function as its solution whatever the value of the constant E and consequently does not restrict the energy-values which the electron is allowed to take (a contrary result would have been hard to swallow!). Assigning the value E/h to the frequency of the sine-wave and the value $E/\sqrt{2mE}$ to its speed, we obtain for the wave-length of the wave-train, "associated with" a free electron (or free particle) of mass m and speed v, this value:

$$\lambda = \frac{E/\sqrt{2mE}}{E/h} = \frac{h}{\sqrt{2mE}} = \frac{h}{mv}. \tag{196}$$

For electrons of such speeds as ordinarily occur in discharge-tubes, these wave-lengths are of the magnitude of X-ray wave-lengths; for instance, a 150-volt electron is associated with a wave-length of very nearly one Angstrom unit.

[25] C. Manneback, *Phys. ZS.*, **28**, pp. 72–84 (1927); J. H. Van Vleck, *Phys. Rev.* (2), **29**, pp. 727–744; **30**, pp. 31–54 (1927). For the classical derivation of formula (194) and meaning of the symbols cf. my article "Ferromagnetism," this Journal, **6** (1927), pp. 351–353.

This coincidence makes one wonder whether, if a stream of such electrons were projected against a crystal such as is used for diffracting X-rays, there would be a semblance of diffraction. Nothing yet said about the waves leads inevitably to such an inference. On the contrary, it might well be argued that we have no greater justification for expecting to observe them in the ordinary world of space and time than for expecting the x and the y of an algebraic equation to come to life before our eyes. It might forcibly be pointed out that while in this instance and the instance of the hydrogen atom the "waves" are defined in ordinary space, there are other instances—supplied for instance by rotators—in which the wave-equation is formally similar to (195) and the theory quite as effective, and yet the alleged "waves" exist only in the configuration-space and indeed in non-Euclidean configuration-space, which is much the same as saying that they do not exist at all. Nevertheless it appears that there is indeed a diffraction of electrons by crystals, and that the wave-length indicated by the diffraction-angles is in accordance with the value given by de Broglie! The first evidence for this amazing and portentous effect will be narrated by its discoverers Davisson and Germer in the following issue of this Journal.[26]

Notice that the speed of the associated wave-train is not the same as that of the flying particle; it is $\sqrt{E/2m}$, that of the particle is $\sqrt{2E/m}$. It is, however, the wave-length of the wave-train which is measured by the diffraction-experiments; not the speed, and not the frequency. This is important; for it is the wave-length which is exempt from the consequences of the essential uncertainty in the value of E. In classical mechanics, energy-differences alone are definite, but the absolute values of the "energy" of a system are not defined; the definition of energy involves an arbitrary additive constant. If now we were to add an arbitrary constant to the kinetic energy of the free electron, and call E the sum of the two, we should alter the frequency and alter the speed assigned to the wave-train; but we should not alter the wave-length, for the wave-length is strictly equal to $h/\sqrt{2m(E - V)}$ with V standing for the potential energy of the free electron, and the added constant would enter into V and vanish by subtraction. Returning to the preceding sections of this

[26] Consult meanwhile their note in *Nature*, **119**, pp. 558–560; 1927. The prediction was first published by W. Elsasser (*Naturwiss.*, **13**, p. 711; 1925). For additional intimations of evidence for undulatory qualities in matter cf. G. P. Thomson, A. Reid (*Nature*, **119**, p. 890; 1927); T. H. Johnson, *Nature*, **120**, p. 191 (1927); E. G. Dymond, *Phys. Rev.* (2), **29**, pp. 433–441 (1927). Schroedinger's elegant treatment of the Compton effect is based upon the conception of electrons as wave-trains (*Ann. d. Phys.*, **82**, pp. 257–264; 1927); for a more elaborate treatment of Compton effect cf. W. Gordon (*ZS. f. Phys.*, **40**, pp. 117–133).

article, we see that Schroedinger calculated the energy-values of the Stationary States by conditions imposed upon the wave-lengths of the waves, not upon their frequencies; the wave-patterns depend only upon the wave-lengths, and the frequency of the light which an atom emits in passing between two Stationary States depends only on the difference between their energy-values. In relativistic mechanics, energy is defined absolutely, and this difficulty never even threatens to arise; yet it is worth while to note that the ambiguity of the concept "energy" in classical mechanics does not interfere with, nor is it resolved by, anything which has been observed in Nature and interpreted by wave-mechanics.

In relativistic mechanics, the wave-equation for the free-flying particle assumes the form:

$$\frac{d^2\psi}{dx^2} + \frac{4\pi^2}{h^2c^2}(E - m_0{}^2c^4) = 0 \qquad \left(E = \frac{m_0c^2}{\sqrt{1 - v^2/c^2}}\right). \qquad (197)$$

The wave-length has the value $h\sqrt{1 - v^2c^2}/m_0v = h/mv$; the frequency is $m_0c^2/h\sqrt{1 - v^2/c^2}$; the speed of propagation of the waves is c^2/v, superior to the speed of light.

I can no more than allude to the strangely suggestive fact, that in general as well as in this special case the speed of the particle and the speed of the associated waves are related to one another in the same way as group-speed and wave-speed in ordinary optics.

Attempt to Find a Meaning for the Symbol Ψ

Thirty-three years ago the Earl of Salisbury, invited by reason of his eminence as a statesman to be the President of the British Association for the Advancement of Science, observed the physicists of his day involved in their fervent speculations over the nature of the æther; and of an address brilliant with felicitous phrases the best-remembered words are those by which he described the outcome of their travail: *The main, if not the only, function of the word* aether *has been to furnish a nominative case to the verb 'to undulate.'* Quite the same thing could now be said of the symbol Ψ, insofar as it serves to determine the energy-values of the Stationary States of the systems devised to represent atoms. When it is used for this purpose, it vanishes just as the final triumph is achieved. Like the variable under the sign of integration in a definite integral, it drops out of sight when the calculations which it proposes are actually performed. Indeed it might be discarded altogether during the process of calculating *Eigenwerte* and energy-values; one might speak exclusively of the "differential operator" $\nabla^2 - 8\pi^2m(E - V)/h^2$; many mathematicians do so.

Schroedinger however conceived the daring, the admittedly tentative and still incomplete but very alluring, idea of seeking in Ψ some measure of the density of electric charge. Specifically, it occurred to him to define the square of the amplitude of the oscillations of Ψ, which the *Eigenfunktionen* prescribe as function of the coordinates—to define this squared amplitude as the density of the electric charge, spreading the electron as it were throughout space.

Let us examine this idea, and see whither it leads.

To avoid avoidable complexities as far as possible, I take the simplest of all cases: the linear oscillator, represented by the imaginary "string" stretched along the x-axis, possessed of a wave-speed varying as $\sqrt{1 - x^2/L^2}$, real from the origin both ways as far as the points $x = \pm L$ and imaginary thenceforward. I will also refer to the still simpler "actual" case which served as an introduction to this one: the problem of the stretched string, clamped at its extremities at $x = \pm L$, possessed of a uniform real wave-speed u at all points between.

In both these cases of the imaginary and the real string, the search for the *Eigenwerte* and the *Eigenfunktionen* leads us to diverse natural modes of vibration, executed with various frequencies $\nu_0, \nu_1, \nu_2, \nu_3 \cdots$ and displaying stationary wave-patterns described by the *Eigenfunktionen:*

$$y_i = f_i(x)(A_i \cos 2\pi\nu_i t + B_i \sin 2\pi\nu_i t); \quad i = 0, 1, 2, 3 \cdots \quad (201)$$

For the real string the functions $f_i(x)$ are sine-functions; for the imaginary strings which are the model of the linear oscillator, they are given by (155). I recall once more that in the latter case we have, not distinct modes of oscillation of one string, but the fundamental modes of as many strings as there are Stationary States.

When the real string is vibrating in the ith mode, or when we are dealing with the ith imaginary string, the function $f_i(x)$ is proportional to its vibration-amplitude. The form of equation (201) shows that this amplitude at any fixed point remains constant in time.

If the square of the vibration-amplitude is to be regarded as the density of electric charge along the string, it follows that when the oscillator is in one of its stationary states, and the string therefore vibrating in one of its modes, the density and the distribution of charge remain everywhere constant. There would be no to-and-fro motion of charges, and no tendency to radiation.

Suppose now that the real string is vibrating simultaneously in two modes, the ith and the jth; or that we have both the ith and the jth imaginary string coexisting (this is where the model is clumsiest!).

The vibrations are described by the following formula (I have put $A_i = A_j = 1$ and $B_i = B_j = 0$, which simplifies without injury to the generality of the result):

$$y = y_i + y_j = f_i(x) \cos 2\pi\nu_i t + f_j(x) \cos 2\pi\nu_j t, \tag{202}$$

which is easily transformed thus:

$$y = C \cos (2\pi\nu_i t - \alpha), \tag{203}$$

in which

$$C^2 = f_i^2 + f_j^2 + 2f_i f_j \cos 2\pi(\nu_i - \nu_j)t, \tag{204}$$

and α = a constant not important for our purpose.

Here we have a vibration in which the amplitude at any fixed point varies with time; the square of the amplitude is the sum of a constant term and a sine-function of time, and the frequency of the sine-function is the difference between the frequencies of the two coexisting modes of vibration.

Identifying the square of the amplitude with the density of electric charge, we see that this charge-density varies at each point with the frequency $(\nu_i - \nu_j)$. We might therefore expect radiation of this frequency.

Now the vibration-frequencies ν_i and ν_j corresponding to the modes of vibration, that is to the Stationary States i and j having energy-values E_i and E_j, are respectively E_i/h and E_j/h.

If therefore—to speak in a vague but suggestive fashion—the linear oscillator were simultaneously in two Stationary States, their energy-values being E_i and E_j, then the square of the amplitude of the oscillations of Ψ would be fluctuating at each point of the "imaginary string" with the frequency $(E_i - E_j)/h$; and if this squared amplitude were to be identified with charge-density, then the system might be expected to emit radiation of the frequency $(E_i - E_j)/h$.

Transition between two states would then signify *coexistence* of the two states.[27]

We proceed a step further in the development of this idea, by forming the following integral:

$$M = \int_{-\infty}^{\infty} xC^2 dx = \int_{-\infty}^{\infty} xf_i^2 dx + \int_{-\infty}^{\infty} xf_j^2 dx + \left\{ 2\int_{-\infty}^{\infty} xf_i f_j dx \right\} \cos 2\pi(\nu_i - \nu_j)t. \tag{205}$$

This integral represents the *electric moment* of the supposed distribu-

[27] I should again recall that in the picture we have, not two distinct coexisting modes of vibration of the same elastic string; but the fundamental (and solitary) modes of vibration of two distinct elastic strings.

tion of "electric charge" along the imaginary string, relatively to its centre at the origin. If it should turn out zero, there would be equal amounts of charge to left and to right of the centre; if it should turn out positive or negative, there would be more charge to the right of the centre than to the left, or more to the left than to the right; if it should turn out variable, if for instance the coefficient of the cosine-term should differ from zero, there would be a *surging of the charge* to and fro across the origin.

The functions $f_i(x)$ have been written down in equation (155), near which it was shown that they are alternately even and odd functions of x; $f_0, f_2, f_4 \cdots$ are even, $f_1, f_3, f_5 \cdots$ are odd. Their squares are consequently even, the products of their squares by x are odd, functions of x; and the first two integrals in the expression (205) for M vanish. As for the integral $\int_{-\infty}^{\infty} x f_i f_j dx$, its integrand is an odd function of x if i and j are both even or both odd, and in either case it vanishes; so that if two wave-patterns corresponding both to even-numbered or both to odd-numbered Stationary States coexist, there is no surging of charge to and fro, and the electric moment of the system remains constant. If i is even and j odd, or vice versa, the conclusion is not so immediate. It follows however from the general properties of the Hermite polynomials [28] that the integral $\int_{-\infty}^{\infty} x f_i f_j dx$ always vanishes unless i and j differ by one unit, so that in every case but this the electric moment is continually zero. This leads us to the rule:

If two modes of vibration i and j coexist, the electric moment of the "imaginary string" representing the linear harmonic oscillator varies sinusoidally with the frequency $(\nu_i - \nu_j)$, *if and only if* $i = j \pm 1$; *otherwise the electric moment is and remains zero.*

This may be interpreted as meaning physically that radiation occurs only when two "adjacent" states—two states for which the quantum-number differs by one unit—coexist; that *transitions are possible only between adjacent states.*

This is a Principle of Selection. It is the principle of selection predicted for the linear harmonic oscillator in the earlier versions of atomic theory, and sustained by observations on those features of band spectra which are attributed to simple-harmonic vibrations of molecules.

Thus in the case of the linear oscillator, the idea of interpreting the square of the amplitude of the Ψ-vibrations as density of electric charge is twice successful. When the oscillator is in one of its sta-

[28] Courant-Hilbert, *Methoden der math. Physik*, p. 76.

tionary states, the distribution of "charge" along the imaginary string which represents it is stationary; when the vibrations corresponding to two distinct Stationary States coexist, the distribution of the "charge" fluctuates, with precisely the frequency which experiment teaches us to expect from a transition between the states in question; when and only when the two coexisting stationary states are adjacent in the ordering, when and only when experiment teaches us to expect transitions, the fluctuation assumes the emphatic character of a bodily surging of the charge to and fro across the centre of the string.[29]

One further desirable result of identifying square of amplitude of Ψ with density of electric charge appears when from one dimension we go over to systems of two or three dimensions. As illustration I take the example of an hydrogen atom exposed to an electric field, represented by an imaginary fluid in three dimensions, the stationary wave-patterns of which correspond to the stationary states of the perturbed atom. If two of these stationary wave-patterns coexist, there may be a bodily surging of charge to and fro, with the frequency belonging to the transition between the stationary states which the wave-patterns represent. If in particular two wave-patterns sharing a common value of the quantum-number s (the "equatorial quantum-number," equation 186) coexist, there is a surging of the "charge," and its to-and-fro motion is parallel to the applied electric field; there is no component of the motion normal to the field. With this result agrees well the fact of experience, that the light emitted by virtue of transitions between stationary states differing only in the quantum-numbers k_1 and k_2, and sharing the same value of s, is polarized with its electric vector parallel to the field. Again, if two wave-patterns for which the values of s differ by one unit coexist, the resultant surging of the charge is perpendicular to the electric field; and it is a fact of experience that the light due to transitions between stationary states having values of s one unit apart is polarized with electric vector normal to the field. Finally, if two wave-patterns for which the values of s differ by two or more units coexist, there is no far-sweeping dis-

[29] Schroedinger has shown that if we conceive a great number of Stationary States with high values of i and artfully chosen relative "amplitudes" (i.e. values of A_i and B_i in equation 201) to exist simultaneously, we find the "charge-density" concentrated into a small region, a sort of knot or bundle which oscillates back and forth across the centre of the string with frequency ν_0 and with approximately the same amplitude of vibration as the original particle (the particle with mass m and restoring-force $-4\pi^2 m\nu_0^2 x$ of which the string is our image in wave-mechanics) would have if its energy were the same as that of the Stationary State which was made most prominent in the summation (*Naturwiss.*, **14**, pp. 664–666; 1926). This is a promising result, suggesting as it does that atoms in highly excited states may be groups of particles which, as the system returns to normalcy, spread out into a sort of fluid haze. The idea can be generalized widely, and merits a thorough analysis.

placement of charge; and in the spectra, the lines which such transitions should cause are missing.

Thus in the case of the hydrogen atom exposed to an electric field, and in other two- and three-dimensional systems as well, the identification of the square of the amplitude of the Ψ-vibrations with density of electric charge is thrice successful. In the picture, we see the electric charge stationary when the system is in a stationary state, fluctuating with the proper frequency when two states coexist; we see it surging back and forth *en masse* when the coexisting states are two between which a transition is "permitted," and otherwise not; we see it surging back and forth along the proper direction to explain the polarization of the light which results from the transition. As a device for picturing the radiation-process, Schroedinger's model is certainly unrivalled. In the earlier atom-models, even the frequencies of the emitted rays of light and the frequencies of the intra-atomic vibrations did not agree. Here at last they do, and when a tube full of hydrogen atoms is pouring out the light of the red Balmer line with its frequency of $4.57 \cdot 10^{14}$, it is permissible at last to imagine each of them as a mechanism, within which something is vibrating $4.57 \cdot 10^{14}$ times in a second.

Even the relative intensities of spectrum lines may fall within the scope of wave-mechanics. We have seen that in the case of the linear oscillator, the vanishing of the integral $\int x f_i f_j dx$ for all pairs of Stationary States for which i and j differ by more than one unit entails the non-occurrence of the corresponding transitions, the inability to emit or absorb the corresponding radiation. May it not be that the intensity of the radiation emitted by reason of the transition between any two states of any system, and polarized parallel to any direction x, is governed by the value of the integral $\int x \psi_i \psi_j dx$ involving the *Eigenfunktionen* ψ_i and ψ_j of the states in question? To develop this idea more assumptions must be introduced than I have yet mentioned, since every *Eigenfunktion* which I have thus far written down might be multiplied by any constant factor without ceasing to be an *Eigenfunktion*, and some rule must be laid down to fix these constant factors. To predict the relative intensities of the components into which certain hydrogen-atom lines are split by electric field, Schroedinger made a simple and natural assumption about these factors; and the results turned out to be in good agreement with the data.[30] I cannot enter further into this topic, except to remark that the point of contact between wave-mechanics and the matrix-mechanics of Heisenberg lies here; for the integrals in question figure as matrix-

[30] Schroedinger, *Ann. d. Phys.*, **80**, pp. 464–478 (1926); *Phys. Rev.* (2), **28**, pp. 1049–1070 (1926).

elements in the latter theory, which indeed appears to be an alternative way of thinking to reach the same conclusions as emerge from the speculations of de Broglie and Schroedinger.[31]

Nevertheless the image is still far from perfect; there is certainly something still lacking, something still to be discovered and added. Radiation may flow forth from the atom when two stationary states coexist, but it does not flow forever; one or the other of the wave-patterns must therefore die out, soon after the radiation commences; yet no agency has thus far been provided to effect the extinction of either. It may not be difficult to insert such an agency into the theory, in the form perhaps of an interaction between the Ψ-waves and the outflowing electromagnetic waves. It may be much more troublesome to extricate ourselves from the paradox into which the identification of square-of-amplitude-of-the-Ψ-vibration with density-of-electric-charge has led us. All of the numerical agreements between this theory of the hydrogen atom and the features of the hydrogen spectrum are obtained by putting $-e^2/r$ for the potential-energy-function of the atom-model. This is the potential-energy-function for a point-nucleus and a point-electron. If we dissolve the electron, spread it out like a cloud in space around the centre of the atom, how can we consistently retain the potential-energy-function derived from the picture of a point-charge? How is it defensible to define electric charge in one way in order to lay the cornerstone of the new theory, and then redefine it in a contrasting way in order to raise the superstructure?

Wave-mechanics, striking as are the pictures which it offers of certain of the processes within the atom, still abounds in conceptual difficulties of which the last is a fair instance; and those who share the view of Lessing that it is more desirable to be approaching truth perpetually than ever to attain it may still find satisfaction in physics. Wave-mechanics still is tentative, not definitive; a plan of campaign, rather than a conquest. The outcome cannot now be foreseen. Yet we may reflect that twenty-five years ago it was universally supposed that light possesses only the qualities of a wave-motion; and then experiment was piled upon experiment which showed that in addition it behaves in many situations as though it were a stream of corpuscles. Perhaps we stand at the beginning of an equally imposing series of experiments, which will show that matter with equal inconsistency partakes of the qualities of particles and of the qualities of waves.

[31] Schroedinger, *Ann. d. Phys.*, **79**, pp. 734–756 (1926); C. Eckart, *Phys. Rev.* (2), **28**, pp. 711–726 (1926).

Power Plants for Telephone Offices

By R. L. YOUNG

Synopsis: The present paper gives a brief discussion of some of the more important problems connected with the supplying of power to telephone offices, and developments which are being perfected to bring about economies. Among the subjects discussed are the use of commercial types of charging generators together with appropriate filters, power factor correction, complete power unit assemblies for small installations, and the development of more nearly automatically controlled power installations with the object of reducing supervision.

I. The Power Problems

The purpose of the telephone power plant is to furnish energy of the required character in proper amount and available 100 per cent of the time. An elaborate telephone system, comprising buildings, central office equipment, outside plant lines and substation apparatus, together with a staff of operators, is rendered useless if the supply of power fails. No conversations can be held. No calls can be made and none received. In a way, the power plant might be termed the "heart" of the system, since every line and connection will be "dead" the moment the supply of power is interrupted.

Continuity

In order to meet the vital need of ever-ready power it is necessary in telephone power plants to arrange for some primary power source which is usually a commercial electric service from outside. The services are investigated with care to determine their reliability and, wherever possible, two services connected to different generating stations or systems are brought into the telephone building. In those cases where a single service only can be secured, a local means of charging such as an engine-generator set may be provided as a reserve on this service.

Even with the best commercial power services short interruptions are experienced, so that it is necessary to provide another source which shall be available at all times to operate the central office during temporary failures of the outside service. This is accomplished by the use of a storage battery of sufficient capacity to carry the load of the office during failure of the sources of power supply, the battery being continuously connected to the circuits so no interruption occurs. Common practice and experience have resulted in batteries of certain sizes being provided, these sizes being sufficient to carry the exchange

load for intervals ranging from a few hours to several days, depending upon conditions. The present practices have been successful in maintaining continuous power supply, and central offices generally throughout the country have been ready to serve, even during periods of storm, fire or other calamities.

Type of Power Needed

Power as furnished by the public service companies is not of the sort suitable for operating telephone power plants, but must be converted from a relatively high voltage alternating or direct current to a lower

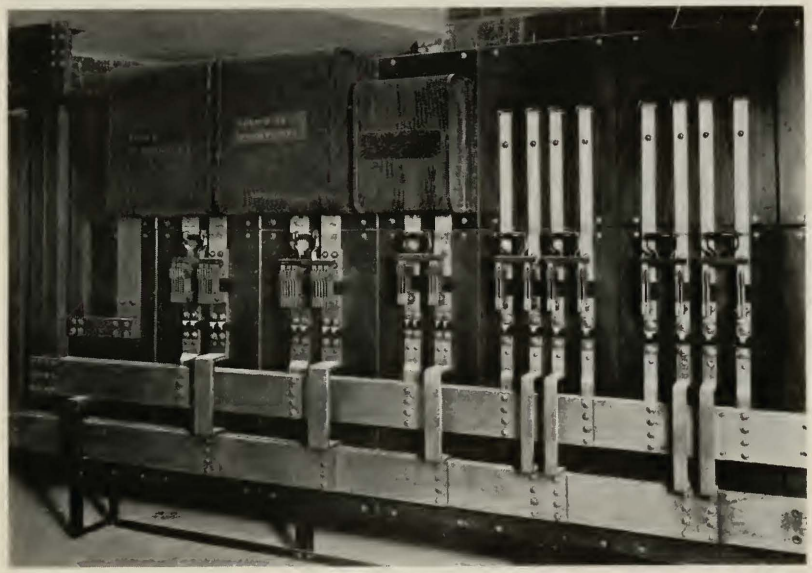


Fig. 1—Incoming direct-current power for large telephone building. About 1,000 h.p. of this is provided to drive motor-generators for reserve central office use, the regular power being alternating current. Both direct- and alternating-current services are duplicated. This panel provides four feeders direct to substation and four to network, capacity 3,480 kw.

voltage direct current for talking, supervisory and signaling purposes and to alternating current of various voltages for signaling. This conversion is commonly made by means of motor-generator sets or some type of rectifier, of either the mercury arc, hot cathode or other types. Since it is impossible to use outside power as furnished, suitable reserve machine equipment must be provided capable of replacing the regular machines before the reserve energy in the central office battery is exhausted.

The low voltage charging generators furnishing the bulk of the power must be electrically quiet so that they will not cause disturbing noises in the telephone circuits. It is, of course, economical to furnish most of the energy required by the telephone equipment directly from the motor-generator sets rather than from the reserve battery, since the conversion efficiency is substantially greater and the battery investment is much less. While various direct-current voltages are required, 24 volts and 48 volts predominate.

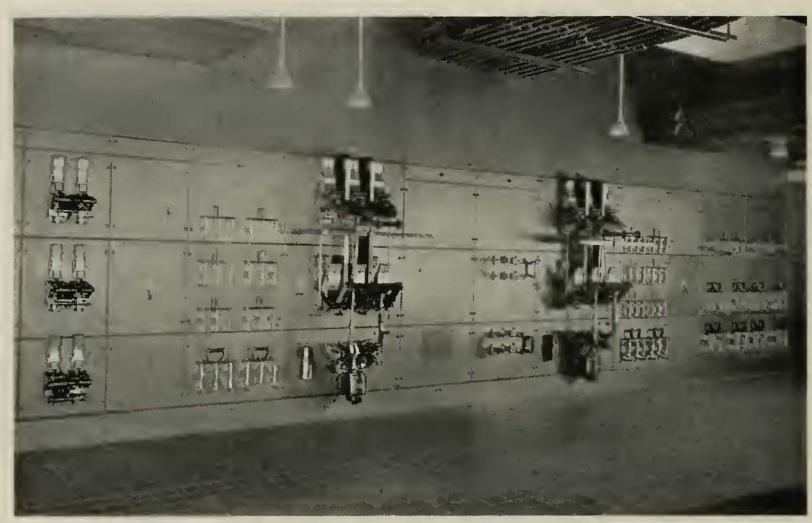


Fig. 2—Building switchboard to distribute incoming power shown in Fig. 1. The 5,000-ampere circuit breakers switch the important load circuits from this panel to a similar reserve panel fronting on a different street.

The signaling machines and batteries, while of relatively small output, are subject to rather exacting performance limitations. Twenty-cycle alternating current of approximately sine-wave form, at nominal voltages of 105, 100, 85, 77, etc., is needed for ringing on different types of circuits. For four-party selective ringing, positive and negative direct currents are superimposed upon the alternating current to secure wave shapes especially suited to the operation of biased ringers. For machine ringing, the 20-cycle current is divided into one or two second ringing periods separated by silent intervals during which direct current is provided for operating the tripping relay and stopping the ringing when the called subscriber answers.

For ringing over composited toll lines a higher frequency, which will not interfere with telegraph operation, is required and 135 cycles is

provided. For other types of toll circuits "voice-frequency" ringing at 1000 cycles must be furnished.

Message registers in manual offices use direct current at 39 volts, coin collect and refunding operations require positive 110-volt and negative 110-volt direct current, while "tones" of approximately 160 and 480 pulsations per second are needed for giving various signals to the operators and the subscribers. A graduated tone, like a siren, is required for the "howler" used to call the attention of a subscriber to a telephone receiver left off the hook. Various flashing signals and combined tones and flashes are also used, such for example as the "busy" signal.

Operation and Maintenance

In addition to being designed for furnishing power of the required characteristics, the machines and apparatus must operate for long periods with a reasonably small amount of operating attention and maintenance. Due to the narrow requirements being placed by the circuits upon the power equipment and the more frequent readjustments required, it is becoming necessary in many cases to furnish automatic voltage regulation. As the cost of labor increases, it will become still more desirable to provide equipment which will largely run and regulate itself.

Sizes of Power Equipment

It has been stated at different times by people connected with the telephone companies and also by outsiders that the amount of power required to carry on telephone conversation is microscopically small, if not negligible. This perhaps is true when considering merely the small amount of alternating current which travels over the line and operates the diaphragms of the receivers. The great sensitivity of this instrument permits operation on very small energy.

There is, however, a large amount of equipment in the central office, including relays, lamps, and other apparatus, which must function in order that this small talking current may be provided and may go from the subscriber who wishes to talk to the subscriber he desires to reach. When this apparatus is multiplied for the thousands or even hundreds of thousands of conversations per day which may be supplied from a power plant serving two, three or perhaps more central offices, the size of equipment needed becomes quite substantial. In these multi-office power plants several of the largest charging generators each driven by an 80 h.p. motor, as well as a number of smaller charging sets may be required, while two batteries of the largest storage battery cells manufactured may be used in parallel to give the necessary battery reserve.

Some of the large telephone buildings house several central offices and, in addition, administrative, engineering, commercial and other departments. A joint incoming power service is often provided for such a building, of which the initial telephone power plant requirements may approximate 500 h.p. with an estimated energy consumption approximating 1,000,000 kw.-hrs. per year. Provision for double this demand in the ultimate may be made.

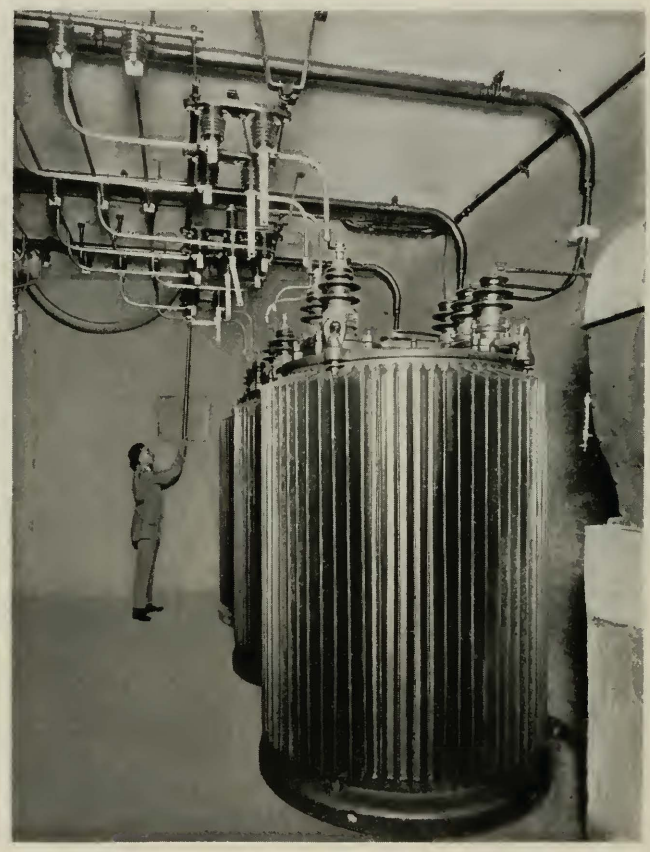


Fig. 3—Typical large transformer installation for breakdown service when two types of alternating current are furnished. Three 333-kv.-a., 11,000/2,200-volt transformers supply frequency changer, these being in addition to the 60-cycle transformers for regular service.

The range of sizes is very great, varying from the above down to the small "magneto" office which operates largely on dry cells and other primary batteries, and may also take 1/8 h.p. to run a magneto ringing machine from the power service. In such offices without

electric power supply, all the equipment must be operated from primary batteries. The "magneto" office, so-called, is one which serves "local-battery" subscribers each of whom has dry cells and a hand ringing generator or magneto. In the larger "common-battery" systems all the power for both talking and signaling is provided from sources at the central office common to all subscribers.

In dial offices it is evident that more power equipment is required, since the processes of connecting through the circuits are performed by machine instead of by operators.

Cost of Power

The cost of power as purchased from the public service companies varies largely, depending upon the location, the amount purchased and to some extent upon the characteristics of the load. In large cities, power is billed at from 2 to 5 cents per kw.-hr. Usually a sliding

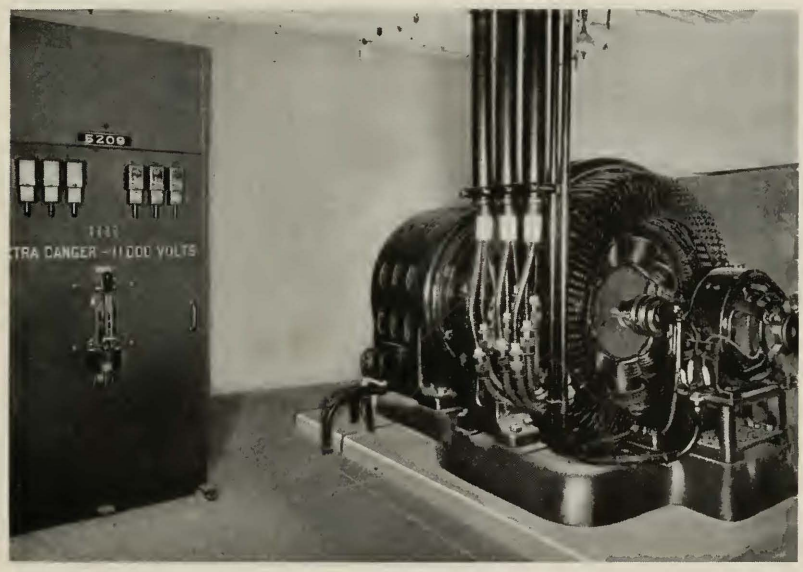


Fig. 4—Frequency changer set for breakdown service, 650-h.p., 25-cycle, 2,200-volt to 600-kv.-a., 60-cycle, 220-volt. Emergency power controlled by 11,000-volt truck switch.

scale is offered and the lower figures apply to purchases of alternating current in large quantities. Cities near sources of soft coal supply or near large water power developments get cheaper rates, in some cases being nearer 1 cent than 2. In small offices, power usually costs between 5 and 10 cents per kw.-hr., running as high as 15 cents in a

small percentage of outlying rural offices. For studies involving the use of power furnished from a telephone power plant, it will, of course, be necessary to consider the cost of the machine and battery equipment, of the floor space and of the operating attendance in addition to the cost of the "raw material" power as purchased. A fair overall figure, including these charges, might approximate 30 cents per kw.-hr. for a typical dial office, or 40 cents for a typical manual office, the higher charge for manual offices, in general, being accounted for by the fact that the quantities purchased and used are less, involving somewhat higher purchase price and overhead. It should, of course, be appreciated that the amounts will vary considerably with local conditions including the type of equipment used and the "load factor," or the distribution of load throughout the day and night. In most telephone power plants this factor is unfavorable for low cost power since most of the traffic is concentrated within a few hours of the twenty-four. The cost of energy varies also during the life of the same office, being higher during the early years and lower when load on the power equipment more nearly approaches capacity.

II. Some Developments to Meet the Power Plant Problems

The objectives toward which development work is directed are improved service, reduced cost, simplification of installation and decreased maintenance. Under these headings one of the most important developments at the present time is the use of commercial type charging generators.

Commercial Type Charging Generators

Charging motor-generator sets furnish most of the energy used in telephone power plants. Up to the present time "telephone generators" have been built to give an electrically smooth direct-current output which will not cause interference with conversations when furnishing current to the telephone circuits. They have also been made mechanically quiet so as not to interfere with nearby testing. They are quite special in construction, including smooth core armature and brass gauze brushes, and are subject to certain limitations which make them larger and considerably more expensive to build than ordinary machines of the same capacity.

Filters consisting of choke coils and high capacity electrolytic condensers have been developed, and with these filters commercial type charging generators can be used to float or charge the central office battery, and this type of generator is now being made available. The purpose of the filter is to make the current from the discharge

leads of the power plant sufficiently quiet for talking battery supply. It is also possible to use a somewhat higher speed machine which is smaller than the present type. The usual slotted mica commutator construction and self-lubricating carbon brushes are employed. While the mechanical noise tends to be greater because of the higher speed and the carbon brushes, it has been found that this is not a factor of importance under present conditions where power plants can usually be located more or less by themselves and well removed from the Wire Chief's testing equipment.

Filters

A new type of choke coil has been developed for the filter used with commercial type generators. It is of the enclosed shell type design having short air gaps, using the materials more economically and having a higher inductance than the coils which have been available

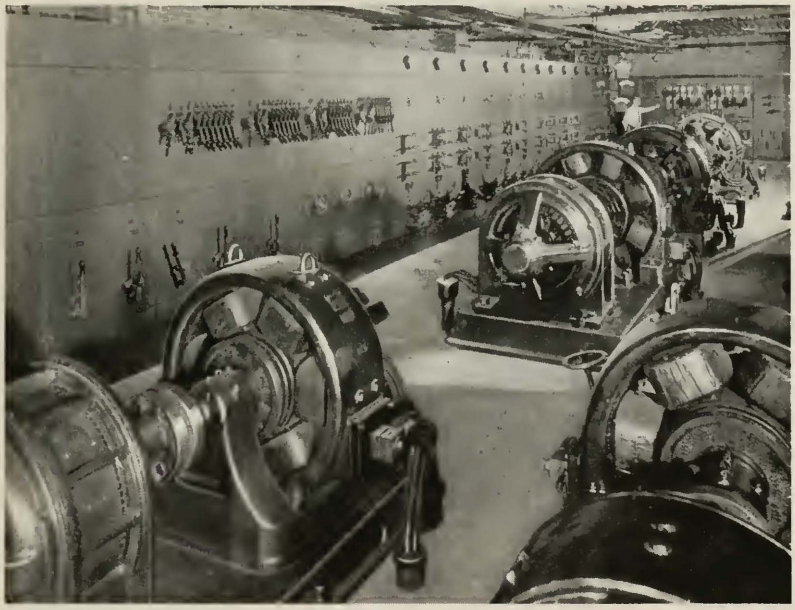


Fig. 5—Part of power room for two large panel units as provided five years ago. Twelve motor-generators for both alternating-current and direct-current service with power switchboard at left and battery control board in background. This was an alternate arrangement to use of large motor-generator to make both kinds of power available.

heretofore. Associated with this coil is a group of electrolytic condensers each of which has upwards of 1,200-mf. capacity on 24 volts, or roughly half this amount on 48 volts.

For one of the larger motor-generator sets the cost of the first set, including a filter, is about the same as that of the former "telephone generator" type outfit. The cost of the additional sets used in the power plant, which under present conditions are not provided with individual filters if a common filter is placed in the talking discharge circuit, is about half that of the present type sets.

Power Factor Correction

In connection with the new generators, synchronous motors are being made available for use where it is desired to improve the power factor of the load. The synchronous motors will be arranged to give



Fig. 6—Generator end of control switchboard, as provided five years ago, rear view. Six 1,000- and 1,500-ampere generators, bus bars terminated for growth when additional units required.

0.8 leading power factor, so all motors in the plant will not need to be of this type. The standard induction motors which are cheaper will be retained, both types thus being available to meet all conditions. For existing installations requiring a moderate amount of correction to avoid the imposition of penalties and where no new motor-generator sets are to be added, static condensers are available to improve the power factor and thereby reduce the excess charges on the power bill.

The addition of these is more economical than new synchronous motor-generators unless the replaced sets can be used to good advantage elsewhere.

Power Switchboards and Bus Bars

Two recent developments are now in use which will reduce the cost of control equipment and will have other advantages. Long power switchboards for large power plants had to be designed in detail and built individually. The use of unit control panels in power boards and battery fuse panels now permits layout of a simple schematic from which a power board can be assembled, using units which may be stocked as demand warrants. Considerable engineering expense is thus avoided since it will be unnecessary to work up the rather elaborate detailed drawings previously required for each major installation.

A further development of this idea is the "semi-remote control system" recently adopted. With this system most of the control equipment for each motor-generator set is located upon unit panels mounted at the set, thus reducing the main power board to small dimensions and giving increased flexibility which is particularly useful in connection with additions. Overhead bus bars and conduit are employed which are not installed till needed. The flexibility also aids in utilizing improvements and changes in the art occurring between the initial equipment installation and the additions made from time to time as the growth of the load requires.

From a production basis it is anticipated that this unit panel design will be easier to manufacture and to stock and that it will also be simpler to install than the earlier arrangements.

Complete Power Unit Assemblies for Small Applications

Where small amounts of power are required, the provision of storage batteries and associated charging equipment has been relatively high in cost of material and of installation. An appreciable cost reduction has been secured by the design of small power plant units complete with batteries mounted in cabinets and assembled with associated charging or floating equipment.

Crosstalk Reduction

In a common battery telephone office all subscribers are furnished with power from a single central office storage battery. There is a tendency toward "crosstalk," that is, mixing of conversations, so that fragments from one conversation might be overheard in another. This tendency is limited by so designing the battery and wiring common to all circuits that it will have very low impedance, particular attention

being given to arrangement of cables and use of large conductors. This imposes certain limitations upon the location of equipment and may involve considerable cost for copper in the larger power plants. However, by means of the electrolytic condensers, previously mentioned, located at battery fuse panels, crosstalk on talking feeders can be reduced to very low values, the limitations on floor plan arrangement can be largely removed and substantial savings in copper can be made.

Battery Reserve

In order to insure continuous telephone service in spite of failure of the primary sources of power, it has been customary, as already mentioned, to provide storage battery capacity sufficient in itself to operate



Fig. 7—Battery room for two large panel units.

the central office equipment for a considerable period. The amount of battery reserve provided depends upon the reliability of the regular outside power service and on the reserve source provided. This battery reserve may range from about three net busy hours for offices in large metropolitan districts to several days in small outlying offices. This reserve in the past has been successful in preventing suspension of telephone service due to failures of the power. With the greatly extended plant and the increasing reliability of the public service supply companies, however, the allowance of battery reserve in some cases can

be safely decreased, permitting advantage to be taken of appreciable savings in the cost of battery equipment.

Simplified Installation

The developments just discussed will, it is believed, simplify the work of installation, although certain of these developments will not necessarily decrease the amount to be done, as some of the work has

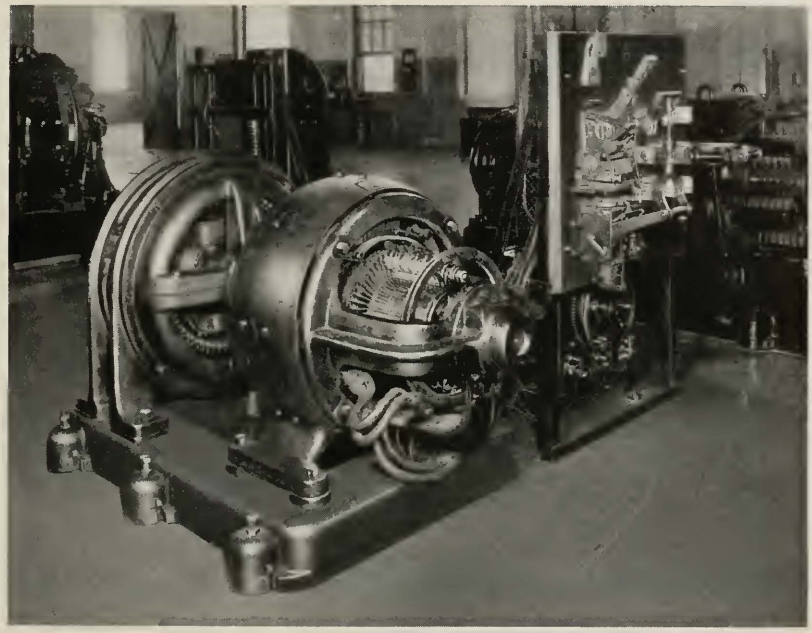


Fig. 8—Commercial type charging generator driven by synchronous motor. 52-kw. output compares with 33 kw. for the larger sized set on the right. 52-kw. "M" type generator in left background belted to gas engine.

been shifted from installer to factory and some in the reverse direction. The unit ringing control panel assemblies, for example, are being furnished wired in the shop with rows of terminal punchings to which the installer connects the wires from the generators. For charging equipment, the panels at the machines will be connected to a common overhead bus bar system, bus bars being shipped in stock lengths and cut by the installer as needed.

For small repeater and similar installations a power plant has been developed which can be placed upon shelves on a rack and connected to the power source and to the distributing bus bars. This compares with the former system of installing a number of separate units and wiring them up upon the job.

Floor plans suitable for the majority of offices are available and reasonably standardized layouts of power equipment to operate certain types of offices are found practicable.

Improved Service

Improved operation of telephone equipment is being made possible by more rigid requirements placed upon the power plant. Automatic voltage regulators for ringing generators have been in use for some time

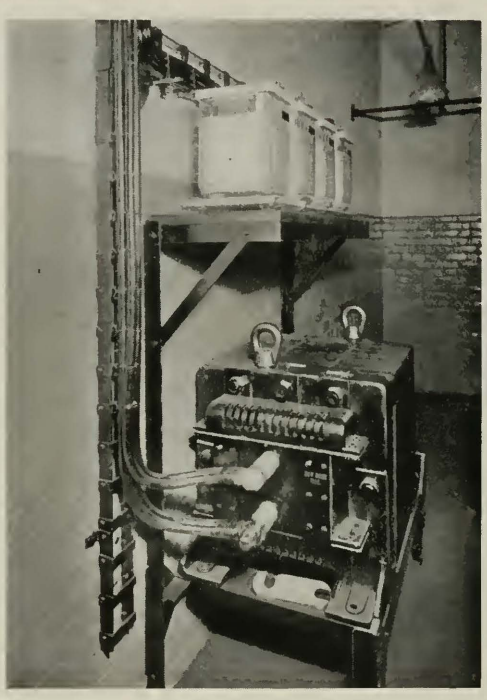


Fig. 9—24-volt discharge lead filter. 800-ampere coil, four 1,200-mf. condensers for use in suppressing generator and other equipment noises.

and the new alternating-current—direct-current system is reducing service troubles. Automatic voltage control equipment for charging generators has been developed and is being introduced. This, together with the full floating system of operating batteries, is capable of holding the main power supply for the central offices at much closer voltage limits than have been found practicable in the past, thus further stabilizing transmission and contributing to even more reliable operation of relays and supervisory equipment.

Reduction of Maintenance

It will be evident that many of the foregoing developments will reduce maintenance of power plant equipment as well as its first cost. The automatic regulating devices substantially reduce or eliminate the attention which would otherwise be required to readjust machines to compensate for load conditions. When this regulation is applied to generators which are floating batteries, it may also result in substantially increasing the life of the batteries, thus deferring replacements. The commercial generators are designed for and equipped with carbon brushes, and will require a minimum of attention.

The introduction of the enclosed type of small battery and the improvements in operating methods of large batteries are decreasing evaporation and spraying, thus reducing additions of water and the repainting of exposed equipment in battery rooms. The new methods also reduce the number of periodic overcharges or eliminate them entirely.

Combining Objectives in Signaling Machine Development

In designing new equipment it is, of course, desirable to accomplish as many objectives as possible. In this connection, mention might be made of a combination machine which may properly lay claim to attaining four important objectives, namely: improved service, reduced cost, simplified installation and reduced maintenance.

Several years ago it was the practice to secure ringing current for subscribers' bells and also for various tones and signals from a small motor-generator set, subject to generator voltage variations amounting to 35 volts as the load on the machine changed and the supply line voltage varied within stated limits. Each large central office unit required these motor-generators, one driven by a line motor and a reserve set driven by a battery motor. Direct current at + 110 and − 110 volts for controlling coin box telephones was furnished by two sets of dry cells or storage cells. In either case a third or spare battery was provided.

To replace the ringing sets and coin control batteries a combined ringing and coin-control motor-generator set has been developed and is being used except in the smallest offices, eliminating the cost and the maintenance of the separate batteries, giving closer voltage and frequency regulation for ringing, and automatically continuing service in spite of outside power failures. A description of the features of this equipment showing what it will do may be of interest as this represents a typical development.

Associated with the generator is a transformer, the primary winding

acting as a balance coil for a three-wire direct-current system and the secondary winding having taps to provide one or more of the four alternating-current voltages used with 20-cycle ringing. The generator can thus serve a toll installation at 105 volts, a dial or manual office

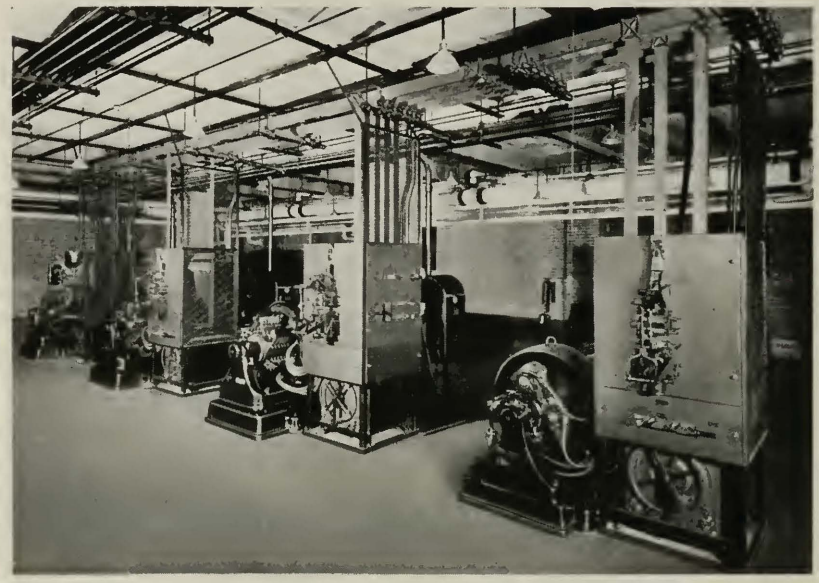


Fig. 10—Charging machines with remote controlled equipment at generator. Overhead bus bar system.

using 100 volts with the alternating-current—direct-current system and two offices, having "superimposed ringing" for party lines, one using 85 volts and the other 77 volts, the voltage used depending upon the type of subscriber sets installed in the district. Positive and negative superimposed currents are obtained by small storage batteries connected in series with the 77- or 85-volt tap of the transformer. All four types of office ringing can be secured from one machine simultaneously, though more than two is unusual. The voltage is controlled automatically within close limits, regardless of load or of normal variations in the voltage and frequency of the supply power.

In addition to ringing, the generator supplies approximately +110 volts direct-current for collecting coins and −110 volts direct-current for refunding coins, the two voltages in combination also exciting the generator field at 220 volts.

Brushes bearing on sectional and solid rings mounted on the generator shaft interrupt battery current and provide a high tone of 480

and a low tone of 160 pulsations per second which are used for various signaling purposes.

Through a 120/1 worm gear reduction an auxiliary shaft is run at 10 r.p.m. Attached to one end of this shaft is a "low-speed inter-

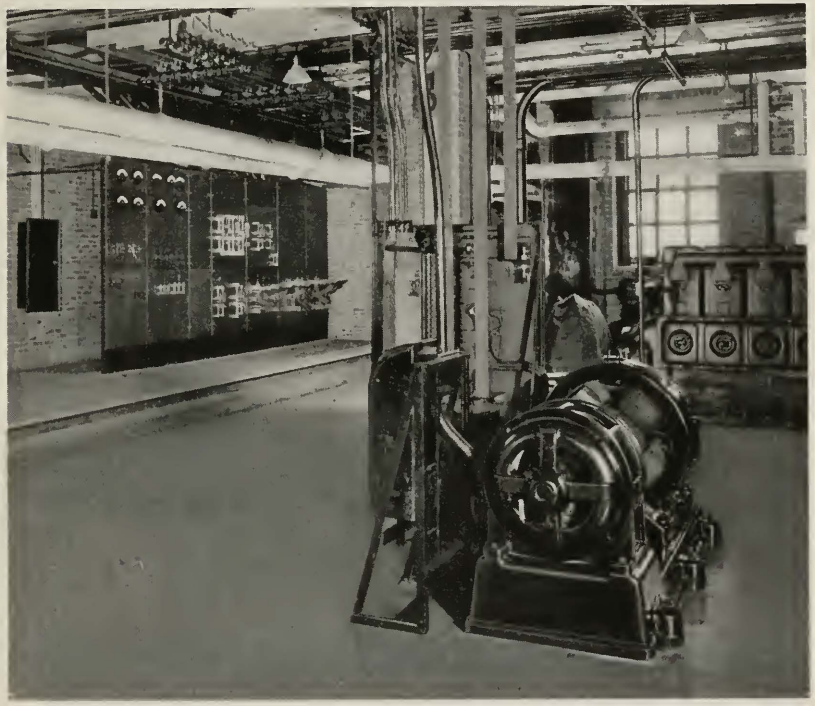


Fig. 11—Master control board for all charging sets and main storage batteries. Charging motor-generator in foreground, emergency engine-alternator set in background. Emergency lighting cabinet to left of control board.

rupter" which provides flashing or tone signals for "busy" and other uses and may provide half a dozen or more different signals. To the other end of this shaft is usually attached a ringing interrupter which divides the constant ringing current from the generator into machine ringing intervals such as 2 seconds ring, 4 seconds silent, or 1 second ring, 1 second silent, 1 second ring, and 3 seconds silent. This interrupter also controls battery current for tripping during the silent interval, and a "pickup" circuit the purpose of which is to prevent ringing the wrong party on party lines.

The generator and all the interrupters are regularly driven by an alternating-current line motor operating upon the outside power

supply. In addition to this, however, the set includes a direct-current motor designed to operate on current from the central office battery but normally not connected to the battery. Automatic relays and magnetic switches close the connection to the battery when the regular power fails so the set continues to operate without interruption

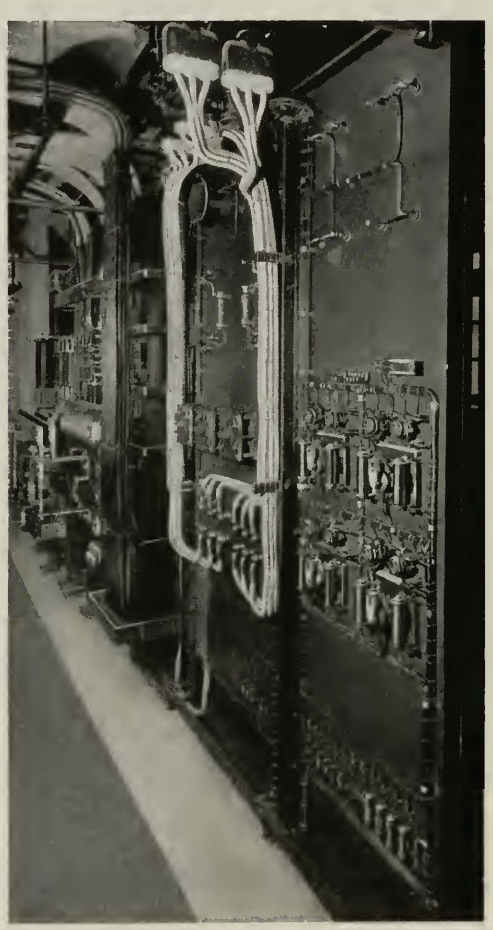


Fig. 12—Rear of control panel for charging sets and batteries.

or change in output. This feature avoids the delay of a few minutes which would otherwise result after a power failure, while the attendant started up a reserve motor-generator and transferred the load circuits, which may run from 17 circuits to twice that number for ringing, coin control, tones and signals. It also avoids accumulation of machine

ringing calls which occur when a ringing generator stops during a busy period and which may overload the motor sufficiently to blow the protective fuses and prevent restarting.

The battery motor is equipped with an automatic speed regulator which keeps the generator frequency within two per cent of rated speed throughout the range of the battery motor supply voltage.

It is obvious that a combination which will do so many things at once costs more than a simpler type of machine. The fact, however, that it will operate several central office units and will replace coin control batteries makes it cheaper than the equipment formerly required to do the work. With fewer machines and no batteries, except those for superimposing, installation is simplified. The closer voltage and speed regulation reduce relay and ringing troubles and, in conjunction with the continuity of operation, improve service from the subscribers' viewpoint, as well as reduce the amount of maintenance required of the attendants.

III. The Future Telephone Power Plant

It may be of interest to consider the direction toward which developments in prospect are leading, that we may learn what the future telephone power plant may be like. It seems probable that further progress will be made in the application of unit panels and unit

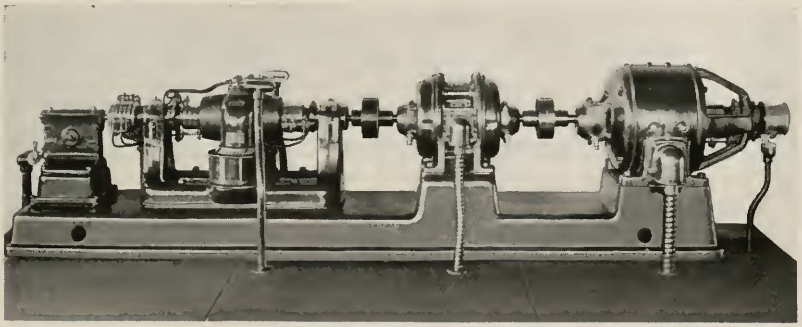


Fig. 13—One type of signaling machine—20-cycle ringing and direct-current 110-volt coin control machine, with reduction gearing for low-speed signals. Driven by a.-c. line motor with reserve battery motor automatically energized upon power failure. Speed controller on battery motor.

assembles or combinations of machines and control equipment. With the better characteristics, much of this equipment, including storage batteries, can be mounted with the circuit apparatus on standard racks, making self-contained units.

Because of the advantages to be obtained in circuit operation from closer voltage regulation and because of the higher cost of manual attendance, automatic regulators for machines and batteries will be used more extensively. Further attention will be given to the automatic operation of power plants, successfully accomplished for private branch exchanges and small offices.

Entirely automatic power plants for large offices could, it is believed, be developed without great difficulty somewhat along the lines of the automatic substations in use by some of the power and traction companies. These should, in general, require attention only periodically for cleaning, replacing worn parts, adjusting, etc., except during a failure of equipment or some other abnormal condition which would be indicated by an alarm. Since full automatic control would probably cost more than that requiring a limited amount of supervision, a study is required to determine how nearly automatic the equipment should be made for offices where an attendant will be required in any case for some of the equipment.

As for machines generally, the tendency will be towards greater use of more nearly commercial designs, construction and finish, eliminating as many as possible of the special features formerly necessary but not now required with changed conditions and the supplementary apparatus which recent developments have made available.

A more extended use of filters in power plant circuits may be expected.

In the direction of power supply, efforts have for some time been applied with some measure of success toward increasing the reliability of the service from outside, which work usually consists of cooperation with the electric supply companies in investigating conditions under which independent duplicate power services can be secured. The securing of reliable duplicate services permits elimination of a local emergency generating plant such as the engine-generator sets. As these efforts become more successful and the public service systems increase in extent and in reliability with the increase in interconnection, it should be found possible to reduce the amount of storage battery reserve in telephone power plants. Experimental introduction of low-voltage alternating-current networks similar to the direct-current networks used in the central parts of some large cities is being watched with interest and some installations are in progress. Although this might be classed as one electrical system, the safeguards against failure and the duplication of equipment is often such as to warrant entire dependence upon this power without a separate emergency source in the building.

With regard to types of batteries, the enclosed type in glass jars will be increasingly introduced wherever suitable because of the lower cost of installation and the subsequent reduced maintenance. The further extension of continuous floating systems of operation also makes it practicable in some cases to use batteries of the pasted plate construction in hard rubber jars, which are cheaper, particularly in first cost. On the general subject of battery operation, the use of the "continuous floating system" is being encouraged where practicable since this usually gives more efficient operation and always results in longer life for the storage batteries and in smaller sizes for equivalent reserve. As an alternate plan a "constant voltage charge system" is in process of adoption for general use where, for any reason, continuous floating is impossible or uneconomical.

The size and cost of power plants is largely controlled by the circuit and apparatus requirements, and improvements in these, such as reductions in current drains for dial equipment and for repeater tubes, are immediately reflected in the telephone power plant which will decrease in size and cost in almost direct proportion.

Quality Control

By W. A. SHEWHART

Introduction

A MANUFACTURER is interested in producing a controlled product—one in which the deviations about the average level of quality are no larger than can be accounted for as a result of chance. The present paper gives simple detailed methods for determining from inspection data whether or not a product is being controlled in the sense of indicating the presence of assignable causes of variation. Naturally the inspection data constitutes a sample of the effects of the manufacturing causes and hence the interpretation of these data in terms of what may be expected in the future is a statistical problem.

A controlled product is defined as one for which the frequency of deviations from the expected quality can be estimated by probability theory. To make such estimates, however, it is necessary to characterize or specify the distribution of quality which the manufacturer wishes to maintain. These specifications of the desired quality must be arrived at by methods customarily used in setting engineering standards, but when once they have been established the statistical methods amplified in this paper make possible the most economical control of this quality.

The limits within which quality may be controlled with a given amount of inspection depend upon the standards adopted for the quality to be maintained.

This paper interprets quality specifications in terms of five different types of constant systems of manufacturing causes. The five types chosen are sufficient to cover the entire range and it is believed that only five types are necessary because sampling theory indicates that little practical advantage would be derived by endeavoring to subdivide one or more of these. It is shown that quality control can be maintained with the fewest number of measurements and within the closest limits through the adoption of Type V.

Specification of Control

One of the principal objects of inspection is the detection of lack of control of manufactured product, that is, the detection of the presence of assignable causes of variation in the quality. A recent paper in this *Journal* [1] describes a quality control chart designed to attain this ob-

[1] Shewhart, W. A., "Quality Control Charts," October 1926.

ject and some of the results obtained through the application of the chart have also been presented.[2] In general the detection of the existence of assignable causes of variation leads to their elimination at a minimum of cost.

As a basis for this chart we start with the conception of a constant system of causes as being one such that the probability of a unit of product having the quality X within the range X to $X + dX$ is independent of time. For convenience in the present discussion we may represent this probability dP as a function f of the quality X and m parameters. Thus

$$dP = f(X, \lambda_1, \lambda_2, \cdots \lambda_m)dX. \tag{1}$$

The present paper presents different ways of specifying the constant system of causes and of detecting lack of control upon the basis of the different specifications principally by setting sampling limits on the parameters. In this way it is shown that the best control can be secured when all of the parameters together with the function f in Eq. 1 are specified. We shall assume, in what follows, one set of specifications after another for the constant system of causes and then show for each set how sampling limits may be established. Nomograms are presented to make the determination of the limits possible without the use of even a slide rule. We shall start with the simplest specification, usually referred to as Type I, which has found extensive use.

Type I often gives a satisfactory basis of control although it makes use, as we shall see, of only a fraction of the information given by the data used in connection with Specification Type V, which is the ideal set wherever the manufacturer is warranted economically in trying to secure the highest degree of control. The choice of specification to be adopted in a given case depends entirely upon the economic advantage attainable through the detection and elimination of assignable causes of variation. In particular the use of Type V specification in the initial stages of the development of the manufacturing process is almost always warranted, because it materially assists in arriving at a controlled process with a minimum of labor.

Specification Type I: The probability of the production of a defective piece of apparatus shall be p'.[3]

To set limits in this case is very simple indeed, particularly if we choose the probability P associated with the limits to exceed .9. It

[2] Jones, R. L., "Quality of Telephone Materials," *Bell Telephone Quarterly*, Vol. 6, pp. 32–46, January 1927.

[3] The primed notation is used throughout to denote parameters of the universe as contrasted with the estimates of these determined from the sample.

has been found satisfactory in many cases to take $P \doteq .99$ and so, upon this basis, we shall present the method of setting limits upon the expected fraction defective in a sample of size n. It is well known that the probability of an observed value of p lying within the limits $p' \pm 3\sigma_{p'}$ is approximately equal to .997 provided the fraction defective p' is approximately equal to the fraction non-defective q', and n is large. It can be shown, however, that irrespective of the magnitudes of p' and n the value of P so determined lies between .95 and 1.00 and for most cases met in practice P does not differ from .997 by as much as 1 per cent.

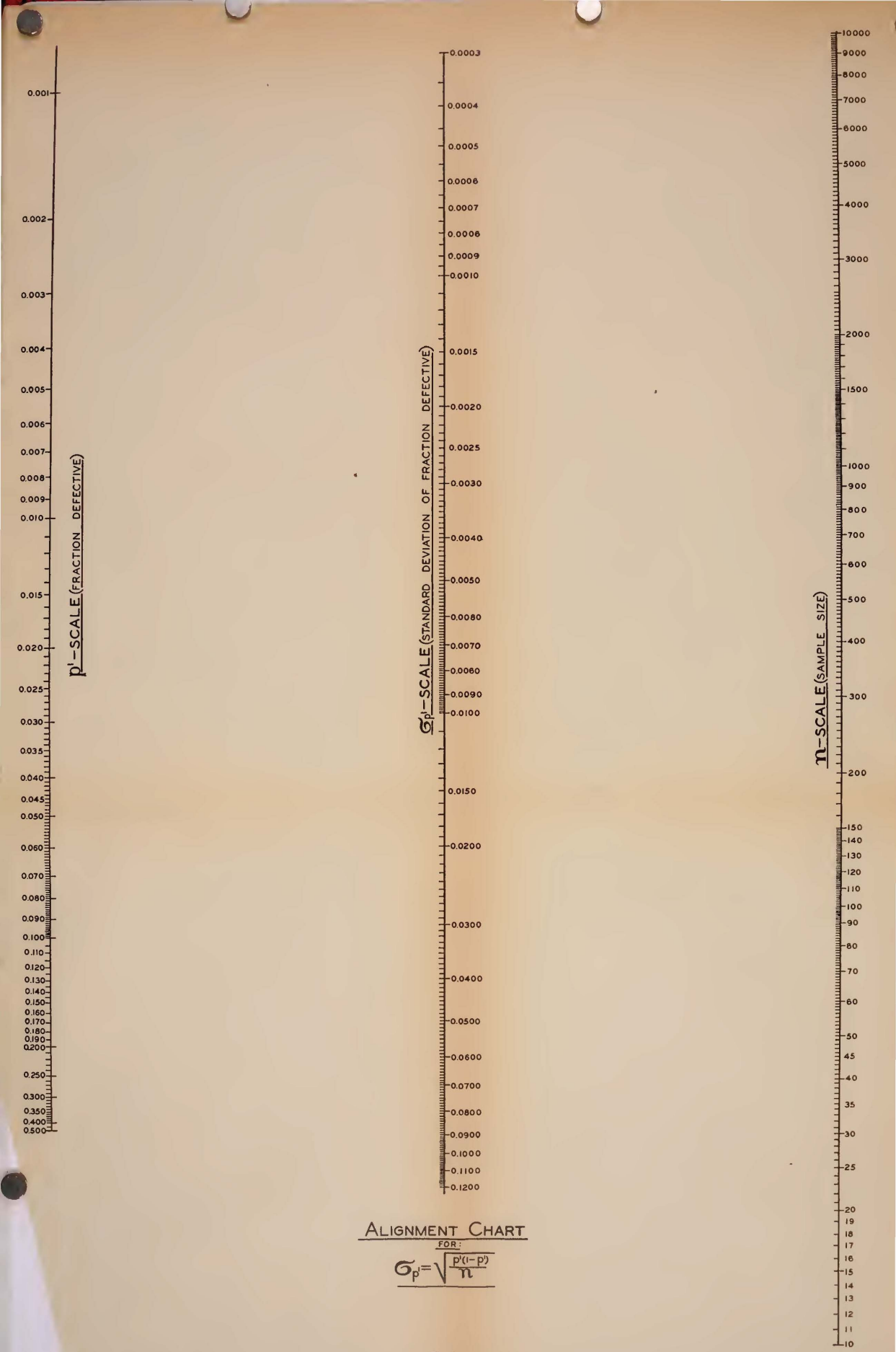
It is obvious, therefore, that, if we construct an alignment chart on which we may read directly the standard deviation $\sigma_{p'}$ when p' and n are given, then the average p' plus or minus three times the standard deviation $\sigma_{p'}$ gives the corresponding values of the limits.

Let us consider a practical problem, see how the question of whether or not a product is controlled really arises and see how control limits can be found from the alignment chart of Fig. 1 to answer this question.

Table 1 represents the observed fraction found defective over a period of 12 months for two kinds of product designated here as Type A and Type B. The table gives for each month the sample size n, the number defective m and the fraction defective $p = m/n$. The average fractions defective for the 12-month period are $\bar{p}_A = .0109$ and $\bar{p}_B = .0095$. Subject to later consideration we shall assume $p'_A = \bar{p}_A$ and $p'_B = \bar{p}_B$.

TABLE 1

Apparatus Type A				Apparatus Type B			
Month	n No. Insp.	m No. Def.	$p = \frac{m}{n}$ Fraction Def.	Month	n No. Insp.	m No. Def.	$p = \frac{m}{n}$ Fraction Def.
Jan.	527	4	.0076	Jan.	169	1	.0059
Feb.	610	5	.0082	Feb.	99	3	.0303
Mar.	428	5	.0117	Mar.	208	1	.0048
Apr.	400	2	.0050	Apr.	196	1	.0051
May	498	15	.0301	May	132	1	.0076
June	500	3	.0060	June	89	1	.0112
July	395	3	.0076	July	167	1	.0060
Aug.	393	2	.0051	Aug.	200	1	.0050
Sept.	625	3	.0048	Sept.	171	2	.0117
Oct.	465	13	.0280	Oct.	122	1	.0082
Nov.	446	5	.0112	Nov.	107	3	.0280
Dec.	510	3	.0059	Dec.	132	1	.0076
Average	483.08	5.25	.0109		149.33	1.42	.0095



ALIGNMENT CHART

FOR:

$$\sigma_{p'} = \sqrt{\frac{p'(1-p')}{n}}$$

Fig. 1

Is there any indication that the observed fluctuations in the fraction defective p could have been produced by other than chance causes? In other words, were apparatus Type A and apparatus Type B con-

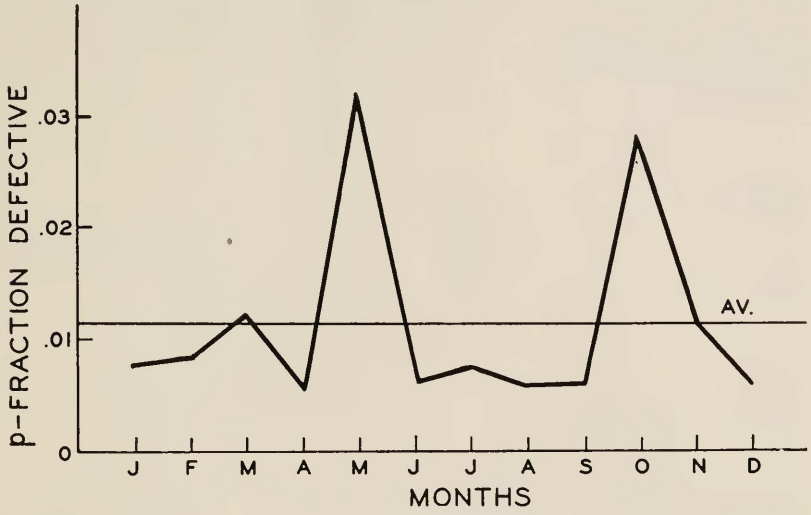


Fig. 2*a*. Apparatus Type A

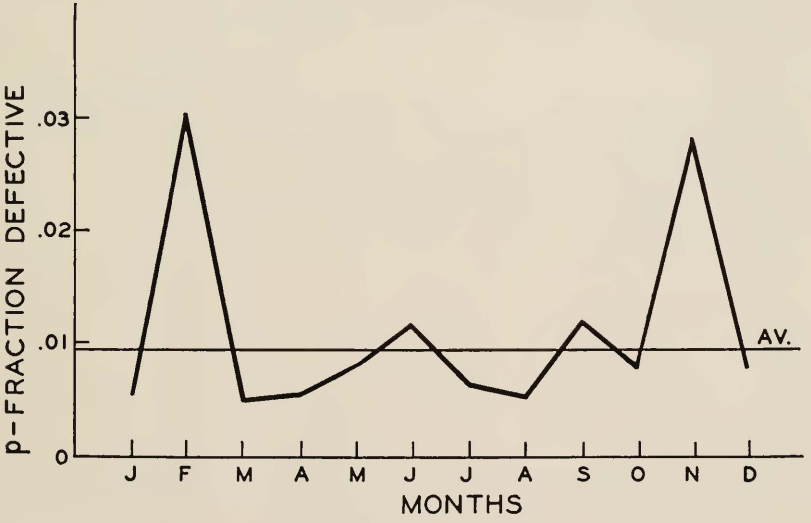


Fig 2*b*. Apparatus Type B.

trolled over the given period? Furthermore, is there any indication that the product could have been improved during this period without changing the process of its manufacture?

Is there any indication that the observed fluctuations in the fraction defective p could have been produced by other than chance causes? In other words, were apparatus Type A and apparatus Type B con-

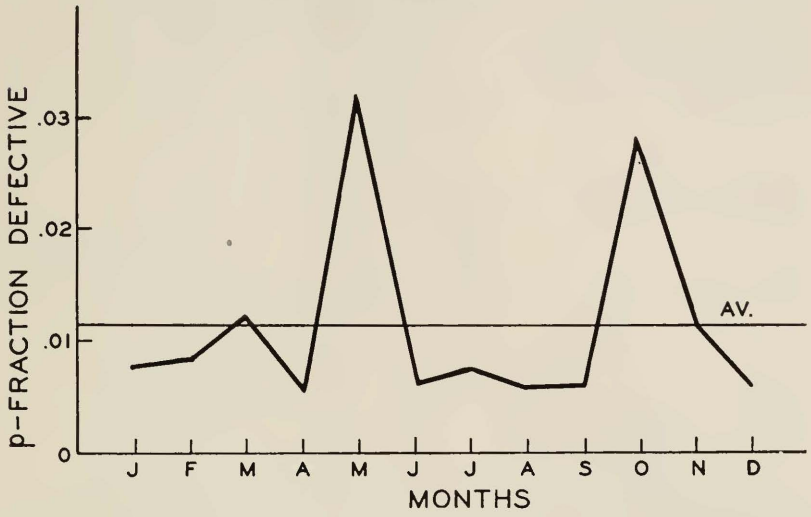


Fig. 2*a*. Apparatus Type A

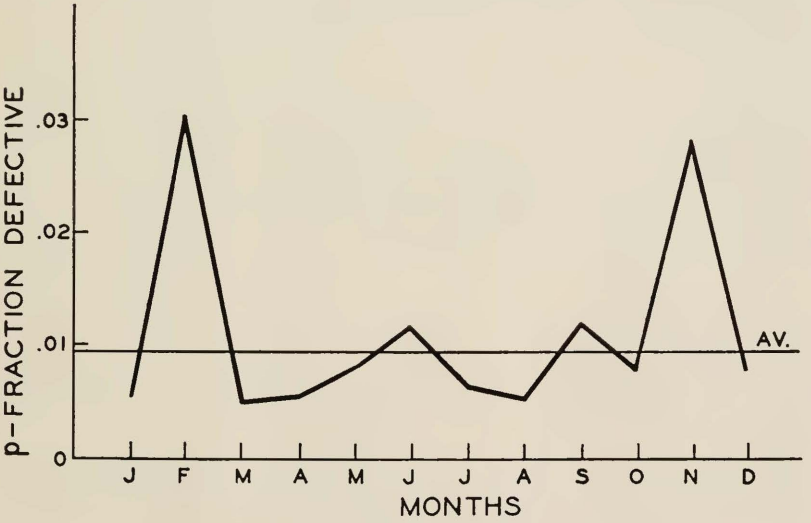


Fig 2*b*. Apparatus Type B.

trolled over the given period? Furthermore, is there any indication that the product could have been improved during this period without changing the process of its manufacture?

To better visualize the fluctuations in p, the data of Table 1 are shown graphically in Fig. 2*a* and Fig. 2*b*. It may appear that during the months of May and October there existed some assignable cause of variation in the production process of Type A apparatus. The same may appear to be true for Type B apparatus during the months of February and November.

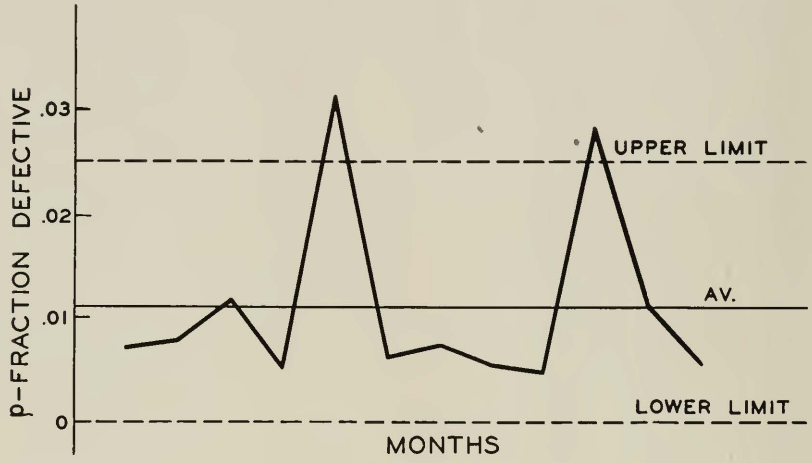


Fig. 3*a*. Apparatus Type A

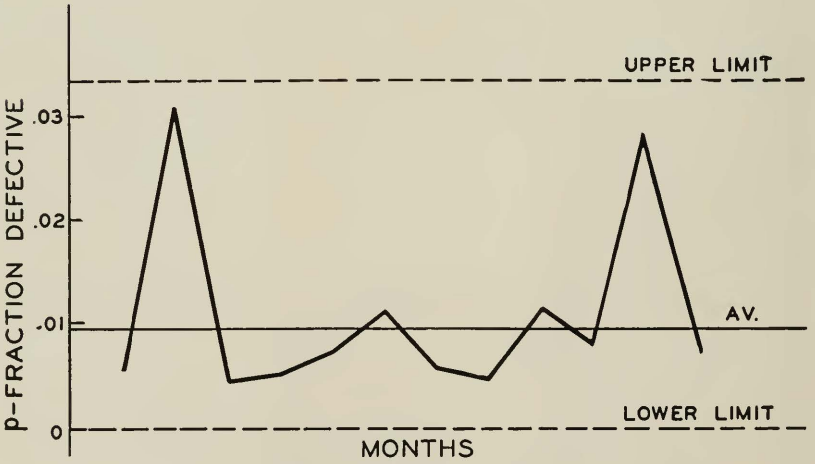


Fig. 3*b*. Apparatus Type B

We shall see upon investigation that there is evidence of lack of control of apparatus Type A but not any evidence of lack of control of apparatus Type B.

Taking n equal to the average sample size (483 for Type A), we connect by a straight line the point 483 on the n scale of Fig. 1 with the point .0109 on the p' scale. We read the intersection of this straight line with the $\sigma_{p'}$ scale as .0047. Hence the upper limit for p' is $.0109 + 3\sigma_{p'} = .0250$, a value which is exceeded during the months of May and October; and the lower limit is $.0109 - 3\sigma_{p'} = -.0032$. Of course negative values of p have no significance; hence we take the lower limit as zero. Following the same procedure for Type B apparatus, we get limits 0 and .0332.

We see that twice during the year Type A apparatus appears to have been out of control whereas at no time during the year can we say this of Type B.[4]

Now, we shall take up successively the method of finding limits corresponding to specifications involving:

a. Only one parameter (Type II).
b. Only two parameters (Type III).
c. Two parameters and a restriction on the function f over a certain range (Type IV).
d. Four parameters and a specific function f (Type V).

We shall find that the limits become progressively smaller in the above order. In fact for Specification Type II no limits can be set and for Specifications Type III and IV the limits are so large as to be in most instances impractical.

Specification Type II: The expected or average quality shall be $\overline{X}'$.

There is an indefinitely large number of constant systems of causes which would meet this requirement. Associated with each constant system of causes there are specific sampling limits. Sufficient information, however, is not called for in the Specification Type II to fix sampling limits on the quality of a single unit or on the expected or average quality.

In other words, Specification Type II is useless insofar as it does not provide for the detection of lack of control in the sense now under discussion.

Specification Type III: The expected or average quality shall be $\overline{X}'$ *and the standard deviation shall be* σ'.

Again there is an indefinitely larger number of different cause systems which would satisfy this requirement. However, it is re-

[4] Strictly speaking statistical theory only shows that two of the observed deviations in p_A are highly improbable upon the assumption that the product had been controlled about p'_A. It should be noted, of course, that the sample size is not the same from month to month and hence that the limits for a given month should really have been based upon the sample size for that month. However, in the present instance, this method of procedure leads to the same conclusion as given above and hence was not introduced because of necessary complications.

markable, even though this be true, that the work of Tchebycheff [5] makes it possible for us to give a lower bound to the probability that a unit of product will be produced with a quality X lying within the range $\overline{X}' \pm L_1$ and also therefore to the probability that an observed average quality of a sample of n units will lie within any given range $\overline{X}' \pm L_n$.

Taking $L_1 = c\sigma'$ $(c > 1)$, the probability $P_{c\sigma'}$ that the constant system of causes Type III will produce a unit of product having a quality X within the range $\overline{X}' \pm L_1$ is given by the expression

$$P_{c\sigma'} \geq 1 - \frac{1}{c^2}. \tag{2}$$

Expression 2 also defines the probability that the average quality of n units of product coming from the constant system of causes Type III will lie within the range $\overline{X}' \pm L_n$ where

$$L_n = \frac{c\sigma'}{\sqrt{n}}.$$

Let us illustrate the method of finding the limits under these specifications. Assume that the specified average resistance $\overline{X}'$ of a relay is 150 ohms and the standard deviation σ' is 5 ohms. What is the range within which we may expect 90 per cent of the product (i.e. $P_{c\sigma'} = .90$) to lie, assuming no assignable causes of variation in product? What is the similar range for the average of 1000 relay windings?

Turning to the nomogram of Fig. 4, we connect by a straight line the point $P_{c\sigma'} = .90$ and the point A near the center of the chart. The point on the c scale fixed by the intersection of the straight line so determined with the c scale is 3.15. The required values of L_1 and L_{1000} are therefore $L_1 = 3.15 \times 5 = 15.75$ ohms and $L_{1000} = \dfrac{3.15 \times 5}{\sqrt{1000}}$ $= .50$ ohm. Hence the limits are 150 ± 15.75 ohms and $150 \pm .50$ ohms.

To avoid the slide rule computations in obtaining $c\sigma'$ and $c\sigma'/\sqrt{n}$ we can use the nomogram of Fig. 5. We enter this nomogram by the value $c = 3.15$ and find a point on the $c/\sqrt{n}$ scale which lies on a straight line with the point $c = 3.15$ on the c scale and $n = 1$ on the n scale. Connecting the point thus fixed on the $c/\sqrt{n}$ scale with the point $\sigma' = 5$, we read on the L scale 15.75 ohms. Carrying through the same procedure, but starting with $n = 1000$ instead of $n = 1$, we read on the L scale .50. These values give the limits found above.

[5] Tchebycheff, *Liouville Journal*, 1867. "Des valeurs moyennes," *Journal de Mathematiques* (2), Vol. 12, pp. 177–84.

0.9997844
0.9996818
0.9995348
0.9993262
0.9990332
0.9986258
0.9980648
0.9973002
0.9962684
0.9948898
0.9930660
0.9906776
0.9875806
0.9836050
0.9785518
3.5
3.0
2.9
2.8
2.7
2.6
2.5
2.4
2.3
0.600
0.650
0.700
0.710
0.720
0.730
0.740
0.750
0.760
0.770
0.780
0.790
0.800
0.810
0.820
0.830
FOR TYPES III AND IV
PROBABILITY P

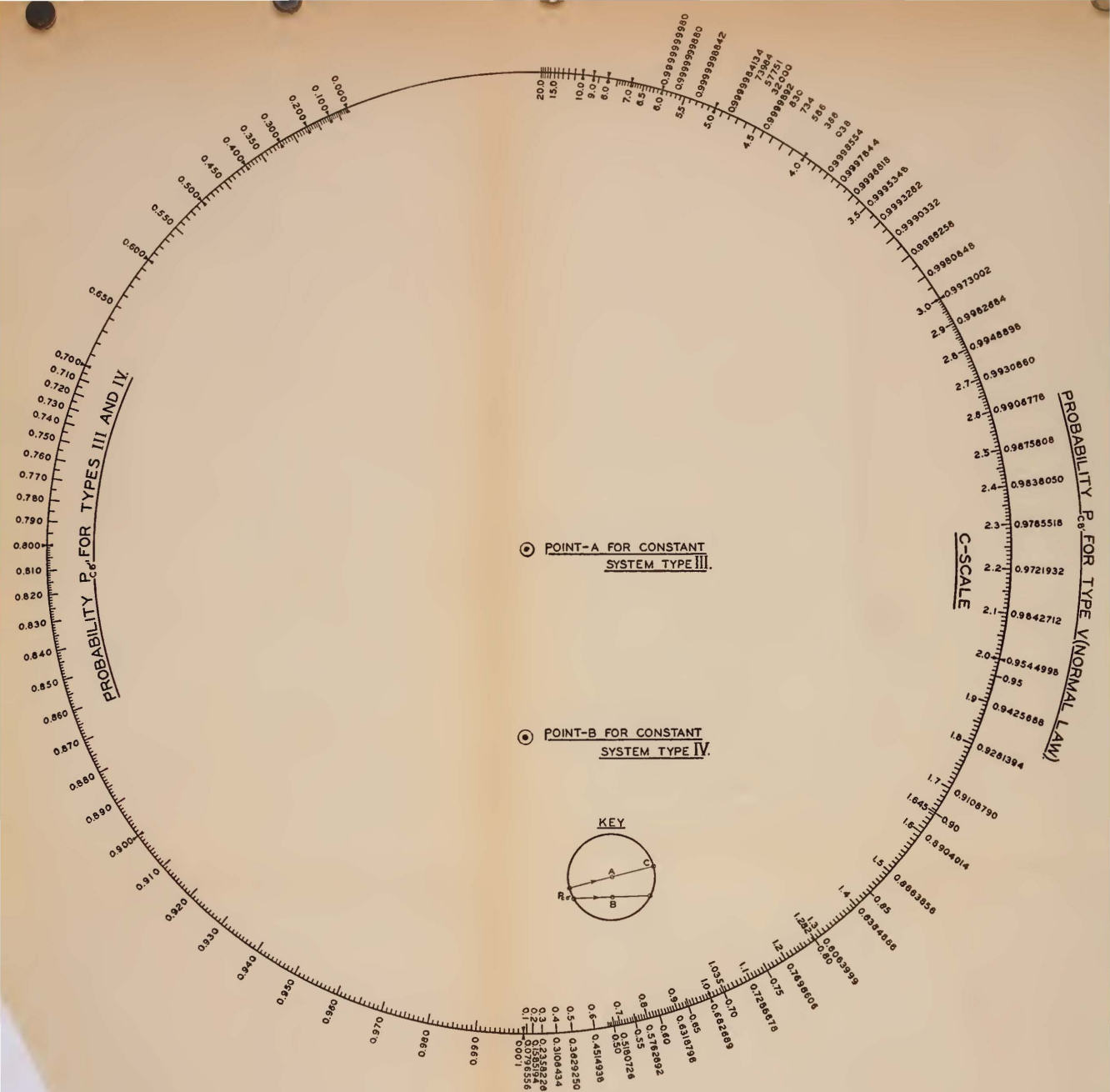
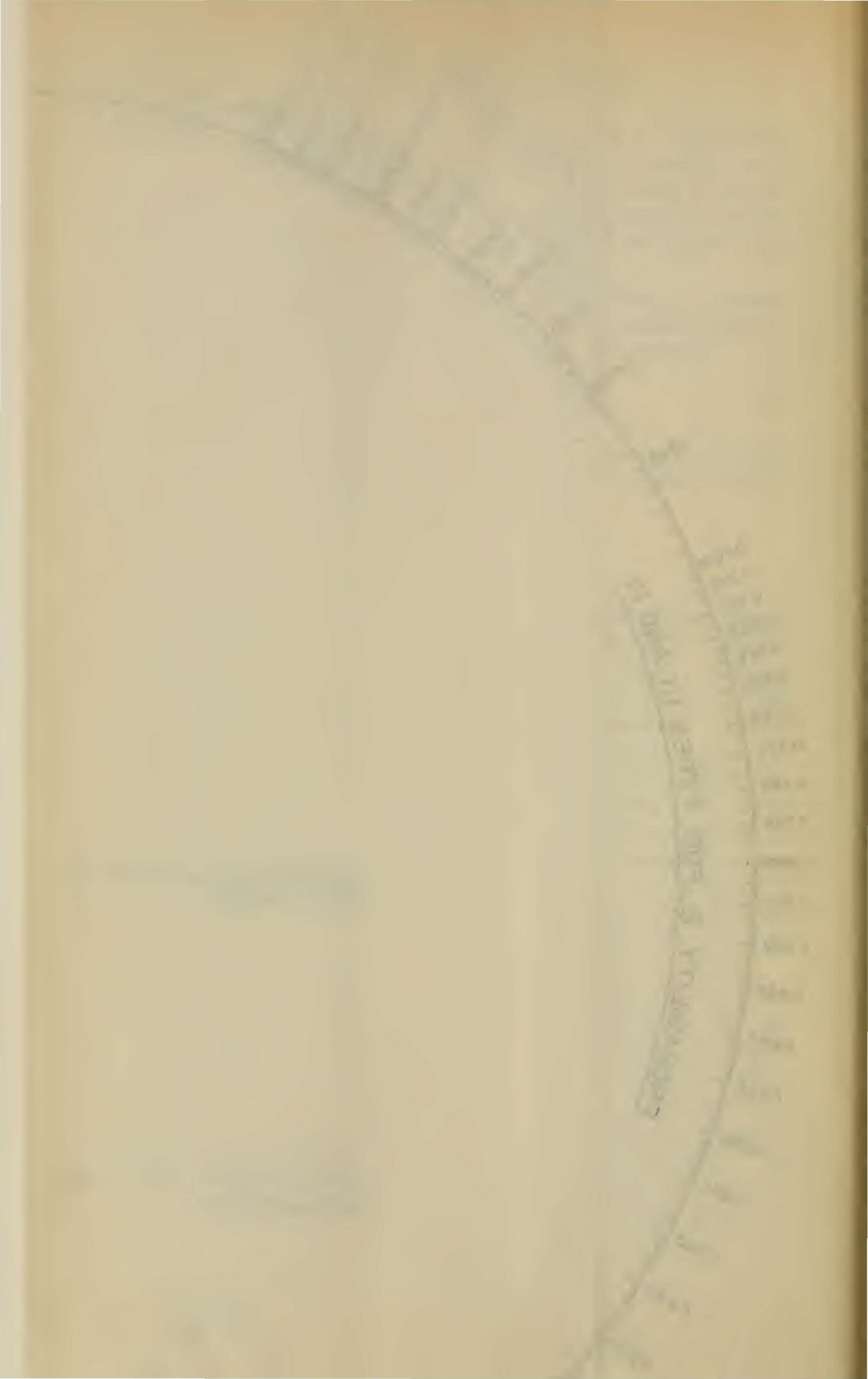

PROBABILITY P_{c8} FOR TYPES III AND IV
PROBABILITY P_{c8} FOR TYPE V (NORMAL LAW)
C-SCALE
POINT-A FOR CONSTANT SYSTEM TYPE III.
POINT-B FOR CONSTANT SYSTEM TYPE IV.
KEY
A
B
C
R_c

Fig. 4

n-SCALE.
G' SCALE.
1.20
1.10
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.400
PER
2000
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
4.5
4.00
3.90
3.80
3.70
3.60
3.50
3.40
3.30
3.20
3.10
3.00
2.90
2.80
2.70
2.60
2.50
2.4
2



n-SCALE.
500
600
PER
2.4
2
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
G SCALE.
2.50
2.60
2.70
2.80
2.90
3.00
3.10
3.20
3.30
3.40
3.50
3.60
3.70
3.80
3.90
4.00
4.5
PER
300
0.400
1.20
1.10
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
√c/ρ

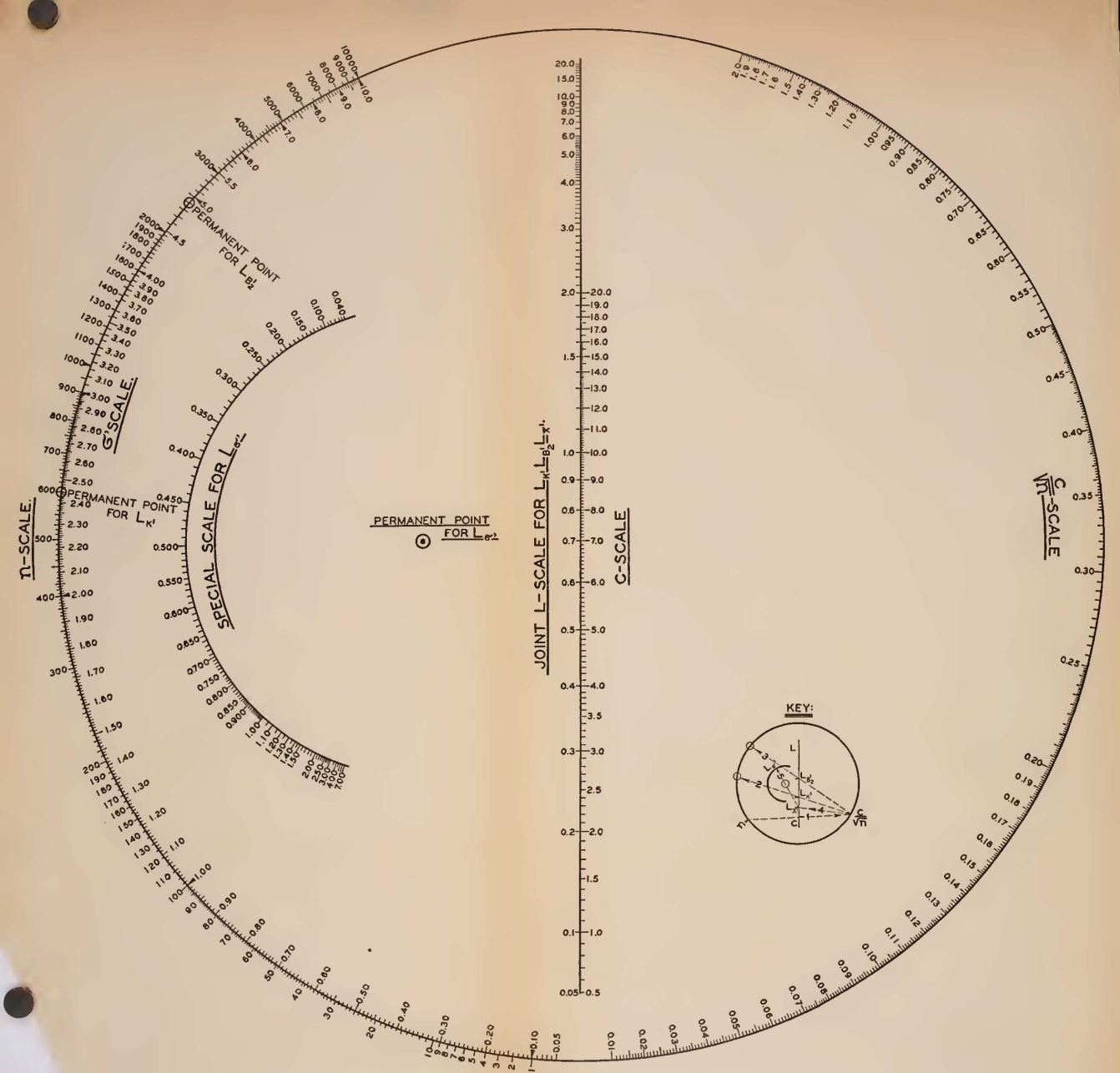

PERMANENT POINT FOR L_{B1}
SPECIAL SCALE FOR L_{B2}
PERMANENT POINT FOR L_{B2}
PERMANENT POINT FOR $L_{K'}$
G-SCALE
n-SCALE
JOINT L-SCALE FOR $L_K L_{B1} L_{B2} L_{K'}$
C-SCALE
$\frac{C}{\sqrt{n}}$-SCALE
KEY:

Fig. 5

Specification Type IV: The expected or average values of quality and standard deviation shall be $\overline{X}'$ *and* σ' *respectively. The expected modal and average qualities shall coincide and the probability function for the constant system of causes shall be monotonically decreasing for all values of* x *where* x *is measured from the mean.*

In this case the lower bound to the probability $P_{c\sigma'}$ is given by the expression [6]

$$P_{c\sigma'} \geq 1 - \frac{1}{2.25c^2}. \tag{3}$$

The limits can be obtained just as in case of Type III except that we use point B in Fig. 4 instead of point A. It may be easily verified by this nomogram that the Type IV values of L_1 and L_n for the special problem considered for Type III are $L_1 = 10.4$ ohms and $L_n = 0.33$ ohm respectively.

This shows that the additional requirements placed upon Type IV over those of Type III make for better control in the sense that the associated sampling limits are thereby decreased. By going further in adding restrictions upon the cause system, we gain even more marked improvements in the condition for control. In fact it is the system now to be described that has been found to be the most useful practical standard where the quality is measured as a variable.

Specification Type V: The system of causes shall yield a product distributed according to the Gram Charlier series [7] *with arithmetic mean* $\overline{X}'$, *standard deviation* σ', *skewness* k' *and kurtosis* β_2'.

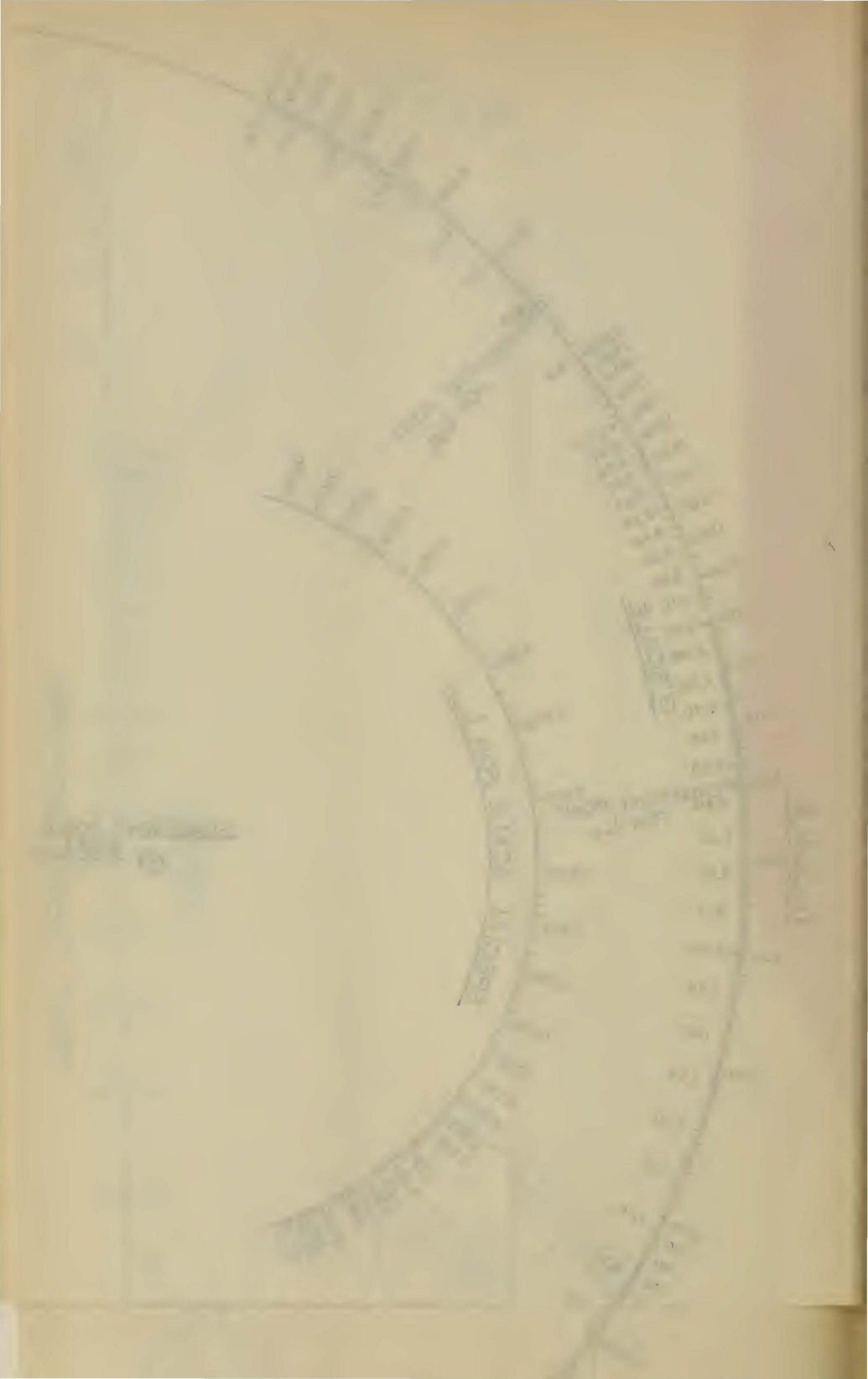
With the use of the four parameters we can detect lack of control of product through the failure of the observed value of any parameter determined from a sample of size n to fall within its sampling limits. It may happen that lack of control will be indicated by deviation beyond the sampling limits for only one of the four parameters. This case has already been illustrated in the article referred to in footnote 1. We shall now present, however, a method of setting these limits which is very easily applied.

As a specific example, let us assume the following expected values:

$$\begin{aligned} \overline{X}' &= 0, \\ \sigma' &= 1, \\ k' &= 0, \end{aligned}$$

[6] Camp, Burton H., "A New Generalization of Tchebycheff's Statistical Inequality," *Bulletin of the Amer. Math. Soc.*, December 1922, pp. 427–432. Eq. 3 is a special case of the general theorem of Camp. This theorem may be extended to determine lower bound to the probability of an error of the average as is done in this paper.

[7] Of course we might use certain other functions involving the same parameters.



Specification Type IV: The expected or average values of quality and standard deviation shall be $\overline{X}'$ *and* σ' *respectively. The expected modal and average qualities shall coincide and the probability function for the constant system of causes shall be monotonically decreasing for all values of* x *where* x *is measured from the mean.*

In this case the lower bound to the probability $P_{c\sigma'}$ is given by the expression [6]

$$P_{c\sigma'} \geq 1 - \frac{1}{2.25c^2}. \tag{3}$$

The limits can be obtained just as in case of Type III except that we use point B in Fig. 4 instead of point A. It may be easily verified by this nomogram that the Type IV values of L_1 and L_n for the special problem considered for Type III are $L_1 = 10.4$ ohms and $L_n = 0.33$ ohm respectively.

This shows that the additional requirements placed upon Type IV over those of Type III make for better control in the sense that the associated sampling limits are thereby decreased. By going further in adding restrictions upon the cause system, we gain even more marked improvements in the condition for control. In fact it is the system now to be described that has been found to be the most useful practical standard where the quality is measured as a variable.

Specification Type V: The system of causes shall yield a product distributed according to the Gram Charlier series [7] *with arithmetic mean* $\overline{X}'$, *standard deviation* σ', *skewness* k' *and kurtosis* β_2'.

With the use of the four parameters we can detect lack of control of product through the failure of the observed value of any parameter determined from a sample of size n to fall within its sampling limits. It may happen that lack of control will be indicated by deviation beyond the sampling limits for only one of the four parameters. This case has already been illustrated in the article referred to in footnote 1. We shall now present, however, a method of setting these limits which is very easily applied.

As a specific example, let us assume the following expected values:

$$\overline{X}' = 0,$$
$$\sigma' = 1,$$
$$k' = 0,$$

[6] Camp, Burton H., "A New Generalization of Tchebycheff's Statistical Inequality," *Bulletin of the Amer. Math. Soc.*, December 1922, pp. 427–432. Eq. 3 is a special case of the general theorem of Camp. This theorem may be extended to determine lower bound to the probability of an error of the average as is done in this paper.

[7] Of course we might use certain other functions involving the same parameters.

and

$$\beta_2' = 3.$$

Also let us assume that the size of the sample n for which the limits are to be established is 1000 and that we wish to establish limits upon the basis of a probability $P \doteqdot .997$.

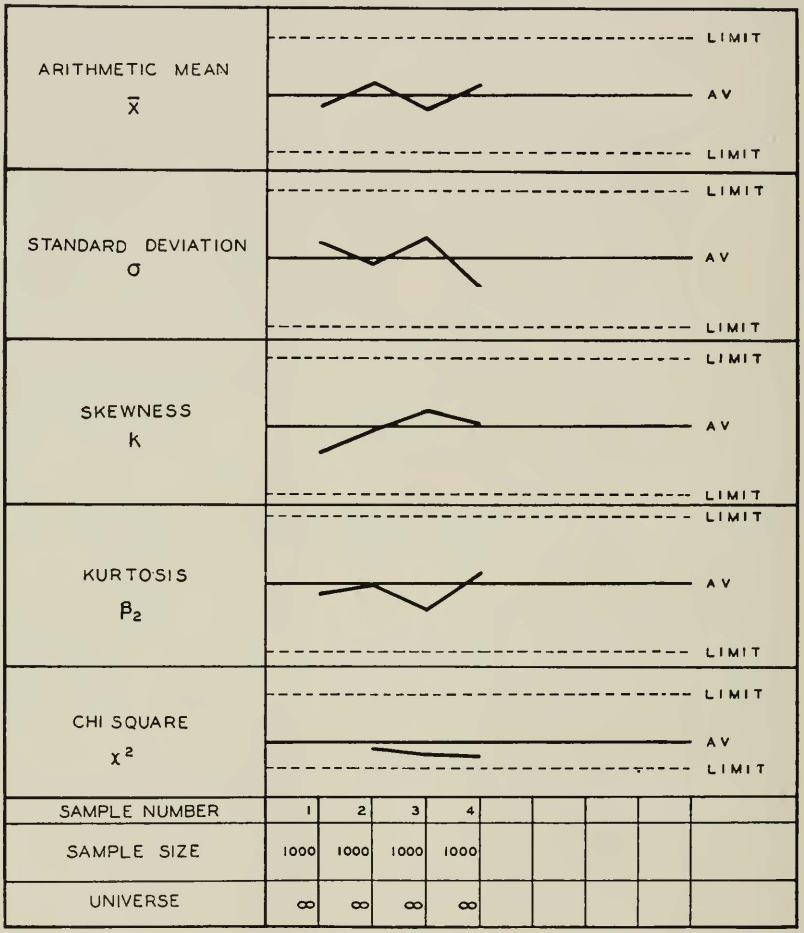


Fig. 6

The nomogram of Fig. 4 gives us immediately that $c = 3.0$ for $P = .997$. Hence we enter the nomogram of Fig. 5 on the c scale $c = 3.0$. The best way in which all four limits can be found by using the nomogram of Fig. 5 is then as follows, where the limits are set in the order $L_{k'}$, $L_{\beta_2'}$, $L_{\bar{X}'}$, and $L_{\sigma'}$. Join the point $n = 1000$ on the n scale

and $c = 3.0$ on the c scale by a straight line and thus find a pivot point on the $c/\sqrt{n}$ scale. Holding the ruler on this pivot point, join it successively with the permanent points of $L_{k'}$ and $L_{\beta_2'}$ and with σ' taken all on the σ' scale and read accordingly $L_{k'}$, $L_{\beta_2'}$, and $L_{\bar{X}'}$, on the L scale. After reading $L_{\bar{X}'}$, release the pivot point and turn the ruler around the $L_{\bar{X}'}$ point so as to join it with the permanent point for $L_{\sigma'}$. Then read the intersection of the ruler with the inner circular scale $L_{\sigma'}$, hereby obtaining the limit for σ'. Thus in five movements of the ruler we find all four limits: [8]

$$0 \pm L_{k'} = 0 \pm .23,$$
$$3 \pm L_{\beta_2'} = 3 \pm .46,$$
$$0 \pm L_{\bar{X}'} = 0 \pm .095,$$
$$1 \pm L_{\sigma'} = 1 \pm .067.$$

Figure 6 presents the graphical representation of the limits thus determined together with limits on χ^2 assuming that the theoretical frequency distribution was broken up into 13 cells.[9] The irregular lines show the fluctuations in the estimates of these parameters determined from four samples of 1000 each drawn under conditions satisfying the specifications just described for $\bar{X}' = 0$, $\sigma' = 1$, $k' = 0$ and $\beta_2' = 3$. Incidentally it should be noted that in every case the observed fluctuations in the estimates of the parameters are well within the sampling limits. This was to be expected because every effort was made in the sampling process to come as close as practicable to the ideal case where no assignable causes of variation were present. In this respect the data of Fig. 6 form an interesting contrast to the data of Fig. 4 of the article referred to in footnote 1, where evidence of lack of control was found.

Figure 7 makes it possible for us to set limits about the average or expected χ^2 corresponding to a probability of either .98 or .80. Thus for the data of Fig. 6 the limits for χ^2 corresponding to probability .98 are approximately 3 and 26 respectively as read from this chart. If limits corresponding to any other probability are desired, they can be readily obtained from tables for goodness of fit.[10]

We are now in a position to consider more in detail the advantages

[8] In case the given data bring the readings on the extreme points of the scale (where $\sigma' > 10$) it is advisable to take $\sigma'/10$ and multiply the final results obtained by ten. It is also helpful to remember that the L-scale on the nomogram of Fig. 5 can be considered as a regular scale of the product of two factors read on σ' scale and $c/\sqrt{n}$ scale.

[9] For the significance of χ^2 as here used, see paper, footnote 1.

[10] Elderton's Tables for Goodness of Fit reproduced in Pearson's "Tables for Statisticians and Biometricians" and also R. A. Fisher's "Tables for Goodness of Fit" given in his recent book "Statistics for Research Workers" will be found very helpful in the construction of curves similar to those of Fig. 7.

from a control viewpoint of Type V specification over the other suggested specifications. We have seen in Fig. 4 of the previous article on the control chart, footnote 1, that evidence of lack of control may be obtained through deviations in one parameter and not in

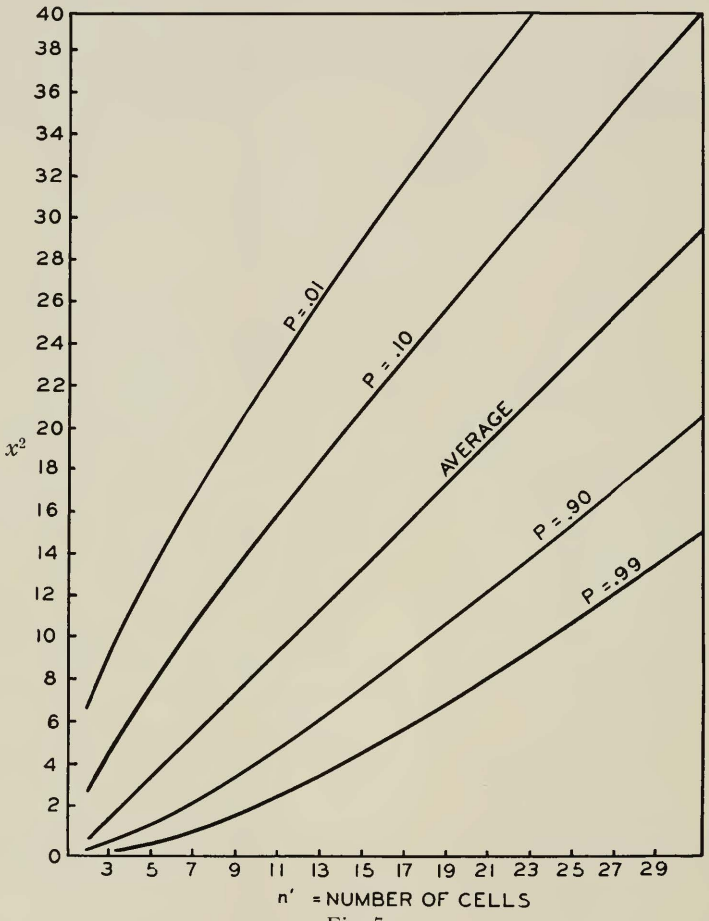


Fig. 7

another. For example, in this figure the per cent defective part of the chart corresponded to Specification Type I. Only 4 of the 12 points on this chart were outside the control limits whereas more than 4 points were outside the control limits for every other parameter and for the χ^2 part of the chart every one of the points was outside the control limits. Of course it is to be expected that the χ^2 test would be much more stringent than the test applied under Type I specification because the control limits established under the Type I specification are merely

the limiting case of the limits set on χ^2 for the case of two cells. We see, however, when samples are actually drawn from a constant system of causes, as was done as nearly as possible in obtaining the data for Fig. 6 of the present paper, all of the estimates of the parameters remain well within the sampling limits at least the expected proportion of the time.

To show that the limits set by means of Specification Type V upon the expected or average value of the data in Fig. 4 of the article on control charts just referred to are much smaller than could have been set by means of either Specification Type III or Specification Type IV, Fig. 8 is given. The limits based upon Specifications Type III and IV

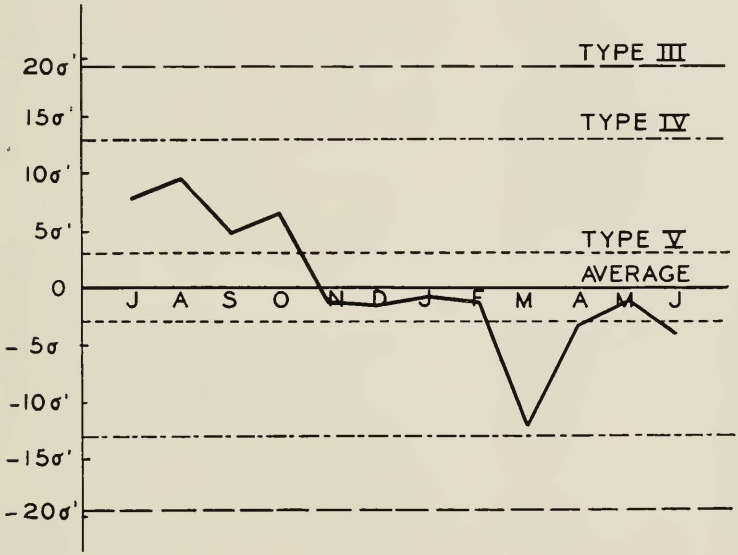


Fig. 8

were obtained directly from the nomogram of Fig. 4. The magnitudes of L_n stand in the order 19.3, 12.8 and 3.0. We see at a glance that lack of control, not indicated at all upon the basis of either Types III or IV, appears probable upon the basis of Type V.

Of course the use of the nomogram of Fig. 5 involves certain assumptions which now should be considered. The sampling limits are based upon the assumption that the sample is drawn from a normal universe. Even under these conditions the distributions of the values of estimates of the four parameters considered above are skew with the exception of that of the average, but approach normality as the size of the sample is increased. Theoretical and practical considera-

tions lead us to believe, however, that satisfactory limits can be established by the method just described making use of the nomogram of Fig. 5 provided the following restrictions as to the size of the sample are made.

(*a*) The expected distribution of the averages of samples of any size n is normal about the expected value $\overline{X}'$.

(*b*) Comparatively small error [11] will be made in fixing the limits on the parameter σ' by means of the nomogram of Fig. 5 provided n is 25 or more.

(*c*) For a sample of size n of 500 or more the nomogram of Fig. 5 may be used in fixing limits on all four parameters.[12]

These limitations do not require necessarily that the distribution of the estimate of a parameter must be normal for n as large or larger than specified; instead they merely require that it may be represented by the first few terms of the Gram Charlier series for which the normal law integral over a range equally divided by the expected value of the parameter is a close approximation to the integral of the Gram Charlier series over the same range.

Fixing the Parameters

There are various ways of arriving at the values of the parameters to be accepted as the basis for quality control. Sometimes they may be fixed by the economics of the problem. Such is the case for the Type I specification when the economic standard fraction defective or p' is known. At other times the parameters are fixed by technical considerations such for example as in the case of an induction coil whose inductance must lie within well-defined limits in order to obtain a proper functioning of the entire circuit, for this would effectively fix $\overline{X}'$ and σ'. In most practical instances the technical considerations tend to fix only the average and standard deviation. At other times we may empirically choose the observed estimates of these parameters determined from the data obtained within the fixed interval of time wherein we have reason to believe the quality has been produced under essentially the same conditions. Irrespective, however, of what period is chosen as a base in fixing p' or any other parameter, the control chart serves to show whether or not the product has been controlled over this period. In any case the parameters are accepted at least as

[11] Pearson, Karl, "On the Distribution of Standard Deviations of Small Samples," *Biometrika*, Vol. X, Part IV, May 1915, pp. 522–529.

[12] Pearson, Karl, and others, "On the Probable Errors of Frequency Constants," *Biometrika*, Vol. XIV, 1903, p. 273 seq., Vol. IX, 1913, p. 22 seq. Isserlis, L., "On the Conditions under Which the Probable Errors of Frequency Distributions Have a Real Significance," *Proceedings of the Royal Society*, Series A, Vol. XCII, 1915, pp. 23–41.

temporary standards. In every case the choice of the fixed values calls for the exercise of engineering judgment. The statistical problem enters after these standards have been fixed. It is to determine whether or not the observed fluctuations in the observed estimates of the parameters are explainable upon the basis of chance. In general, the method of fixing the limits closely corresponds to that whereby a manufacturer sets up specifications for any kind of product.

It should be noted that from a statistical standpoint the control charts are based upon a priori reasoning. The type of cause system specified by the engineer is taken as a standard a priori system which is accepted as an ideal which the manufacturer hopes to maintain. The control chart thus makes it possible to differentiate between deviations in quality which can reasonably be accounted for on the basis of sampling and those deviations which cannot be so accounted for.

It will have been noted that the limits are a function of the size of the sample n. The question is therefore often raised: How large a sample shall be chosen?

So long as we are willing to risk our engineering judgment that the system of causes is controlled, we need take no samples. If, however, we have reason to believe that the quality has not been controlled or at least wish to make sure that it is being controlled to the extent that the deviation introduced by the assignable cause shall not escape detection if greater than some chosen value, it is necessary for us to take a sample of sufficient size to reduce the limits of sampling fluctuations in the particular parameter under study to just less than this same value.

In those cases where customary practice calls for the inspection of a certain number of units of product for reasons other than control, these data may be used in the manner outlined above to indicate the degree of control. In many instances the number of units of product to be inspected is so fixed as to insure with a known degree of probability that the apparatus passing from one stage of the manufacturing process to another meets a given tolerance for defects. This practice serves to fix the number to be inspected in order to maintain a given quality of apparatus as it passes through the stages of the manufacturing process. The use of the data so obtained in the form of a control chart serves to fix attention upon the assignable causes of variation in the quality. The presence of these causes having been detected, it generally becomes a comparatively simple matter to find and eliminate them. In this way we can secure a controlled product usually requiring less inspection and hence involving the lowest cost of manufacture.

I am indebted to Mr. V. A. Nekrassoff for the construction of the nomograms presented in this paper.

The New York-London Telephone Circuit

By S. B. WRIGHT and H. C. SILENT

SYNOPSIS: This paper discusses the special provisions which are in use on the transatlantic telephone to compensate for the variability of the wire and ether paths, for the radio noise, and for the fact that two-way transmission is effected upon a single wave-length. So-called technical operators are in attendance at each end of the radio path and are equipped to adjust the magnitude of the speech currents entering the radio transmitters to such a value as to load these transmitters to capacity. The amplification introduced at the radio receivers can also be adjusted to compensate for changes in the transmission efficiency of the radio paths. Finally, voice-operated relays together with suitable delay circuits are provided which so control the apparatus that at any given time it can transmit in but one direction. By this arrangement, a speaker's voice upon leaving his transmitting station cannot operate his own receiver although this is tuned to the transmitting wave-length.

TO the telephone subscribers who use the New York-London circuit the procedure of making a call and carrying on a conversation is as simple as that of any long distance telephone call. Even to the telephone operator who establishes a transatlantic connection there is little to differentiate the New York-London "wire-radio-wire" circuit from the hundreds of other circuits which appear as mere jacks on the switchboard in front of her. Beyond this point, however, there is an organization of physical plant, personnel and procedure very much different from the usual long distance telephone circuit.

Without going into any description of the radio portion of the New York-London circuit, which has been adequately treated in previous articles, this paper describes some of the interesting features of the circuit, including the method of electrical operation which has been worked out for making possible two-way talking in the usual way, in spite of difficulties introduced by "static," transmission variations and difficulties brought about by the use of the same "frequency band" for transmission in both directions. The method of operation involves manual adjustments of controls at the radio stations and at the circuit terminals, and automatic switching by means of vacuum tube-operated relays controlled by the voice currents of the telephone subscribers. The voice-operated relay system is particularly interesting, and is, therefore, rather fully described.

Before the operation of the circuit is described a brief general picture of the system will be given. Fig. 1 shows its geographical layout, and gives an idea of the relative lengths of wire and radio circuit involved.

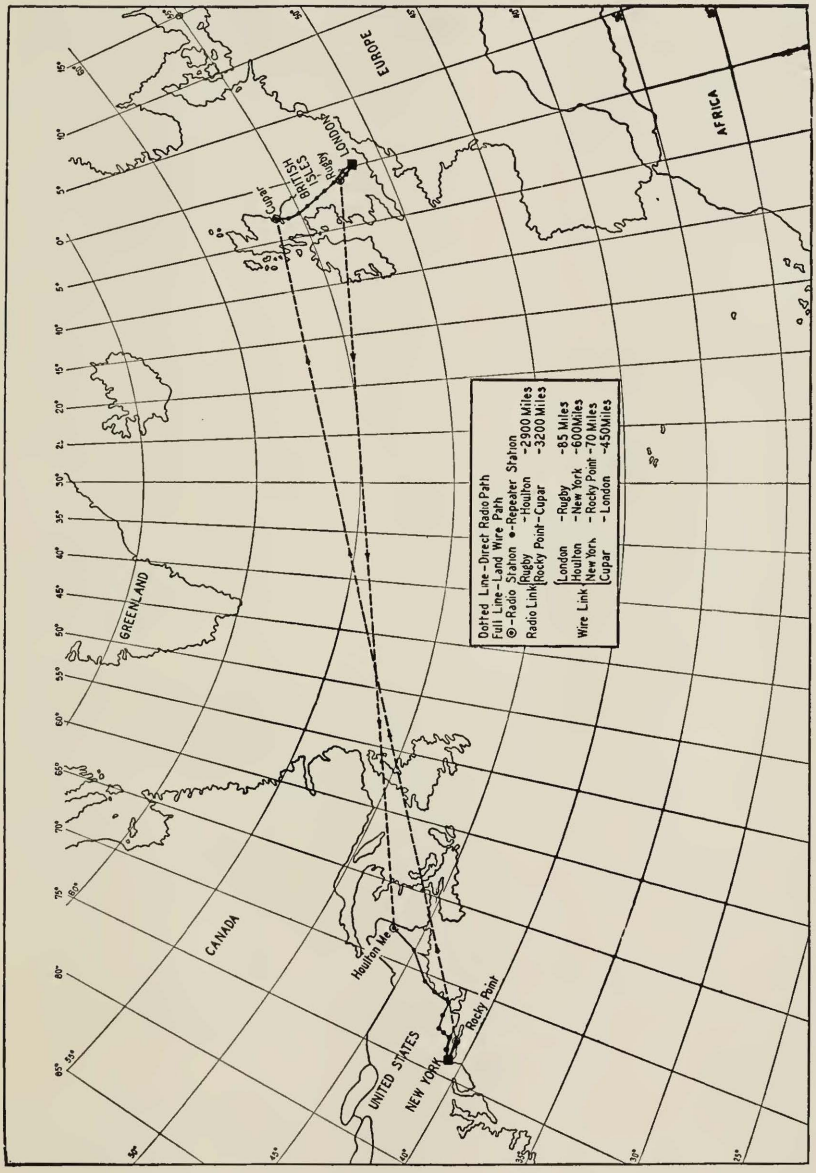

EUROPE
AFRICA
GREENLAND
CANADA
UNITED STATES
BRITISH ISLES
LONDON
Rugby
Cupar
Houlton Me.
NEW YORK
Rocky Point
Dotted Line–Direct Radio Path
Full Line–Land Wire Path
◎–Radio Station ●–Repeater Station
Radio Link Rugby –Houlton –2900 Miles
Rocky Point–Cupar –3200 Miles
Wire Link London –Rugby –85 Miles
Houlton –New York –600Miles
New York –Rocky Point–70 Miles
Cupar –London –450Miles

Fig. 1.

Fig. 2 (*A*) is a schematic circuit diagram emphasizing the land wire sections to permit showing the locations of intermediate repeater stations and terminal apparatus. This figure shows that the transatlantic is similar to a long "four-wire" land telephone circuit in which speech travels over different paths in the two directions. These two branches are combined into a single circuit at the terminals where special apparatus, including automatic switching devices, and specially trained men known as "technical operators" are stationed. The usual long distance girl operator establishes connections to subscribers.

As will appear shortly, the duties of a technical operator have nothing to do with the setting up of connections but require him to be continuously attentive to the electrical operation of the circuit, and to make adjustments of the amplification in the wire lines whenever the strength of voice currents bound for the radio transmitter changes. He is enabled to do this by watching the indicating needle of a sensitive vacuum tube-operated meter, called a "volume indicator." The volume indicator shows the strength or weakness of the electrical speech waves in the line. Alongside of this meter are located the dials with which he controls the amplification. Fig. 3 shows a technical operator watching the meter at the New York terminal. The apparatus shown on the panels in this picture includes the necessary terminal amplifiers and devices for adjusting and maintaining various parts of the wire and radio system.

Fig. 2 (*B*) shows the relative strength of voice waves or "electrical volumes" at various points in the circuit when a telephone subscriber in England is talking to one in the United States. The broken lines in this diagram indicate the magnitude of variations in the electrical volumes delivered to the circuit and received from it, as well as transmission variations in the radio section or "link." The relative values of electrical volume in both directions of transmission are, of course, essentially similar. The voice currents require about 1/15 of a second to travel from either terminal to the other over the circuit. It is interesting to consider that only about one fourth of this time is occupied in traversing the radio link, although radio constitutes about 85 per cent of the total length of the circuit, the remainder being in the wire lines and terminal apparatus.

It is important to note from Fig. 2 (*B*) that the ratio of the strongest to the weakest electrical volumes sent into the circuit at a terminal may be as much as 1,000 times. This is indicated at (*a*) in the figure. The variation is due partly to the different ways in which the subscribers talk, and partly to the variation in losses in the lines which connect the subscribers to the circuit. The technical operator adjusts

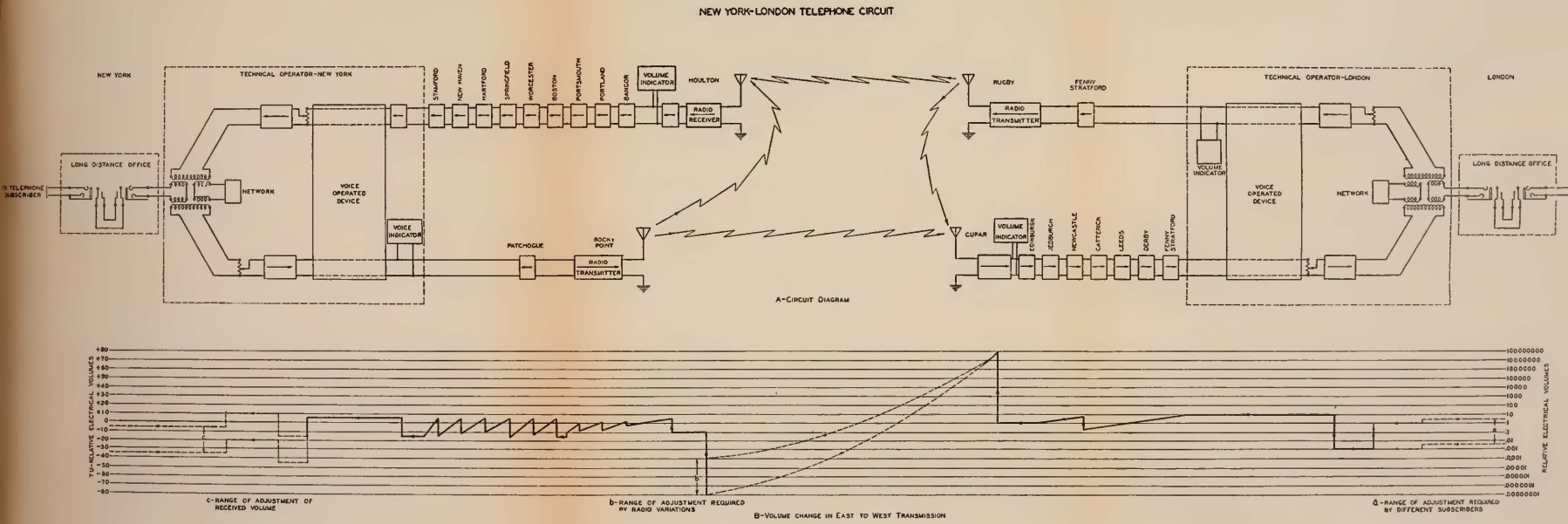

NEW YORK-LONDON TELEPHONE CIRCUIT
NEW YORK
TECHNICAL OPERATOR-NEW YORK
LONG DISTANCE OFFICE
TO TELEPHONE SUBSCRIBER
NETWORK
VOICE OPERATED DEVICE
VOICE INDICATOR
STAMFORD
NEW HAVEN
HARTFORD
SPRINGFIELD
WORCESTER
BOSTON
PORTSMOUTH
PORTLAND
BANGOR
VOLUME INDICATOR
HOULTON
RADIO RECEIVER
PATCHOGUE
ROCKY POINT
RADIO TRANSMITTER
RUGBY
RADIO TRANSMITTER
FENNY STRATFORD
CUPAR
VOLUME INDICATOR
EDINBURGH
JEDBURGH
NEWCASTLE
CATTERICK
LEEDS
DERBY
FENNY STRATFORD
TECHNICAL OPERATOR-LONDON
VOLUME INDICATOR
VOICE OPERATED DEVICE
NETWORK
LONDON
LONG DISTANCE OFFICE
TO TELEPHONE SUBSCRIBER
A-CIRCUIT DIAGRAM
TU-RELATIVE ELECTRICAL VOLUMES
+80 +70 +60 +50 +40 +30 +20 +10 0 -10 -20 -30 -40 -50 -60 -70 -80
RELATIVE ELECTRICAL VOLUMES
100000000 10000000 1000000 100000 10000 1000 100 10 1 .1 .01 .001 .0001 .00001 .000001 .0000001 .00000001
C-RANGE OF ADJUSTMENT OF RECEIVED VOLUME
b-RANGE OF ADJUSTMENT REQUIRED BY RADIO VARIATIONS
a-RANGE OF ADJUSTMENT REQUIRED BY DIFFERENT SUBSCRIBERS
B-VOLUME CHANGE IN EAST TO WEST TRANSMISSION

Fig. 2.

the amplifiers so as to keep the electrical volumes reaching the radio transmitter at a predetermined value. The technical operator also adjusts the received volume over the range indicated at (*c*) to give the best operation under different conditions of static, and for different types of connections.

An interesting fact to be observed here is that the voice power is amplified about 100,000,000 times in the radio transmitter and anywhere from 30,000 to 300,000,000 times at the radio receiver and associated amplifiers, depending on the loss in the radio path at any particular time as indicated at (*b*). Including the amplification which it is necessary to use on the wire "links," the total power amplification in either direction is approximately 10^{40}. Although more amplification is used in this circuit at a single point, such as at the radio transmitter, than at a single point in any other commercial telephone circuit, the total power amplification is less than in one of the telephone cable circuits from New York to St. Louis, where it is approximately 10^{50}.

Having in mind the foregoing facts, one can appreciate the difficulties which had to be faced in the way of operating this circuit and which have been successfully overcome. The more important may be summarized as follows:

(1) The transmission losses through the ether in the radio links vary from time to time in an irregular manner at intervals which preclude the possibility of making predetermined or systematic compensating adjustments of the amplification at the radio receivers.

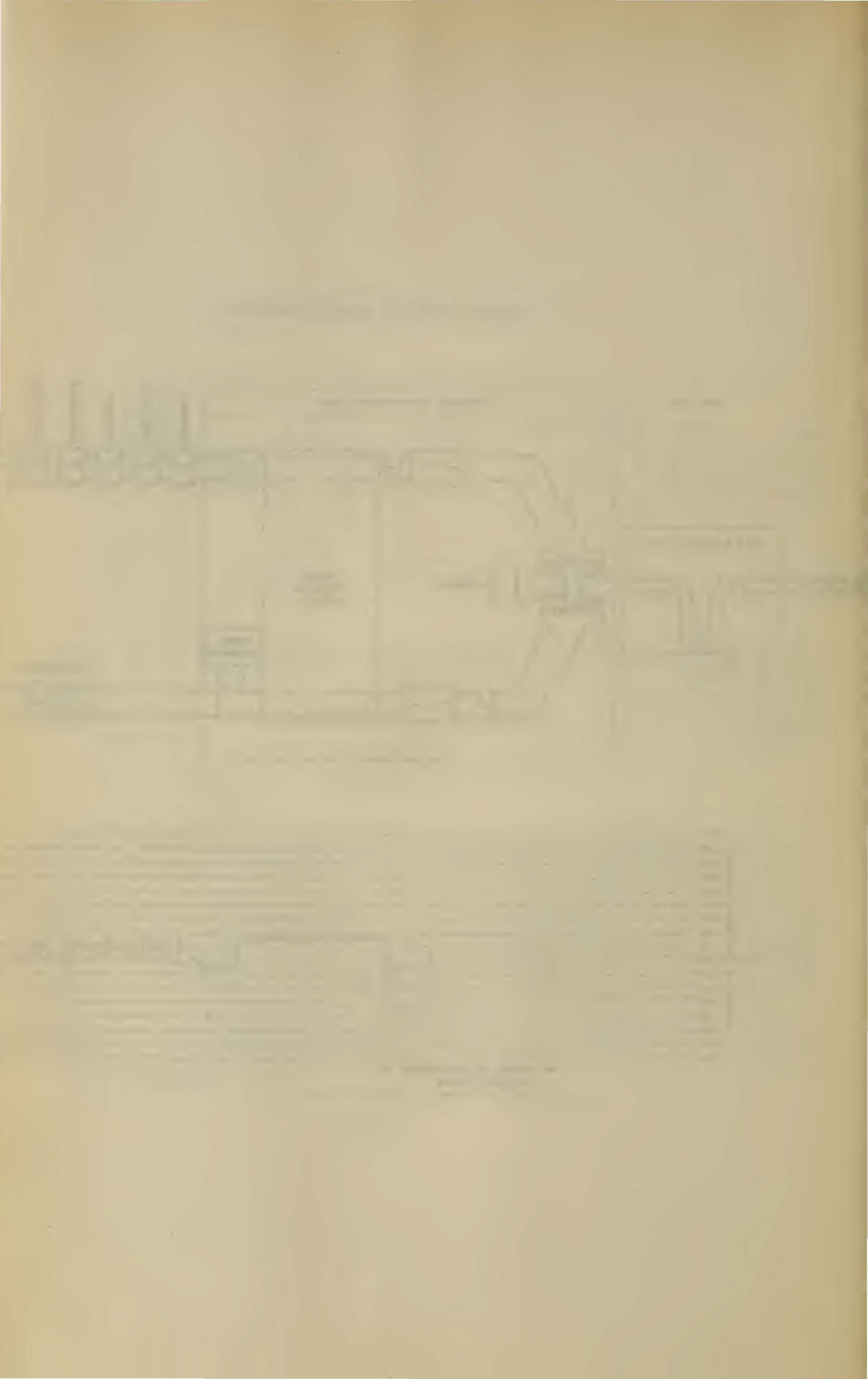
(2) The radio links are frequently more noisy than wire circuits. This noise consists principally of stray electric waves (static) and varies considerably from time to time.

(3) The tendency for strong echo currents to exist in this circuit is considerably greater than in ordinary wire circuits. This is due partly to the methods employed for overcoming the difficulties brought about by (1) and (2), and partly to the fact that radio transmission in the two directions is carried out in the same frequency band.

These difficulties have been overcome by the following means:

(1) To overcome the variations in the transmission efficiency of the radio links, adjustments are made from time to time of the amplification in the radio receivers. Radio operators are in constant attendance at the receiving stations in order to make these adjustments.

(2) To assist in overcoming the effect of radio noise, adjustments are



the amplifiers so as to keep the electrical volumes reaching the radio transmitter at a predetermined value. The technical operator also adjusts the received volume over the range indicated at (*c*) to give the best operation under different conditions of static, and for different types of connections.

An interesting fact to be observed here is that the voice power is amplified about 100,000,000 times in the radio transmitter and anywhere from 30,000 to 300,000,000 times at the radio receiver and associated amplifiers, depending on the loss in the radio path at any particular time as indicated at (*b*). Including the amplification which it is necessary to use on the wire "links," the total power amplification in either direction is approximately 10^{40}. Although more amplification is used in this circuit at a single point, such as at the radio transmitter, than at a single point in any other commercial telephone circuit, the total power amplification is less than in one of the telephone cable circuits from New York to St. Louis, where it is approximately 10^{50}.

Having in mind the foregoing facts, one can appreciate the difficulties which had to be faced in the way of operating this circuit and which have been successfully overcome. The more important may be summarized as follows:

(1) The transmission losses through the ether in the radio links vary from time to time in an irregular manner at intervals which preclude the possibility of making predetermined or systematic compensating adjustments of the amplification at the radio receivers.

(2) The radio links are frequently more noisy than wire circuits. This noise consists principally of stray electric waves (static) and varies considerably from time to time.

(3) The tendency for strong echo currents to exist in this circuit is considerably greater than in ordinary wire circuits. This is due partly to the methods employed for overcoming the difficulties brought about by (1) and (2), and partly to the fact that radio transmission in the two directions is carried out in the same frequency band.

These difficulties have been overcome by the following means:

(1) To overcome the variations in the transmission efficiency of the radio links, adjustments are made from time to time of the amplification in the radio receivers. Radio operators are in constant attendance at the receiving stations in order to make these adjustments.

(2) To assist in overcoming the effect of radio noise, adjustments are

made of the amplification in the wire links so that the radio transmitter is fully loaded up. This permits radiation of full power regardless of how loudly or weakly the subscriber talks, and regardless of the length of the circuit between the subscriber and the transatlantic terminals. This keeps the radio speech waves as large as possible compared to the noise at all times. These adjustments are made by the technical operators under the guidance of the "volume indicators."

(3) To suppress echo effects, a system of voice-operated switching relays has been devised whose function is to interrupt, when not in use, any transmission path which may double back to its source and give rise to echoes or singing in the circuit.

The manual adjustments of controls required in (1) and (2) should require no further explanation.

Before describing the voice-operated switching system of (3), it will be desirable to explain what this system is required to do. As previously stated, the adjustments employed to eliminate the two radio effects—namely, variability and noise—tend to increase the severity of echo effects. This follows from the fact that such adjustments result in a net transmission loss from terminal to terminal which is not constant as in ordinary telephone circuits, but which varies from time to time depending on the loss in the ether path and the strength of the voice currents which are delivered to the circuit terminal. The overall transmission of the circuit may vary from a loss to a considerable gain. If means were not taken to prevent it, this gain would set up between the two subscribers, circulating currents of rather large amplitude producing either severe electrical echo effects or the totally inoperative condition known as "singing."

A further echo difficulty was brought about through the use of a common frequency band or group of wave-lengths for transmitting in both directions. This was highly desirable to reduce the amount of frequency space occupied in the ether since there is but a limited suitable frequency space available. The radio waves at the frequencies used (namely, 58.5 to 61.5 kilocycles) cannot be directed to a definite point or confined to a single path. The radio receiver cannot, of course, when both transmitters are operating at the same frequency, select one transmitter from the other by any ordinary tuning means. Referring to Fig. 1, since the distance from the receiver at Houlton, for example, to Rocky Point is much less than the distance from Houlton to Rugby, the antenna at Houlton is exposed to a signal from the transmitter in America which is much stronger than the signal from

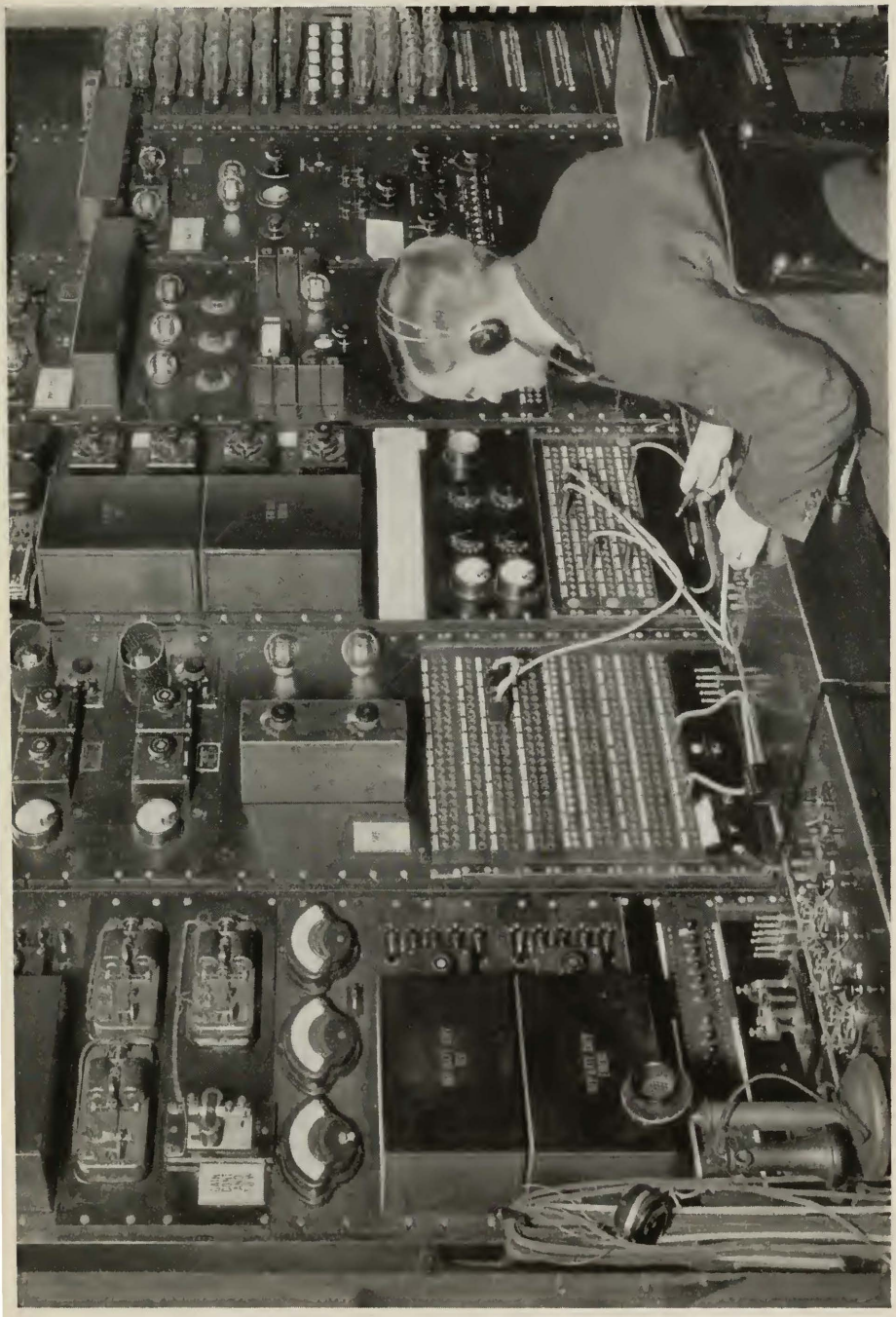


Fig. 3.

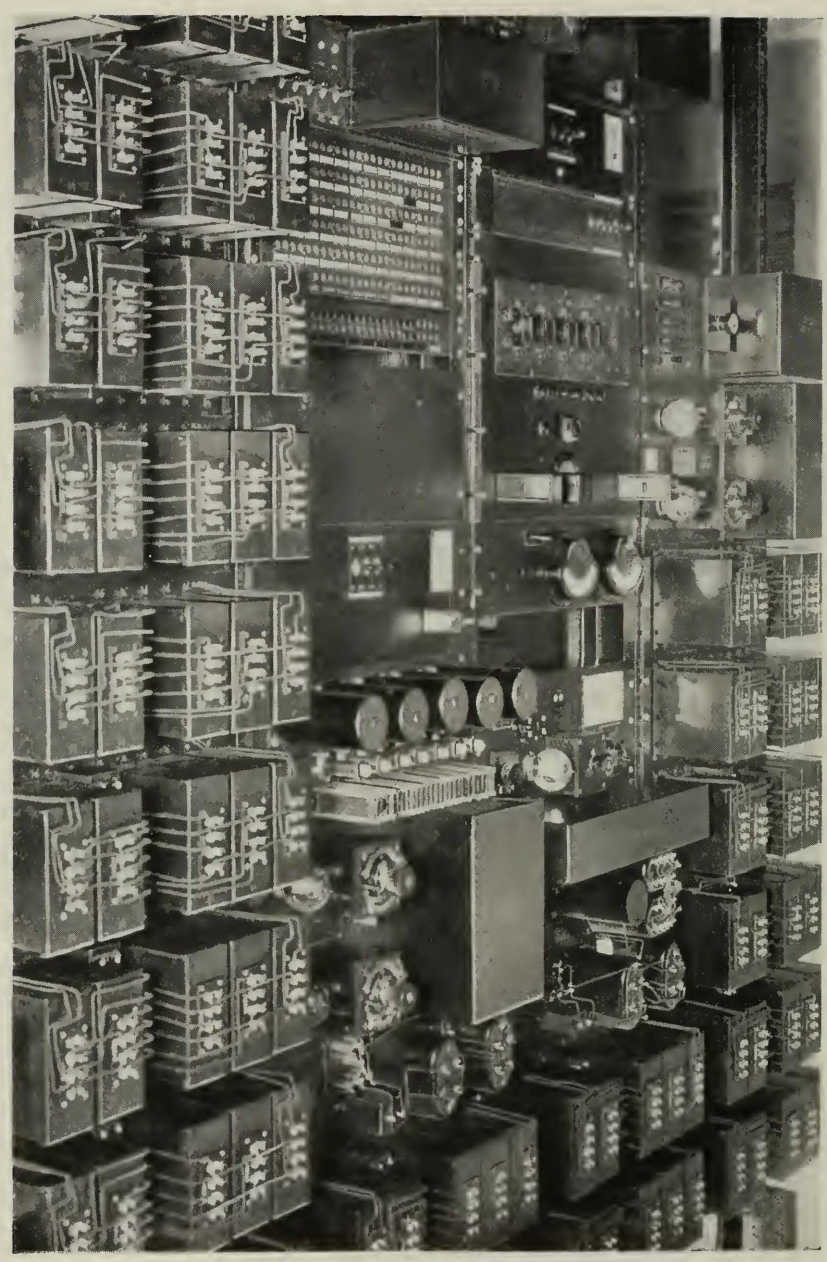


Fig. 4.

England.[1] Were it not for the use of the voice-operated devices, this would result in a very strong echo being returned to the local talker with disconcerting effects, and, dependent upon the adjustment of other parts of the circuit, might even result in violent singing of the circuit.

Referring again to Fig. 2 (*A*), it will be seen that there are three paths capable of giving rise to objectionable echoes or singing, one path at each end of the circuit through the wire lines, radio transmitter and local receiver, and a third path from end to end of the circuit and back again. The first two paths are introduced by using the same frequency band for transmission in both directions. The third path which depends upon the impedance unbalance between the two-wire lines at the terminals and their respective networks is similar to the one which gives rise to echoes in long four-wire land telephone circuits. All three paths are affected by the amplification adjustments.

Suppression of echoes and singing in the circuit requires that all three of these echo paths be kept blocked at all times against unwanted transmission. Furthermore, since there is no single point common to all the echo paths, the system for suppressing echoes comprises two separate installations—one of which is located in New York and the other in London. The devices used to control the echo paths are operated by the voice currents of the two telephone subscribers, in such a manner as to allow transmission to pass first in one direction when one subscriber is talking, and then in the other direction when the second subscriber replies. Transmission in the opposite direction to that in which the waves are traveling is blocked. When no one is talking, the outgoing transmission paths at both ends of the circuit are blocked. The necessary functions at the New York end of the circuit are performed by a combination of electro-magnetic relays, vacuum tube detectors and delay circuits. A photograph of the installation is shown in Fig. 4. At London a device performing similar functions has been developed by engineers of the British General Post Office.[2]

A schematic diagram of the device employed at the New York end is shown in Fig. 5. By tracing the action of the relays it will be seen that for all conditions of the relays, the echo paths shown are blocked at the proper times. Thus, Fig. 5, which shows the conditions when no one is talking, indicates that the circuit from the radio receiver to the terminal is normally in a receiving condition but the transmitting branch of the circuit is kept inoperative by relays *SS* and *CS*.

[1] Directive antenna systems with a blind spot might be used to overcome this effect, but their directive properties would not then be available for use against static and other interference. The general directivity of the receiving systems used, however, reduces the unwanted transmissions about 100-fold.

[2] C. A. Beer and G. T. Evans, "The Post Office Differential Voice-Operated Anti-Singing Equipment," *P. O. E. E. Jnl.*, April, 1927.

APPLICATION OF VOICE OPERATED DEVICES TO NEW YORK–LONDON TELEPHONE CIRCUIT

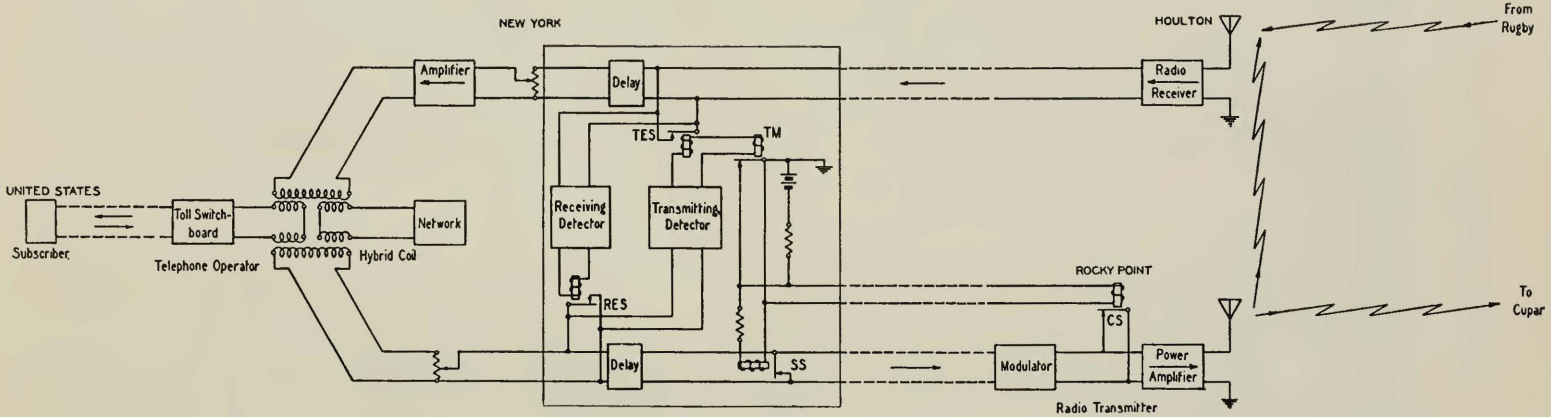


Fig. 5.

When the United States subscriber speaks, a small portion of his voice currents enters a detector, operating relays *TM* and *TES*. The action of relay *TM* causes the operation of relays *SS* and *CS*, thus clearing the outgoing line. The operation of relay *TES* short-circuits the receiving line. The main part of the outgoing voice currents passes on through the delay circuit, wire line and radio transmitter to the distant subscriber. Any transmission picked up by the radio receiver is blocked by relay *TES*. When the subscriber stops talking, the relays are restored to the normal condition.

While relays *SS* and *TES* are sufficient to switch the speech paths back and forth and prevent singing, Fig. 5 shows that there are two other relays which also interrupt undesired transmission. One of these is relay *RES*, which operates when a subscriber in England is speaking and short-circuits the United States transmitting line. This short-circuit prevents the transmitting relays from being operated by the echo of received currents returned from the local subscriber's line. The other relay, shown in Fig. 5 as *CS*, is not used at present but was needed when the circuit was first opened due to the fact that the radio transmitter and receiver in the United States were much closer together than they now are. The action of relay *CS* is similar to that of *SS*, but it was located at the radio transmitting station for an interesting reason. Although the radio transmitter is of a type which should radiate energy only during the actual transmission of speech, it would, were it not for relay *CS*, put a certain amount of noise energy into the air. While this noise, which originates partly in the radio transmitter and partly in the wire lines connecting it to the terminal, is too weak to be heard at the distant terminal, it was strong enough when picked up by the radio receiver at Riverhead, Long Island, to interfere with reception of the distant transmitter. Relay *CS* suppressed any such noise when the transmitter was idle, that is, when no one was speaking from New York. There are a number of ways of operating relay *CS*; either by voice currents rectified at the radio transmitter or via a wire circuit from the action set up by voice currents at the terminal. This latter method is shown in Fig. 5. When the United States radio receiver was moved to Houlton, Maine, the use of relay *CS* was discontinued, as the noise currents picked up there from Rocky Point were negligible.

A graphical representation of the time functions of two of the relays on the transmitting side of the voice-operated device is shown in Fig. 6. This illustrates the action during a representative spoken syllable. It will be noted that relay *SS* does not operate at the exact beginning of the speech. As shown at *a* in Fig. 6, it is necessary for the speech wave

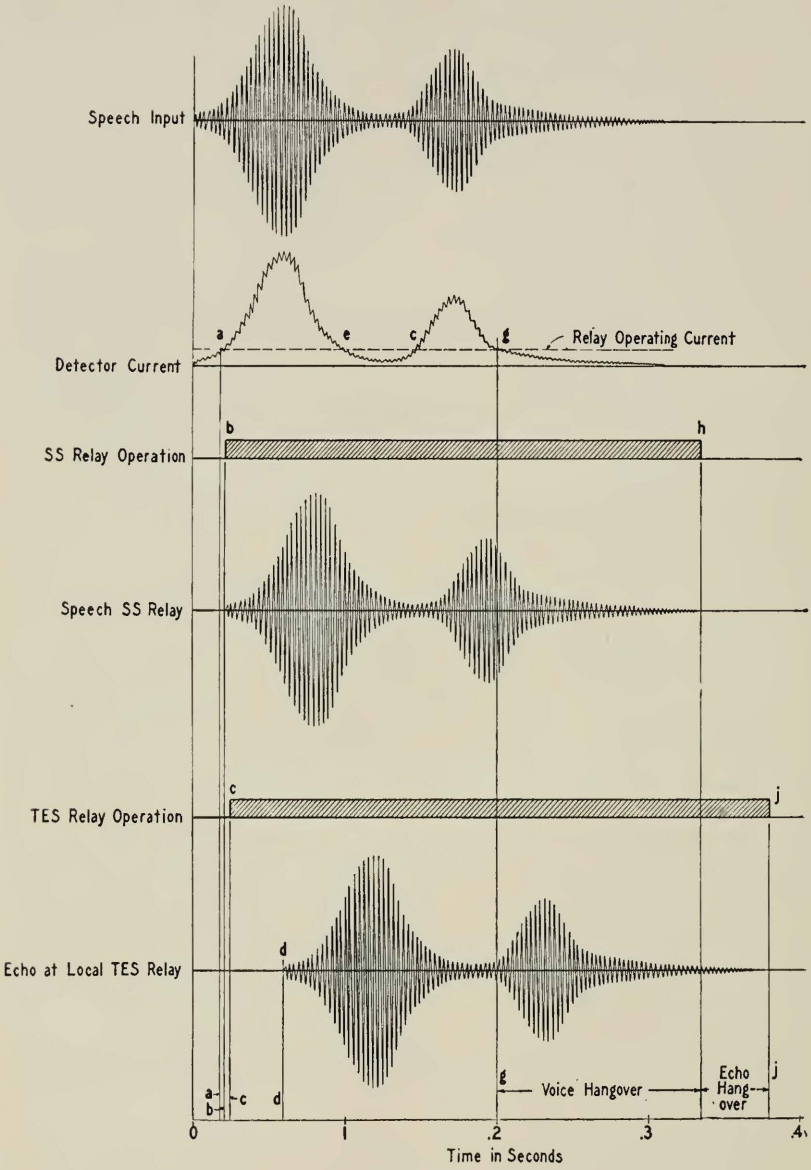

Speech Input
Detector Current
a
e
c
g
Relay Operating Current
SS Relay Operation
b
h
Speech SS Relay
TES Relay Operation
c
j
Echo at Local TES Relay
d
a
b
c
d
g
Voice Hangover
Echo Hang-over
j
0
1
.2
.3
.4
Time in Seconds

Fig. 6.

to build up to a certain definite amplitude before operation can begin. After this a finite time *a–b* is required for the relay to operate. During the interval from *O* to *b*, the voice currents are passing through the delay circuit. Thus the relay clears the path in time to transmit fully all necessary parts of this speech wave even though they have a very weak beginning.

The final trailing weak part of the speech wave is allowed to pass, by making the *SS* and *CS* relays release slowly by means of suitable circuits not shown on the drawing. This effect is shown from *g* to *h* in Fig. 6, as the "voice hangover" action. This action also functions to hold relay *SS* operated during any momentary weak lapses of speech between parts of the syllable as shown at *e–f* in the speech wave in Fig. 6.

As previously mentioned, a strong echo of the outgoing transmission is picked up at the radio receiver and suppressed by relay *TES*, the operation of which is indicated at *c* in Fig. 6. This echo is delayed an amount represented by *O* to *d* in being transmitted over the wire circuits. Relay *TES* has a releasing time slower than that of relay *SS* by the amount *h–j*, which is sufficient to care for the time that the echo is delayed.

In the operation of this system it is necessary for relays *SS* and *CS* and the devices which operate them to be sufficiently sensitive to operate on all parts of the outgoing speech sounds in order that none may be lost. On the other hand, relay *RES* need operate only on impulses which, if allowed to be transmitted across the multi-winding transformer ("hybrid" coil) as echoes, would be strong enough to falsely operate the relays associated with the transmitting side. Use of a relay on the receiving side which normally blocks reception would be possible, but this would require very much greater sensitivity. Due to the noise (static) introduced by the radio links, the use of such a sensitive relay is undesirable. Therefore, the device has been made to have a transmission path normally free in the direction in which the noise is high and a normally blocked path in the direction in which only the noise on the two-wire line need be combated.

If it were not for the delay circuit on the transmitting side, it would be necessary to increase the sensitivity of the voice-operated relay device so that the relays would obtain enough current to cause their operation at the very beginning of the speech wave, rather than allow the wave to build up to an appreciable amplitude before operation occurs. This delay circuit, therefore, permits appreciable reduction in the sensitivity of the transmitting side of the device, reducing the probability of operation by noise from the connected subscriber's line.

It may be said in passing that by an increase in sensitivity it is perfectly possible, with the extremely fast relays used, to omit this delay circuit and obtain satisfactory operation. This would, however, make the device more subject to noise effects.

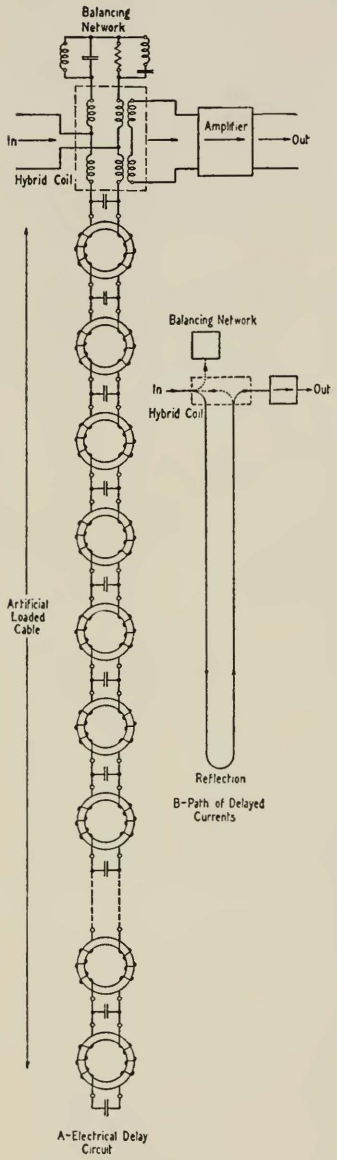


Fig. 7.

Referring to Fig. 5, a delay circuit is also included in the receiving branch of the circuit which, however, performs a somewhat different function from that in the transmitting branch. This delay circuit serves to delay transmission across the hybrid coil, thereby permitting the relay *RES* to operate and apply its protecting short-circuit before the echo from the connected circuit arrives at the input to the transmitting detector. In suppressing the echo from the radio receiver by relay *TES* a similar action is performed by the delay in transmission over the wire lines to the radio stations.

The type of delay circuit used in the voice-operated device just described is shown in Fig. 7 (*A*). This consists of an electrical network by means of which transmission sent into it is received at its output after a finite time interval. To obtain this delay action a loaded artificial line having low attenuation is employed, in conjunction with a network which balances its "surge" impedance, and a hybrid coil. An interesting feature of the arrangement is that the principle of reflection, which tends to cause objectionable echoes in telephone circuits, is here usefully employed to pass the voice currents through the artificial line twice. This results in a considerable saving of apparatus. Fig. 7 (*B*) shows the path of transmission through the delay circuit. Alternating current entering the hybrid coil divides equally between the delay circuit and the balancing network. That part which enters the balancing

network is dissipated; that part which enters the delay circuit is transmitted through it, with small attenuation, to the end. Here it meets with a reflecting termination and is transmitted back whence it came. Reaching the hybrid coil, half of it goes back toward the input and half of it is transmitted on in the desired direction. The half which goes to the input meets the output side of a one-way amplifier and is dissipated. The remaining half passes through an amplifier which makes up for the transmission loss of the delay circuit and the loss due to the two divisions of energy at the hybrid coil.

The desirability of maintaining the proper relationships between the time actions of the relays and the delays in the other parts of the system will be apparent from the foregoing. A circuit for measuring the time of operation of the relays is provided which in combination with a detector and a relay may also be used for measuring the time required for alternating currents to travel through a delay network or other telephone circuit. This device is capable of measuring directly time intervals as short as .0001 second and up to 1 second in length. The measuring device is conveniently located along with the apparatus comprising the voice-operated device, as shown in Fig. 4.

In conclusion, it should be pointed out that the method of operation that has been described is expensive and has disadvantages which make it undesirable on any but a very special and necessarily complex telephone circuit, such as the transatlantic. It has, however, proved satisfactory in this service. The more interesting effects which this method of operation accomplishes may be restated as follows:

Given the condition of an anti-singing circuit such as the New York-London radio circuit, it is possible to make the amount of power radiated from the radio transmitting stations independent of the strength of the voice currents arriving over the land lines. For example, a subscriber speaking from a suburb of Chicago is heard just as loudly in London as another person speaking from the terminal of the circuit at New York.

As a result, voices of all talkers, strong or weak, are heard with the same freedom from static.

Both of the above effects result from the adjustment of the strength of outgoing speech so as to load the radio transmitter to maximum output for all messages. If the circuit were operated with amplification fixed at a value required by the strongest talkers, then the voices of weak speakers would often be lost unless the power of the radio transmitter were increased several hundred fold.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

Thermal Agitation in Conductors.[1] H. NYQUIST. At the December, 1926, meeting of the American Physical Society, J. B. Johnson reported the discovery and measurement of an e.m.f. due to the thermal agitation in conductors. The present paper outlines a theoretical derivation of this effect. A non-dissipative transmission line is brought into thermodynamic equilibrium with conductors of a definite temperature. The line is then isolated and its energy investigated statistically. The resultant formula is $E_\nu^2 d\nu = 4kTRd\nu$ for the r.m.s. e.m.f. E_ν contributed in a frequency range one cycle wide by a network whose resistance component at the frequency ν is R. T and k are the absolute temperature and the Boltzmann constant. Experimental data are available for the audible range and there the agreement between the formula and the data is good. It will be observed that neither the charge nor mass nor any other property of the carrier of electricity enters the formula explicitly. They enter indirectly through R. The formula above is based on the equipartition law. If the quantum distribution law is used, the expression becomes

$$E_\nu^2 d\nu = [4h\nu R/(e^{h\nu/kT} - 1)]d\nu.$$

The two expressions are indistinguishable in the range of the measurements.

Light Waves in Metals.[2] THORNTON C. FRY. When a wave of plane polarized light falls obliquely upon a conducting surface, it gives rise to a disturbance inside the conductor which has, among others, the following peculiarities:

(*a*) It is neither plane nor elliptically polarized, but belongs to a third distinct category;

(*b*) It does not travel with what is customarily called "the speed of light";

(*c*) Its velocity varies with the angle of incidence.

There are similar light waves in dielectrics and in free space.

Transatlantic Telephony.[3] F. B. JEWETT. This paper discusses in rather popular terms some of the outstanding problems which

[1] *Phys. Rev.*, Vol. 29, p. 614, April, 1927.

[2] *Opt. Soc. Amer. Jl.*, Vol. 14, p. 473, June, 1927.

[3] *Scientific Monthly*, August, 1927.

were met and solved in the course of the development of commercial transatlantic telephony. The discussion covers the use of single side band transmission, directive receiving antennæ and voice-operated relays which permit of two-way operation upon a single wave-length. The possibilities brought to light by the extended study of receiving conditions are also described.

Some Possibilities and Limitations in Common Frequency Broadcasting.[4] DeLoss K. Martin, Glenn D. Gillett, Isabel S. Bemis. Radio broadcast stations assigned to transmit on the same carrier frequency may cause audible beat notes to be produced when their signals are received simultaneously, due to the inaccuracies in the frequency adjustments of the transmitters. The radio broadcast transmission results that might be obtained from two or more stations transmitting on the same frequency with sufficient accuracy in frequency adjustment to eliminate audio-frequency beat notes are presented briefly in this paper.

Two cases are considered, the first case where there is a difference in frequency of a few cycles and the second case where the frequency of the carrier signal for all stations transmitting on the same frequency is determined by a common oscillator.

The results of preliminary experimental tests with the signals from a station in New York City and a station in Washington, D. C., are given.

[4] *Proceeding Institute of Radio Engineers*, Vol. 15, Number 3, p. 213, March, 1927

Contributors to this Issue

HERBERT E. IVES, B.S., University of Pennsylvania, 1905; Ph.D., Johns Hopkins, 1908; assistant and assistant physicist, Bureau of Standards, 1908–09; physicist, Nela Research Laboratory, Cleveland, 1909–12; physicist, United Gas Improvement Company, Philadelphia, 1912–18; U. S. Army Air Service, 1918–19; research engineer, Western Electric Company and Bell Telephone Laboratories, 1919 to date. Dr. Ives' work has had to do principally with the production, measurement and utilization of light.

FRANK GRAY, B.S., Purdue, 1911; instructor and graduate student in physics at the University of Wisconsin, Ph.D., 1916; member of the Naval Experimental Station during the war. Mr. Gray entered the Bell Telephone Laboratories—then the Engineering Department of the Western Electric Company—in 1919 and has been closely associated with Dr. Ives in his studies on light.

J. W. HORTON, B.S., Massachusetts Institute of Technology, 1914; instructor in physics, 1914–16; Engineering Department of the Western Electric Company and Bell Telephone Laboratories, 1916–. Mr. Horton has been closely connected with the development of apparatus for carrier current communication.

R. C. MATHES, B.Sc., University of Minnesota, 1912; E.E., 1913; Engineering Department of the Western Electric Company and Bell Telephone Laboratories, 1913–. Mr. Mathes has played an important part in the repeat development program and in the design of vacuum tube amplifiers for a wide variety of uses.

H. M. STOLLER, degree of E.E. from Union College, 1913; M.S. in electrical engineering, 1915; Engineering Department of the Western Electric Company and later Bell Telephone Laboratories, 1914 and 1916–. Most of Mr. Stoller's work has dealt with special problems connected with electrical power machinery, particularly voltage and speed regulators. He designed a multi-frequency generator which is now employed in the voice frequency carrier telegraph system.

E. R. MORTON, M.E., Stevens, 1917; Western Electric Company and Electric Test Lab., 1917–18; Air Service, U. S. Army, 1918; development and manufacturing control of color motion pictures, 1919; Harvard, 1920, electro-chemical investigations of nickel;

Fairchild Aerial Camera Corp., 1921–22; Western Electric Company and Bell Telephone Laboratories, 1923–. Since entering the Laboratories, Mr. Morton's work has related principally to vacuum tube regulators.

D. K. GANNETT, B.S. in engineering, 1916, University of Minnesota; E.E., 1917; American Telephone and Telegraph Company, Engineering Department, 1917–19, and Department of Development and Research, 1919–. During the war, Mr. Gannett was engaged in development work on amplifiers for submarine cable telegraphy. Since then, his work has been in connection with the transmission development problems of toll line signaling, vacuum tubes, telephotography and television.

E. I. GREEN, A.B., Westminster College (Fulton, Mo.), 1915; University of Chicago, 1915–16; professor of Greek, Westminster College, 1916–17; Captain, U. S. Army, 1917–19; B.S. in electrical engineering, Harvard University, 1921; Department of Development and Research, American Telephone and Telegraph Company, 1921–. Mr. Green has been engaged largely on work in connection with carrier transmission problems and development.

EDWARD L. NELSON, B.S. in electrical engineering, Armour Institute of Technology, 1914; Engineering Department of Western Electric Company, 1917; technical expert, Bureau of Engineering, Navy Department, 1917–18; Lieut. U. S. N. R. F., 1918–19; Engineering Department of Western Electric Company and Bell Telephone Laboratories, 1919–. Mr. Nelson has been closely connected with the development of radio apparatus, particularly transmitting equipment for ship-to-shore telephony and for broadcasting.

KARL K. DARROW, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D. in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company and Bell Telephone Laboratories, 1917–. Dr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

R. L. YOUNG, B.S. in electrical engineering, University of Pennsylvania, 1907; Westinghouse Electric and Manufacturing Company, 1907–09; Carnegie Institute of Technology, instructor of mathematics, 1909–10; McGraw Publishing Company, district manager Pittsburgh and assistant manager New York of *Metallurgical and Chemical Engineering*, and editorial board *Electrical World*, 1909–11; Engineer-

ing Department, American Telephone and Telegraph Company, 1911, and Department of Development and Research, 1911–. Mr. Young has been in charge of Power Development work since 1911.

S. B. Wright, M.E. in E.E., Cornell University, 1919; Department of Development and Research, American Telephone and Telegraph Company, 1919–. Mr. Wright has been engaged in transmission development work on long distance voice frequency telephone circuits.

H. C. Silent, E.E., Cornell University, 1921; Department of Development and Research, American Telephone and Telegraph Company, 1921–. Mr. Silent has been associated with the development work on voice-operated switching devices, and spent some time in England in connection with their application to the New York-London telephone circuit.

Walter A. Shewhart, A.B., University of Illinois, 1913; A.M., 1914; Ph.D., University of California, 1917; Engineering Department, Western Electric Company and Bell Telephone Laboratories, 1918–. Mr. Shewhart is making a special study of the application of probability theories to inspection engineering.

Index to Volume VI

A

Amplifiers, Application of Vacuum Tube, to Submarine Telegraph Cables, *Austen M. Curtis*, page 425.

Amplifiers, Modulation in Vacuum Tubes Used as, *Eugene Peterson* and *Herbert P. Evans*, page 442.

Analyzer for Complex Electric Waves, *A. G. Landeen*, page 230.

Analyzer for the Voice Frequency Range, *C. R. Moore* and *A. S. Curtis*, page 217.

Application of the Theory of Probability to Telephone Trunking Problems, *Edward C. Molina*, page 461.

Application of Vacuum Tube Amplifiers to Submarine Telegraph Cables, *Austen M. Curtis*, page 425.

Automatic Printing Equipment for Long Loaded Submarine Telegraph Cables, *A. A. Clokey*, page 402.

B

Bown, Ralph, Transatlantic Radio Telephony, page 248.

Bridge, Shielded, for Inductive Impedance Measurements at Speech and Carrier Frequencies, *W. J. Shackelton*, page 142.

Bridge, Measurement of Inductance by the Shielded Owen, *J. G. Ferguson*, page 375.

Broadcast Coverage of City Areas, Radio, *Lloyd Espenschied*, page 117.

C

Cables, Automatic Printing Equipment for Long Loaded Submarine Telegraph, *A. A, Clokey*, page 402.

Cables, Application of Vacuum Tube Amplifiers to Submarine Telegraph, *Austen M. Curtis*, page 425.

Cables, Determination of Electrical Characteristics of Loaded Submarine Telegraph, *J. J. Gilbert*, page 387.

Cables, Location of Opens in Toll Telephone, *P. G. Edwards* and *H. W. Herrington*, page 27.

Carson, John R., Electromagnetic Theory and the Foundations of Electric Circuit Theory, page 1.

Carson, John R., Propagation of Periodic Currents over a System of Parallel Wires. page 495.

Circuit Theory, Electromagnetic Theory and the Foundations of Electric, *John R. Carson*, page 1.

Circuit Theory, Propagation of Periodic Currents over a System of Parallel Wires, *John R. Carson* and *Ray S. Hoyt*, page 495.

Clokey, A. A., Automatic Printing Equipment for Long Loaded Submarine Telegraph Cables, page 402.

Contemporary Advances in Physics—XII, Radioactivity, *Karl K. Darrow*, page 55.

Contemporary Advances in Physics—XIII, Ferromagnetism, *Karl K. Darrow*, page 295.

Contemporary Advances in Physics—XIV, Introduction to Wave-Mechanics, *Karl K. Darrow*, page 653.

Copper Wire, Developments in the Manufacture of, *John R. Shea* and *Samuel McMullan*, page 187.

Crandall, Irving B., Dynamical Study of the Vowel Sounds, Part II, page 100.

Curtis, A. S., Analyzer for the Voice Frequency Range, page 217.

Curtis, Austen M., Application of Vacuum Tube Amplifiers to Submarine Telegraph Cables, page 425.

D

Darrow, Karl K., Contemporary Advances in Physics—XII, Radioactivity, page 55.

Darrow, Karl K., Contemporary Advances in Physics—XIII, Ferromagnetism, page 295.

Darrow, Karl K., Contemporary Advances in Physics—XIV, Introduction to Wave-Mechanics, page 653.

Davidson, J., Toll Switchboard No. 3, page 18.

Determination of Electrical Characteristics of Loaded Submarine Telegraph Cables, *J. J. Gilbert*, page 387.

Developments in the Manufacture of Copper Wire, *John R. Shea* and *Samuel McMullan*, page 187.

Dynamical Study of the Vowel Sounds, Part II, *Irving B. Crandall*, page 100.

E

Edwards, P. G., Location of Opens in Toll Telephone Cables, page 27.

Electric Circuit Theory, Electromagnetic Theory and the Foundations of, *John R. Carson*, page 1.

Electric Waves, Analyzer for Complex, *A. G. Landeen*, page 230.

Electromagnetic Theory and the Foundations of Electric Circuit Theory, *John R. Carson*, page 1.

Espenschied, Lloyd, Radio Broadcast Coverage of City Areas, page 117.

Evans, Herbert P., Modulation in Vacuum Tubes Used as Amplifiers, page 442.

F

Ferguson, J. G., Measurement of Inductance by the Shielded Owen Bridge, page 375.

Ferromagnetism, Contemporary Advances in Physics—XIII, *Karl K. Darrow*, page 295.

Filters, Study of the Regular Combination of Acoustic Elements, with Application to Recurrent Acoustic Filters, Tapered Acoustic Filters, and Horns, *W. P. Mason*, page 258.

G

Gannett, D. K., Wire Transmission System for Television, page 616.

Gilbert, J. J., Determination of Electrical Characteristics of Loaded Submarine Telegraph Cables, page 38 .

Gray, Frank, Production and Utilization of Television Signals, page 560.

Green, E. I., Wire Transmission System for Television, page 616.

H

Herrington, H. W., Location of Opens in Toll Telephone Cables, page 27.

Horns, A Study of the Regular Combination of Acoustic Elements, with Application to Recurrent Acoustic Filters, Tapered Acoustic Filters, and Horns, *W. P. Mason*, page 258.

Horton, J. W., Production and Utilization of Television Signals, page 560.

Hoyt, Ray S., Propagation of Periodic Currents over a System of Parallel Wires, page 495.

I

Inductance, Measurement of, by the Shielded Owen Bridge, *J. G. Ferguson*, page 375.

Inductive Impedance Measurements at Speech and Carrier Frequencies, Shielded Bridge for, *W. J. Shackelton*, page 142.

Inspection, Quality Control, *W. A. Shewhart*, page 722.
Ives, Herbert E., Television, page 551.

L

Landeen, A. G., Analyzer for Complex Electric Waves, page 230.
Loaded Submarine Telegraph Cables, Automatic Printing Equipment for Long, *A. A. Clokey*, page 402.
Loaded Submarine Telegraph Cables, Determination of Electrical Characteristics of, *J. J. Gilbert*, page 387.
Location of Opens in Toll Telephone Cables, *P. G. Edwards* and *H. W. Herrington*, page 27.

M

Mason, W. P., Study of the Regular Combination of Acoustic Elements, with Application to Recurrent Acoustic Filters, Tapered Acoustic Filters, and Horns, page 258.
Mathes, R. C., Production and Utilization of Television Signals, page 560.
McMullan, Samuel, Developments in the Manufacture of Copper Wire, page 187.
Measurement of Inductance by the Shielded Owen Bridge, *J. G. Ferguson*, page 375.
Measurements at Speech and Carrier Frequencies, Shielded Bridge for Inductive Impedance, *W. J. Shackelton*, page 142.
Modulation in Vacuum Tubes Used as Amplifiers, *Eugene Peterson* and *Herbert P. Evans*, page 442.
Molina, Edward C., Application of the Theory of Probability to Telephone Trunking Problems, page 461.
Moore, C. R., Analyzer for the Voice Frequency Range, page 217.
Morton, E. R., Synchronization of Television, page 604.

N

Nelson, Edward L., Radio Transmission System for Television, page 633.
New York-London Telephone Circuit, *S. B. Wright* and *H. C. Silent*, page 736.

O

Offices, Power Plants for Telephone, *R. L. Young*, page 702.

P

Peterson, Eugene, Modulation in Vacuum Tubes Used as Amplifiers, page 442.
Physics, Contemporary Advances in—XII, Radioactivity, *Karl K. Darrow*, page 55.
Physics, Contemporary Advances in—XIII, Ferromagnetism, *Karl K. Darrow*, page 295.
Physics, Contemporary Advances in—XIV, Introduction to Wave-Mechanics, *Karl K. Darrow*, page 653.
Power Plants for Telephone Offices, *R. L. Young*, page 702.
Printing Equipment for Long Loaded Submarine Telegraph Cables, Automatic, *A. A. Clokey*, page 402.
Probability, Application of the Theory of, to Telephone Trunking Problems, *Edward C. Molina*, page 461.
Production and Utilization of Television Signals, *Frank Gray, J. W. Horton* and *R. C. Mathes*, page 560.
Propagation of Periodic Currents over a System of Parallel Wires, *John R. Carson* and *Ray S. Hoyt*, page 495.

Q

Quality Control, *W. A. Shewhart*, page 722.

R

Radio, New York-London Telephone Circuit, *S. B. Wright* and *H. C. Silent*, page 736.

Radioactivity, Contemporary Advances in Physics—XII, *Karl K. Darrow*, page 55.

Radio Broadcast Coverage of City Areas, *Lloyd Espenschied*, page 117.

Radio Telephony, Transatlantic, *Ralph Bown*, page 248.

Radio Transmission System for Television, *Edward L. Nelson*, page 633.

S

Shackelton, W. J., Shielded Bridge for Inductive Impedance Measurements at Speech and Carrier Frequencies, page 142.

Shea, John R., Developments in the Manufacture of Copper Wire, page 187.

Shewhart, W. A., Quality Control, page 722.

Shielded Bridge for Inductive Impedance Measurements at Speech and Carrier Frequencies, *W. J. Shackelton*, page 142.

Silent, H. C., New York-London Telephone Circuit, page 736.

Speech, Dynamical Study of the Vowel Sounds, Part II, *Irving B. Crandall*, page 100.

Stoller, H. M., Synchronization of Television, page 604.

Study of the Regular Combination of Acoustic Elements, with Application to Recurrent Acoustic Filters, Tapered Acoustic Filters, and Horns, *W. P. Mason*, page 258.

Submarine Telegraph Cables, Determination of Electrical Characteristics of Loaded, *J. J. Gilbert*, page 387.

Submarine Telegraph Cables, Automatic Printing Equipment for Long Loaded, *A. A. Clokey*, page 402.

Submarine Telegraph Cables, Application of Vacuum Tube Amplifiers to, *Austen M. Curtis*, page 425.

Switchboard No. 3, Toll, *J. Davidson*, page 18.

Synchronization of Television, *H. M. Stoller* and *E. R. Morton*, page 604.

T

Telephone Offices, Power Plants for, *R. L. Young*, page 702.

Telephony, Transatlantic Radio, *Ralph Bown*, page 248.

Television, Wire Transmission System for, *D. K. Gannett* and *E. I. Green*, page 616.

Television Signals, Production and Utilization of, *Frank Gray, J. W. Horton* and *R. C. Mathes*, page 560.

Television, *Herbert E. Ives*, page 551.

Television, Radio Transmission System for, *Edward L. Nelson*, page 633.

Television, Synchronization of, *H. M. Stoller* and *E. R. Morton*, page 604.

Toll Switchboard No. 3, *J. Davidson*, page 18.

Toll Telephone Cables, Location of Opens in, *P. G. Edwards* and *H. W. Herrington*, page 27.

Transatlantic Telephony, New York-London Telephone Circuit, *S. B. Wright* and *H. C. Silent*, page 736.

Transatlantic Radio Telephony, *Ralph Bown*, page 248.

Trunking Problems, Application of the Theory of Probability to Telephone, *Edward C. Molina*, page 461.

Tubes Used as Amplifiers, Modulation in Vacuum, *Eugene Peterson* and *Herbert P. Evans*, page 442.

V

Vacuum Tubes Used as Amplifiers, Modulation in, *Eugene Peterson* and *Herbert P. Evans*, page 442.

Voice Frequency Range, Analyzer for the, *C. R. Moore* and *A. S. Curtis*, page 217.

Vowel Sounds, Part II, Dynamical Study of the, *Irving B. Crandall*, page 100.

W

Waves, Analyzer for Complex Electric, *A. G. Landeen*, page 230.

Waves, Analyzer for the Voice Frequency Range, *C. R. Moore* and *A. S. Curtis*, page 217.

Wave-Mechanics, Introduction to, Contemporary Advances in Physics—XIV, *Karl K. Darrow*, page 653.

Wire, Developments in the Manufacture of Copper, *John R. Shea* and *Samuel McMullan*, page 187.

Wires, Propagation of Periodic Currents over a System of Parallel, *John R. Carson* and *Ray S. Hoyt*, page 495.

Wire Transmission System for Television, *D. K. Gannett* and *E. I. Green*, page 616.

Wright, S. B., New York-London Telephone Circuit, page 736.

Y

Young, R. L., Power Plants for Telephone Offices, page 702.