The Bell System Technical Journal

January, 1928

The Measurement of Acoustic Impedance and the Absorption Coefficient of Porous Materials

By E. C. WENTE and E. H. BEDELL

SYNOPSIS: Various ways of determining the acoustic impedance and the absorption coefficient of porous materials from measurements on the standing waves in tubes are discussed. In all cases the material under investigation is placed at one end of the tube and the sound is introduced at the other end. Values of the coefficient of absorption of a number of commonly used damping materials as obtained by one of the methods are given. Several types of built-up structures are shown to have a greater absorption coefficient for low frequency sound waves than is conveniently obtainable by a single layer of material.

THE most commonly used method of determining the sound absorption coefficient of a material is that devised by the late Professor W. C. Sabine. In this method the reverberation time of a room is measured before and after the introduction of a definite amount of the material. This method has the great merit that the values so determined usually apply to the materials precisely as they are ordinarily used in rooms for damping purposes. However, it is tedious and requires a very quiet room and large samples of the materials. A simpler scheme has been devised by H. O. Taylor,[1] in which the absorbing material is placed at one end of a tube. The coefficient of absorption is determined from a measurement of the ratio of maximum to minimum pressures of the standing waves within the tube when sound is introduced at the open end. Thus only a small sample of the material is required and with suitable apparatus the measurements can be made with great facility. In this paper several modifications of Taylor's tube method are discussed; in addition, it is shown that by a similar method it is possible to determine not only the absorption coefficient but also the acoustic impedance, a quantity which is playing an important part in present day applied acoustics.

GENERAL THEORY

Consider a tube of length l, which is filled with a medium having a propagation constant $P = \alpha + i\beta$ and a characteristic acoustic im-

[1] *Phys. Rev.*, II, 1913, p. 270.

pedance [2] equal to Z_0 per unit area. At one end, O, let the velocity be uniform over the whole cross-section and equal to $\dot{\xi}_1 e^{i\omega t}$. At a distance l from O let the tube be terminated by the material which is to be investigated, and the acoustic impedance of which may be rep-

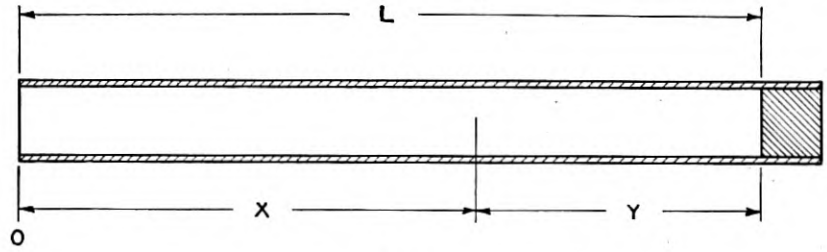


Fig. 1

resented by $Z_2 = R_2 + iX_2$ per unit area. Under these conditions, the pressure, p, at any point in the tube at a distance x from O, is by analogy with the electrical transmission line

$$p = \dot{\xi}_1 e^{i\omega t} \left[\frac{Z_2 \cosh Pl + Z_0 \sinh Pl}{Z_0 \cosh Pl + Z_2 \sinh Pl} \cosh Px - \sinh Px \right] Z_0.^{3} \quad (1)$$

If there is no attenuation along the tube, we get, on dropping the time factor,

$$p = R\dot{\xi}_1 \left[\frac{Z_2 \cos \beta l + iR \sin \beta l}{R \cos \beta l + iZ_2 \sin \beta l} \cos \beta x - i \sin \beta x \right], \quad (2)$$

where $R = c\rho$, the product of the velocity of propagation along the tube and the density of the medium, and

$$\beta = \frac{2\pi f}{c}.$$

Equation (2) indicates numerous possible ways of determining Z_2, e.g., from the values of $\dot{\xi}_1$ and of p at any point in the tube; from the pressures for two values of either x or l, if $\dot{\xi}_1$ is constant; from the pressures at any point in the tube for the unknown and for a known value of Z_2; from the magnitude of p as a function of either x or l. However, we shall confine our discussion to three methods, which appear to be most practicable.

[2] The term acoustic impedance as here used may be defined as the ratio of pressure to volume velocity; the characteristic impedance is this impedance if the tube were of infinite length.

[3] J. A. Fleming, "Propagation of Electric Currents in Telephone and Telegraph Conductors," page 98; 3d Ed.

(a) *Pressure Measured at Two Points in the Tube*

It has already been pointed out that the impedance Z_2 can be determined if the relative phase and magnitude of the pressures at any two points in the tube are known. However, from the standpoint of convenience and precision it appears best to measure the pressures at the reflecting surface and at a point a quarter of a wave-length away. We then have at the reflecting surface $x = l$ and

$$p_2 = R\xi_1 \left[\frac{R_2 + iX_2}{R \cos \beta l + iZ_2 \sin \beta l} \right],$$

and for the point $x = l - \frac{\lambda}{2} = l - \frac{\pi}{2\beta}$,

$$p_1 = R\xi_1 \left[\frac{iR}{R \cos \beta l + iZ_2 \sin \beta l} \right],$$

so that

$$\frac{p_2}{p_1} = \frac{X_2 - iR_2}{R} \equiv Ae^{i\varphi}.$$

Hence

$$\left.\begin{aligned} R_2 &= -AR \sin \varphi, \\ X_2 &= AR \cos \varphi, \\ |Z_2| &= AR. \end{aligned}\right\} \tag{3}$$

If the coefficient of reflection is expressed as [4]

$$Ce^{i\Psi} = \frac{Z_2 - R}{Z_2 + R}, \tag{4}$$

we get

$$C = \left[\frac{1 + 2A \sin \varphi + A^2}{1 - 2A \sin \varphi + A^2} \right]^{1/2},$$

where

$$\varphi = \tan^{-1} \frac{2A \cos \varphi}{A^2 + 1}. \tag{5}$$

The absorption coefficient, which is generally defined as the ratio of absorbed to incident power, is equal to $1 - |C|^2$.

(b) *Tube of Constant Length; the Absolute Value of the Pressure Measured at Points along the Tube*

The method discussed under this section is that adopted by H. O. Taylor for measuring the absorption coefficient of porous materials.

[4] I. B. Crandall, "Theory of Vibrating Systems and Sound," page 168.

For the absolute value of the pressure at any point in the tube we get from equation (2)

$$|p| = \left[\frac{R_2^2 + X_2^2 + R^2 + (R_2^2 + X_2^2 - R^2)\cos 2\beta y + 2X_2R \sin 2\beta y}{R_2^2 + X_2^2 + R^2 - (R_2^2 + X_2^2 - R^2)\cos 2\beta l - 2X_2R \sin 2\beta l}\right]^{1/2} R\xi_1, \quad (6)$$

where $y = l - x$.

$|p|$ has maximum or minimum values when

$$\tan 2\beta y = \frac{2X_2R}{X_2^2 + R_2^2 - R^2}; \quad (7)$$

for the maximum value $2\beta y$ lies in the first and for the minimum, in the third quadrant. We therefore get

$$\frac{|p|_{\max}}{|p|_{\min}} = \left[\frac{X_2^2+R_2^2+R^2+\sqrt{(X_2^2+R_2^2-R^2)^2+4X_2^2R^2}}{X_2^2+R_2^2+R^2-\sqrt{(X_2^2+R_2^2-R^2)^2+4X_2^2R^2}}\right]^{1/2} \equiv A. \quad (8)$$

Let y_1 be the value of y for which the pressure is a maximum; we then have from (7) and (8) and (4)

$$R_2 = \frac{2AR}{(A^2 + 1) - (A^2 - 1)\cos 2\beta y_1}, \quad (9)$$

$$X_2 = \frac{R(A^2 - 1)\sin 2\beta y_1}{(A^2 + 1) - (A^2 - 1)\cos 2\beta y}, \quad (10)$$

$$C_1 = \frac{A - 1}{A + 1}, \quad (11)$$

$$\Psi = 2\beta y_1.$$

The relation (11) can be derived more simply on the classical theory, as it was done by H. O. Taylor. A derivation of (11) is given by Eckhardt and Chrisler,[5] which differs from that of H. O. Taylor. From their derivation it would appear that for (11) to be valid the length of the tube should be adjusted for resonance and that the change in phase at the reflecting surface should be small. The derivation here given shows that (11) is general; it implies only that the waves be plane and that there be no dissipation of power along the tube.

[5] Scientific Paper of the Bureau of Standards, No. 526, page 56.

(*c*) *Tube of Variable Length. Pressure Measured at the Source*

The absolute value of the pressure at the driving end of the tube according to (2) is

$$|p_1| = \left[\frac{R_2^2 + X_2^2 + R^2 + (R_2^2 + X_2^2 - R^2)\cos 2\beta l + 2X_2R \sin 2\beta l}{R_2^2 + X_2^2 + R^2 + (R_2^2 + X_2^2 - R^2)\cos 2\beta l - 2X_2R \sin 2\beta l}\right]^{1/2} R\xi_1$$

and $|p_1|$ is a maximum or a minimum when

$$\tan 2\beta l = \frac{2X_2R}{R_2^2 + X_2^2 - R^2}.$$

For the maximum value $2\beta l$ lies in the first and for the minimum, in the third quadrant. We therefore have

$$\frac{|p_1|_{\max}}{|p_1|_{\min}} = \frac{X_2^2 + R_2^2 + R^2 + \sqrt{(X_2^2 + R_2^2 - R^2)^2 + 4X_2^2R^2}}{X_2^2 + R_2^2 + R^2 - \sqrt{(X_2^2 + R_2^2 - R^2)^2 + 4X_2^2R^2}} \equiv A.$$

By analogy from the equations derived in section (*b*) above, we see that

$$R_2 = \frac{2\sqrt{A}R}{(A+1) - (A-1)\cos 2\beta l_1},$$

$$X_2 = \frac{R(A-1)\sin 2\beta l_1}{(A+1) - (A-1)\cos 2\beta l_1},$$

$$C = \frac{\sqrt{A} - 1}{\sqrt{A} + 1},$$

$$\Psi = 2\beta l_1,$$

where l_1 is the length of the tube when p_1 has a maximum value.

Discussion of the Precision of the Methods

Of the three methods of measuring impedance discussed above, the first is undoubtedly the simplest and most convenient, if an a.c. potentiometer is available. Theoretically, in this case the impedance may be determined with a high degree of precision. However, the method presupposes that the points where the pressures are measured are exactly a quarter of a wave-length apart; a more detailed analysis shows that, if A is small, variations in this distance will have a large effect on both the ratio of the pressures and their phase difference. It therefore is necessary to keep the temperature of the tube accurately constant or else to determine the distance corresponding to a quarter

of a wave-length before each measurement. A precise determination of the point a quarter of a wave-length from the reflecting surface may be made by placing a smooth metal block at the reflecting end and finding then the position in the tube at which the pressure is a minimum.

In the other two methods it is relatively less important that the temperature be maintained constant, for the ratio of pressures is affected very little by any temperature variations. In the third method, where the length of the tube is varied, the expressions for R_2 and X_2 are the same as in (*b*), except that in place of the ratio of pressures they involve the square root of this ratio. For small values of pressure ratios the precision is therefore somewhat greater. However, for high values of reflection the ratio becomes very large and great care is required in the experimental set up to prevent errors creeping into the measurements through extraneous vibrations and stray electromotive forces in the measuring circuit. The main advantage of the method in which the pressure at the source only is measured is that a short length of exploring tube is required. If measurements down to a frequency of 60 cycles are made, the tube length must be at least 8 feet. An exploring tube reaching the whole length would ordinarily introduce too much attenuation if it were of sufficiently small bore to prevent resonance effects at the lower frequencies.

Experimental Procedure

In the case of the experimental results here reported the measurements were all made by the method outlined in section (*c*), i.e., the pressures were measured at the source while the length of the tube was varied. The experimental set up is shown in Fig. 2. A piece of Shelby

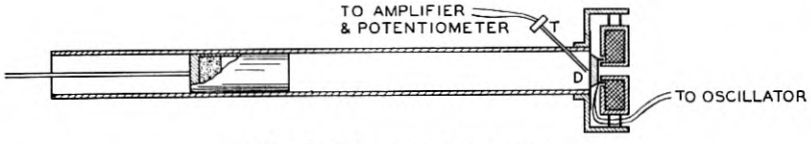


Fig. 2—Diagram of apparatus

steel tubing, 9 feet long, of 3″ internal diameter, and with 1/4″ wall, was fitted with a piston carrying the absorbing material. This piston was made up of a brass tube one foot long with a wall 1/64″ thick, the far end of which was closed with a one-inch brass block. To insure the propagation of plane waves and a constant velocity at the source, the diaphragm at *D* had a diameter of 2⅞″, and a mass of about 100 grams. This was driven with a coil 2″ in diameter situated in a radial magnetic field. The annular gap between the edge of the diaphragm

and the interior of the tube was closed by a flexible piece of leather. To prevent vibrations of the magnet from getting to the tube, the magnet was held in position by flexible supports. The exploring tube t was about 5″ long with a 1/16″ bore which led to the transmitter, T. The voltages generated by the transmitter were measured with an amplifier and an a.c. potentiometer. The potentiometer was used because with it small voltages can be measured and errors due to harmonics are avoided. The proper functioning of the apparatus was determined by measuring the coefficient of reflection with no absorbing material in the piston. Theoretically the reflection should then be practically 100 per cent. The pressure ratios that were actually observed were of the order of 12,000 which corresponds to a reflection coefficient of 98 per cent. Evidently some extraneous pressures or voltages were still present. However, no attempt was made to reduce these further as the materials tested had a reflection coefficient considerably less than this value.

Experimental Results

A brief study was made of the absorption of hair felt, as there is an appreciable variation in the data given by various investigators on the absorption frequency characteristic of felts of presumably the same

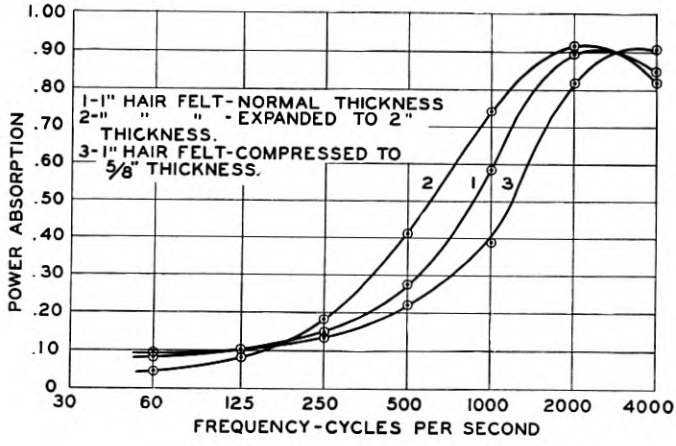


Fig. 3—Power absorbed by hair felt

type. After measurements on several samples it was evident that concordant results could not be expected as the absorption varied considerably with the packing of the felt. This point is illustrated by the curves shown in Fig. 3. These curves were all obtained on the same

piece of hair felt but with different degrees of packing. It is thus evident that a felt which has become loosened by handling may have an absorption frequency characteristic quite unlike that of a new piece.

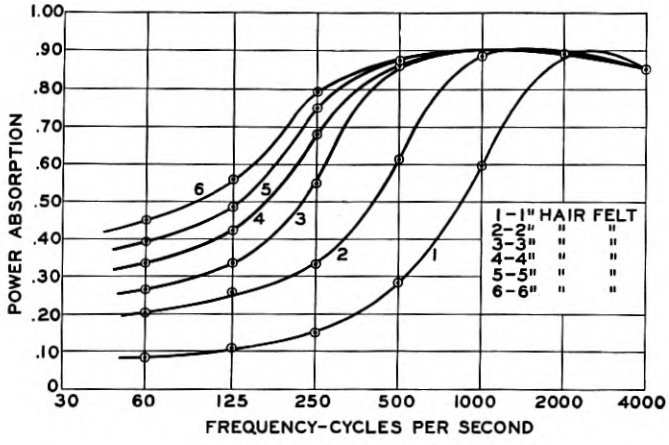


Fig. 4—Power absorbed by hair felt

In Fig. 4 are given the absorption coefficients for various thicknesses of hair felt. These values are in general agreement with those obtained by the reverberation method according to published results. Exact agreement is not to be expected, for the values here given apply only to sound waves having a perpendicular incidence on materials solidly backed by a hard surface. When the materials are applied in a room, the support is often more flexible and the absorption is partly due to inelastic bending. However, the agreement between the sets of values is sufficiently good to show that the results obtained by the simpler tube method may be used to get a good approximation to the values of absorption of the materials when applied in rooms for damping purposes.

Measurements have been made on a large number of porous materials. Although most of these materials are very good absorbers at the higher frequencies, none of them were found to be very efficient in the lower frequency region. Uniform absorption over most of the frequency range was found only in materials which are relatively inefficient absorbers. High absorption at the lower frequencies was obtained only when the thickness of the material was greatly increased. This fact is typically illustrated by the curves of absorption for hair felt given in Fig. 4.

When a sound wave of low frequency is reflected from a wall covered

TABLE I

ABSORPTION COEFFICIENTS FOR VARIOUS FREQUENCIES

FREQUENCY C.P.S.

	60	125	250	500	1000	2000	4000
1″ Acoustic tile	.07	.08	.11	.18	.48	.76	.47
½″ Asbestos hair felt	.08	.10	.15	.24	.49	.84	.66
Axminster rug	.07	.11	.14	.20	.33	.52	.82
Felted wood fibre	.08	.12	.18	.33	.67	.92	.91
½″ Building board	.13	.14	.15	.17	.20	.26	.29
Flax wool	.07	.09	.18	.48	.73	.50	.33
Structure No. 1	.12	.18	.36	.71	.79	.82	.85
" " 2	.16	.24	.46	.77	.92	.89	.85
" " 3	.23	.37	.62	.88	.91	.78	.84
" " 4	.19	.28	.51	.81	.92	.90	.84
" " 5	.15	.25	.44	.75	.77	.71	.80
" " 6	.22	.41	.87	.74	.81	.59	.83
" " 7	.17	.39	.82	.94	.92	.91	.85
" " 8	.24	.39	.83	.82	.64	.59	.80
" " 9	.22	.37	.79	.91	.82	.89	.86
" " 10	.30	.55	.92	.69	.83	.86	.86

Structure No. 1	Fiber building board —no air space—felt
" " 2	Fiber building board —1″ air space—felt
" " 3	Fiber building board —2″ air space—felt
" " 4	Fiber building board —no air space—2″ felt
" " 5	Fiber building board —1″ air space—Fiber building board
" " 6	Fiber building board —1″ air space— Fiber building board—1″ air space—Fiber building board
" " 7	Fiber building board —1″ air space—felt—1″ air space—felt
" " 8	Fiber building board —1″ air space—felt—1″ air space—Fiber building board
" " 9	Fiber building board —1″ air space— Fiber building board—1″ air space—felt
" " 10	Fiber building board —1″ air space— Fiber building board—1″ air space—Fiber building board—felt

with a porous material, the velocity of the air particles near the reflecting surface is small and hence there can be but little absorption. We may look at the phenomenon of reflection in still another way. In order to have a small coefficient of reflection the mechanical impedance of the wall per unit area should, as nearly as possible, be equal to the acoustic impedance of the air per unit area. The reason for the high reflection at low frequencies by a rigid wall covered with a porous material lies in its high stiffness reactance. At a given frequency this reactance can be compensated by loading the air near the reflecting surface. This may be accomplished in various ways. One of these ways is to place at a short distance from the wall a second wall which is porous or perforated. This arrangement has the effect of covering the wall with a multiplicity of resonators, which may be given any desired resonance frequency by properly proportioning the size, length and number of perforations and the spacing of the walls. The surface of the walls forming the air space should be absorbing or else the space should be provided with absorbing material.

To get a wider absorption band two or more perforated walls with proper spacing may be used, as this arrangement is equivalent to an aggregate of multiple resonators. The values of absorption coefficients of a number of structures of this type are given in the accompanying table. The measurements refer to sound which is incident from right to left as the structures are given in the table. The building board referred to in the table is a commercial type of insulating-board one inch thick with 400 1/4 inch by 3/4 inch holes per square foot. The felt in all cases is one-inch hair felt. These values show that relatively high absorption may be obtained at low as well as at high frequencies without an excessive amount of absorbing material. The use of combinations of absorbing materials, such as are given in the table, offers the advantage that more uniform damping at all frequencies can be obtained, and the degree of damping can be readily controlled by covering the proper area of surface. These two factors have become increasingly important in studio and auditorium design, with improved technique in recording and reproducing speech and music.

The Rigorous and Approximate Theories of Electrical Transmission Along Wires

By JOHN R. CARSON

THE theory of electrical transmission along straight parallel guiding conductors is of fundamental importance to the communication engineer. In its original, and largely in its present day form, it involves only relatively simple concepts which go back to the early work of Kelvin and Heaviside. In accordance with these concepts the transmission phenomena are completely determined by the self and mutual impedances of the conductors and the self and mutual capacities (together with the dielectric leakage). As a consequence, the phenomena are completely expressed in terms of the propagation constants and corresponding characteristic impedances of the possible modes of propagation deducible from these underlying concepts.

The elementary theory sketched above is of beautiful simplicity and great value. It is, however, admittedly approximate, and in two respects is not altogether adequate. Its first defect is that it represents the transmission phenomena correctly only at some distance from the physical terminals of the system or at some distance from points of discontinuity. This defect is ordinarily of small practical significance when the conductors all consist of wires of small cross section. When, however, conductors of large cross sections, or the ground, form part of the transmission system, the elementary theory may be quite inadequate. The theoretical questions here involved were briefly discussed by the writer in a previous paper.[1] The mathematics involved in this problem are extremely complicated and the further work of the writer has not as yet been carried to a point which justifies publication.

With the extension of transmission theory discussed in the preceding paragraph the present paper has no concern, and it is to be expressly understood that we are dealing with the transmission phenomena at a sufficient distance from the physical terminals, such that the "end effects" are negligible. The problems here dealt with may be stated as follows: First to investigate the conditions under which the specification of the system by means of its self and mutual impedances is valid and secondly to provide a general method for calculating these circuit parameters from the geometry and electrical constants of the system.

[1] "The Guided and Radiated Energy in Wire Transmission." Trans. A. I. E. E., 1924.

As regards the first phase of this problem it is found that the complete specification of the system in terms of its self and mutual impedances and capacities is only rigorously valid for the ideal case of perfect conductors embedded in a perfect dielectric, and that it becomes quite invalid if either the conductors or the dielectric are too imperfect. Fortunately, however, it is valid to a high degree of approximation for all systems which could be employed for the *efficient* transmission of electrical energy.

Under the circumstances where the approximations discussed in the preceding paragraph are valid it is shown that the electric and magnetic field in both dielectric and conductors are derivable from two wave functions. The first of these is determined as a linear function of the conductor charges by the solution of a well-known two-dimensional potential problem, while the second is determined as a linear function of the conductor currents by the solution of a generalization of the two-dimensional potential problem. The latter problem is believed to be novel, in its general form, and to possess both practical and mathematical interest. For detailed application of the theory to specific problems, the following papers may be consulted.

"Wave Propagation over Parallel Wires: The Proximity Effect." *Phil. Mag.*, April 1921.

"Transmission Characteristics of the Submarine Cable." *Jour. Frank. Inst.*, Dec. 1921.

"Wave Propagation in Overhead Wires with Ground Return." *B. S. T. J.*, Oct. 1926.

I

Maxwell's equations are the set of partial differential equations which formulate the relations between the electric intensity E and the magnetic intensity H in terms of the frequency $\omega/2\pi$ and the electrical constants of the medium. Let λ, μ and k denote the conductivity, permeability and dielectric constant of the medium; let it be supposed that all quantities vary with the time t as $e^{i\omega t}$, and let

$$v = 1/\sqrt{k\mu},$$
$$\nu^2 = 4\pi\lambda\mu i\omega - \omega^2/v^2,$$
$$i = \sqrt{-1}.$$

Then if we introduce the vector

$$M = \mu i\omega \cdot H,$$

Maxwell's equations for a *continuous homogeneous* medium may be written in the compact form [2]

$$\begin{aligned} \text{curl } E &= -M, \\ \text{curl } M &= \nu^2 E, \\ \text{div } E &= 0, \\ \text{div } M &= 0. \end{aligned} \tag{1}$$

From this set of equations it is easily shown that each component of the vectors E and M individually satisfies the *wave equation*

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \nu^2\right) f = 0 \tag{2}$$

or in vector notation

$$(\nabla^2 - \nu^2) f = 0.$$

Here f denotes any vector component; thus in Cartesian coordinates f may stand for E_x, E_y, E_z; M_x, M_y, M_z, all of which separately satisfy (2).

Given the electrical constants and geometry of the conducting system and dielectric media, the general problem is to find solutions of (1) and (2) which also satisfy the *boundary conditions* at the surfaces of separation of the different media. These boundary conditions are that the tangential components of E and H shall be continuous over such surfaces of separation. These boundary conditions, as may be seen from (1), necessitate also the continuity of the *normal* components of M and $(\nu^2/\mu)E$.

If we introduce a vector potential $A(A_x, A_y, A_z)$ and a scalar potential Φ, it is easily shown that (1) may be replaced by

$$\begin{aligned} M &= \text{curl } A, \\ E &= -A - \text{grad } \Phi, \end{aligned} \tag{3}$$

with the further relation

$$\text{div } A + \nu^2 \Phi = 0. \tag{4}$$

Φ and the components of the vector A individually satisfy the wave equation; thus

$$\begin{aligned} (\nabla^2 - \nu^2)\Phi &= 0, \\ (\nabla^2 - \nu^2) A &= 0. \end{aligned} \tag{5}$$

In Cartesian coordinates these equations are

[2] Note that in this form the constants of the medium appear explicitly only through the parameter ν^2.

$$
\begin{aligned}
M_x &= \frac{\partial}{\partial y} A_z - \frac{\partial}{\partial z} A_y, \\
M_y &= \frac{\partial}{\partial z} A_x - \frac{\partial}{\partial x} A_z, \\
M_z &= \frac{\partial}{\partial x} A_y - \frac{\partial}{\partial y} A_x, \\
E_x &= -A_x - \frac{\partial}{\partial x}\Phi, \\
E_y &= -A_y - \frac{\partial}{\partial y}\Phi, \\
E_z &= -A_z - \frac{\partial}{\partial z}\Phi,
\end{aligned}
\tag{6}
$$

and

$$
\frac{\partial}{\partial x} A_x + \frac{\partial}{\partial y} A_y + \frac{\partial}{\partial z} A_z + \nu^2\Phi = 0. \tag{7}
$$

In technical transmission problems we are largely concerned with propagation along a uniform transmission system, composed of straight parallel conductors. That is to say, the transmission system does not vary geometrically or in its electrical constants along the axis of transmission, taken as the axis of Z. It is known that in such transmission systems *exponentially* [3] propagated waves exist. We therefore modify the general equations by assuming that the wave (and all vector components) vary with t and z as $\exp(i\omega t - \gamma z)$, γ being entitled the *propagation constant.* As a consequence of this assumption it is easily shown that the vectors E and M are derivable from the wave functions F, Φ, Θ as follows:

$$
\begin{aligned}
M_x &= \frac{\partial}{\partial y} F - \gamma \frac{\partial}{\partial x}\Theta, \\
M_y &= -\frac{\partial}{\partial x} F - \gamma \frac{\partial}{\partial y}\Theta, \\
M_z &= -(\nu^2 - \gamma^2)\Theta, \\
E_x &= -\frac{\partial}{\partial x}\Phi - \frac{\partial}{\partial y}\Theta, \\
E_y &= -\frac{\partial}{\partial y}\Phi + \frac{\partial}{\partial x}\Theta, \\
E_z &= -\frac{\partial}{\partial z}\Phi - F = \gamma\Phi - F.
\end{aligned}
\tag{8}
$$

The *wave functions* F and Φ are not independent but are connected by the relation

$$
\nu^2\Phi = \gamma F. \tag{9}
$$

[3] This means that the wave involves the axial coordinate z only exponentially.

Another useful formulation of the field equations equivalent to and directly deducible from (8) is

$$\begin{aligned}
(\nu^2 - \gamma^2)M_x &= -\nu^2 \frac{\partial}{\partial y} E_z + \gamma \frac{\partial}{\partial x} M_z, \\
(\nu^2 - \gamma^2)M_y &= \nu^2 \frac{\partial}{\partial x} E_z + \gamma \frac{\partial}{\partial y} M_z, \\
(\nu^2 - \gamma^2)E_x &= \gamma \frac{\partial}{\partial x} E_z + \frac{\partial}{\partial y} M_z, \\
(\nu^2 - \gamma^2)E_y &= \gamma \frac{\partial}{\partial y} E_z - \frac{\partial}{\partial x} M_z.
\end{aligned} \tag{10}$$

In this formulation the problem is reduced to the determination of the *wave functions* E_z and M_z, and the *propagation constant* γ.

It will be observed that, by virtue of the assumption that the wave functions of (8), (9) and (10) involve t and z only through the common factor $\exp(i\omega t - \gamma z)$, we can write

$$\begin{aligned}
F &= f(x, y) \cdot \exp(i\omega t - \gamma z), \\
\Phi &= \phi(x, y) \cdot \exp(i\omega t - \gamma z), \\
E &= e(x, y) \cdot \exp(i\omega t - \gamma z), \text{ etc.},
\end{aligned} \tag{11}$$

where f, ϕ, e, etc., are two-dimensional functions of x and y alone, and satisfy the two-dimensional wave equations

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) f = (\nu^2 - \gamma^2) f, \text{ etc.} \tag{12}$$

In the following, therefore, we shall regard the wave functions F, Φ, E, etc., as two-dimensional functions with the understanding that the common factor $\exp(i\omega t - \gamma z)$ is omitted for convenience.

II

Before taking up the discussion of the general problem in the light of equations (8) and (10) we shall first consider a type of plane wave propagation to which the transmission phenomena closely approximate in an efficient transmission system. We consider the ideal transmission system composed of any number of straight parallel *perfectly conducting* conductors imbedded in a *perfect* dielectric. For such a system we *assume* the possibility of plane wave propagation by supposing that E_z and M_z are everywhere zero. By virtue of the assumption of perfect conductivity, the electric force must vanish inside the conductors, and at the surface the tangential component

must vanish. *In the dielectric* reference to equations (10) shows that if $E_z = M_z = 0$, a finite solution requires that

$$\nu^2 - \gamma^2 = 0$$

or, since $\lambda = 0$ in the dielectric,

$$\gamma = i\omega/v.$$

That is to say, the plane wave is propagated with the velocity of light v, without attenuation.

Reference to equations (8) and (9) shows that the boundary conditions can be satisfied by setting $\Theta = 0$, writing

$$\Phi = \phi \cdot \exp(i\omega t - i\omega z/v),$$

and determining the function ϕ which satisfies Laplace's equation in two dimensions,

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\phi = 0,$$

and is constant over the cross section of the conductors.

From the relation $F = (i\omega/v)\Phi$ it is also easily shown that the electric and magnetic forces are both in planes normal to Z and that these vectors are normal to each other and in time phase. The flow of energy is therefore parallel to the Z-axis everywhere. We therefore have a pure plane guided wave of unit power factor; the ideal for the electrical transmission of energy.

III

We now take up the much more complicated problem arising when the conductivity λ of the conductors is finite and when the dielectric media themselves may be dissipative. In attacking this general problem we shall be guided throughout by the fact that the wave solution we are seeking must approximate, more or less closely, to the ideal plane wave [4] if the system is to efficiently transmit electrical energy. We shall therefore introduce *ab initio* approximations which must be valid in all efficient transmission systems. These approximations cannot be all justified *a priori;* their justification must come *a posteriori* from the fact that the final solution satisfies the original assumptions and approximations.

[4] It is to be noted that the solution sought is the *principal wave.* (See "The Radiated and Guided Energy in Wire Transmission," *Trans. A. I. E. E.*, 1924.) This wave does not, in general, completely represent the phenomena, except at a considerable distance from the physical terminals of the transmission system, and then only in the neighborhood of the conductors.

First we have to define what we mean by conductor and by dielectric; the significance of these definitions will appear in the course of the analysis. *A conducting medium* is one in which ω^2/v^2 is very small compared with $4\pi\lambda\mu\omega$; while a *dielectric medium* is one in which $4\pi\lambda\mu\omega$ is very small compared with ω^2/v^2. The intermediate cases will not be discussed in the present paper; in the following it will be assumed that the conductors and dielectrics satisfy these definitions.[5]

The assumptions which we make at the outset in the approximate solution may now be listed and qualitatively justified as follows:

1. The propagation constant γ is an extremely small quantity and its real part is not large compared with its imaginary part. Since $|\gamma|$ is of the order of magnitude of $\omega \cdot 10^{-10}$, it is evident that γ is very small even for frequencies of millions of cycles per second. As regards the second restriction, if the real part of γ is large compared with the imaginary, the wave will be damped out in a few wave-lengths, and the system cannot efficiently transmit energy.

2. *In the conductors* the axial electric intensity E_z is large compared with the component normal to Z. This restriction means that the dissipation in the conductors due to the axial currents is large compared with the dissipation due to the charging currents. Evidently this restriction is necessary for the efficient transmission of energy.

3. *In the dielectric* the axial electric intensity is small compared with the normal electric intensity. The justification of this assumption is as follows: The propagation of energy occurs in the dielectric, and is normal to the direction of the electric intensity. Since the usefully transmitted energy is propagated along the axis of transmission and the propagation normal to the axis simply means dissipation, the axial electric intensity must be small compared with the normal component for efficient transmission.

4. The axial magnetic intensity H_z is everywhere small compared with the normal intensity. The justification of this assumption depends on the same arguments as (3).

As regards (3) and (4) it will be remarked that in the ideal plane wave propagation both E_z and M_z are zero. In the case of imperfect conductors E_z in the dielectric is not zero but may be regarded as a first order small quantity. M_z on the other hand is to be regarded as a second order small quantity because it not only vanishes for the case of perfect conductors but also vanishes for the case of imperfect conductors for the case where the wave is made up of a set of compo-

[5] In accordance with these definitions, conductors and dielectrics depend for their classifications on the frequency, as well as their electrical constants. The definition of *conductor* means that the displacement current is negligible compared with the conduction current.

nent radially symmetrical waves oriented on the axes of the conductors; to this the actual wave approximates in important transmission systems.

We shall now introduce the consequences of the foregoing assumptions into the differential equations of the problem.

IV

Referring to equations (10), these may be replaced *in the conductors only* where γ^2 is very small compared with ν^2 and γ is a very small quantity, by the approximation:

$$\begin{aligned} M_x &= -\frac{\partial}{\partial y} E_z, \\ M_y &= \frac{\partial}{\partial x} E_z, \\ E_x &= \frac{\gamma}{\nu^2}\left\{\frac{\partial}{\partial x} E_z + \frac{1}{\gamma}\frac{\partial}{\partial y} M_z\right\}, \\ E_y &= \frac{\gamma}{\nu^2}\left\{\frac{\partial}{\partial y} E_z - \frac{1}{\gamma}\frac{\partial}{\partial x} M_z\right\}, \end{aligned} \tag{13}$$

Therefore *in the conductors* the vector components M_x, M_y are derivable by spatial differentiation from E_z. E_x, E_y are not in general so derivable on account of the factor $1/\gamma$, a very large quantity, which appears with M_z. (It appears that γE_z and M_z may be of comparable orders of magnitude.) We assume, however, for reasons discussed above, that both E_x and E_y are very small compared with E_z *in the conductors*.

In the dielectric, where ν^2 and γ^2 are of comparable orders of magnitude, the foregoing approximations are not valid and the rigorous equations must be employed. Returning to equations (10) and writing for convenience $\gamma^2/\nu^2 = \beta$, we have

$$\begin{aligned} M_x &= -\frac{1}{1-\beta}\left\{\frac{\partial}{\partial y} E_z - \frac{\beta}{\gamma}\frac{\partial}{\partial x} M_z\right\}, \\ M_y &= \frac{1}{1-\beta}\left\{\frac{\partial}{\partial x} E_z + \frac{\beta}{\gamma}\frac{\partial}{\partial y} M_z\right\}, \\ E_x &= \frac{\beta}{\gamma}\frac{1}{1-\beta}\left\{\frac{\partial}{\partial x} E_z + \frac{1}{\gamma}\frac{\partial}{\partial y} M_z\right\}, \\ E_y &= \frac{\beta}{\gamma}\frac{1}{1-\beta}\left\{\frac{\partial}{\partial y} E_z - \frac{1}{\gamma}\frac{\partial}{\partial x} M_z\right\}. \end{aligned} \tag{14}$$

In equations (13) and (14), x and y may be any orthogonal coordinate system. Let us suppose that they are so chosen that x is

tangential to the conductor surface; M_y is therefore the normal component of M at the surface of the conductor and must there be continuous. E_z and $(\partial/\partial x)E_z$ are also continuous. Consequently, we must have by equating M_y as given by (13) and (14),

$$\frac{1}{\gamma}\left(\frac{\partial}{\partial y}M_z\right)_e = -\frac{\partial}{\partial x}E_z, \tag{15}$$

the subscript e indicating the value of $(\partial/\partial y)M_z$ outside the conductor. But from the expression for E_x, as given by (14), this is precisely the condition that makes $E_x = 0$ at the surface of the conductor. Consequently we arrive at the very important proposition that, subject to the approximations involved in (13), *the tangential component of E in the xy-plane vanishes at the conductor surfaces.*

We shall now find it convenient to express the field in the dielectric in accordance with (8) in terms of the wave functions F, Φ, Θ. Writing

$$\Theta = \theta \cdot \exp(i\omega t - \gamma z), \tag{16}$$

θ satisfies the differential equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\theta = (\nu^2 - \gamma^2)\theta. \tag{17}$$

Now, *in the dielectric*, ν^2 and γ^2 are both exceedingly small quantities which are nearly equal, so that $\nu^2 - \gamma^2$ is the difference of two very small and nearly equal quantities. We therefore replace it by zero, so that

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\theta = 0. \tag{18}$$

θ is therefore a two-dimensional potential function. Consequently *a conjugate two-dimensional potential function ψ exists, such that*

$$\begin{aligned} \frac{\partial}{\partial x}\theta &= \frac{\partial}{\partial y}\psi, \\ \frac{\partial}{\partial y}\theta &= -\frac{\partial}{\partial x}\psi. \end{aligned} \tag{19}$$

Writing

$$\Psi = \psi \cdot \exp(i\omega t - \gamma z),$$

equations (8) become

$$M_x = \frac{\partial}{\partial y}(F - \gamma\Psi),$$

$$
\begin{aligned}
M_y &= -\frac{\partial}{\partial x}(F - \gamma\Psi), \\
E_x &= -\frac{\partial}{\partial x}(\Phi - \Psi), \\
E_y &= -\frac{\partial}{\partial y}(\Phi - \Psi), \\
E_z &= \gamma(\Phi - \Psi) - (F - \gamma\Psi).
\end{aligned} \tag{20}
$$

Introducing new wave functions

$$
\begin{aligned}
F' &= F - \gamma\Psi, \\
\Phi' &= \Phi - \Psi,
\end{aligned} \tag{21}
$$

we have (dropping primes)

$$
\begin{aligned}
M_x &= \frac{\partial}{\partial y}F, \\
M_y &= -\frac{\partial}{\partial x}F, \\
E_x &= -\frac{\partial}{\partial x}\Phi, \\
E_y &= -\frac{\partial}{\partial y}\Phi, \\
E_z &= \gamma\Phi - F,
\end{aligned} \tag{22}
$$

where now Φ and F are *independent* wave functions.

If the foregoing analysis has been carefully followed, the important advantage of equations (22) as compared with (8) will be appreciated. The transformation of (8) into (22) is strictly dependent upon and conditioned by the legitimacy of neglecting $\nu^2 - \gamma^2$ in the dielectric, whereby the wave functions are essentially reduced to two-dimensional potential functions. It is evident that the whole engineering theory of transmission involves this approximation.

V

We are now prepared to sketch the general solution of the problem,[6] employing equations (13) *in the conductors*, and equations (22) *in the dielectric*. The procedure is as follows:

1. At the surfaces of the conductors the tangential component E_τ in the xy-plane of E vanishes, as shown above. That is,

$$
E_\tau = -\frac{\partial}{\partial \tau}\Phi = 0 \tag{23}
$$

[6] For detailed applications of this method of solution to specific problems, the published papers referred to in the introduction to this paper may be consulted.

at the surface of each conductor.[7] In the dielectric outside the conductor, the potential Φ satisfies Laplace's equation in two dimensions; hence

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\Phi = 0 \tag{24}$$

in the dielectric; and

$$\frac{\partial}{\partial \tau_j}\Phi = 0, \qquad (j = 1, 2, \cdots n) \tag{25}$$

at the surface of the jth conductor. Also

$$\oint E_n{}^j d\tau_j = -\int \frac{\partial}{\partial n_j}\Phi d\tau_j = \frac{4\pi}{k}Q_j, \qquad (j = 1, 2, \cdots n) \tag{26}$$

the integration being carried around the surface of the jth conductor, Q_j being the charge per unit length on the jth conductor.

The determination of Φ from (24)–(26), when the geometry of the conductors is specified, is a well-known two-dimensional potential problem, for the solution of which very general methods are available. The solution results in the form

$$\Phi = \phi_1(x, y)Q_1 + \phi_2(x, y)Q_2 + \cdots + \phi_n(x, y)Q_n. \tag{27}$$

That is, Φ is a linear function of the conductor charges $Q_1 \cdots Q_n$, and the coefficients $\phi_1 \cdots \phi_n$ are unique functions of the geometry of the transmission system and are determinable by the usual methods of two-dimensional potential theory.

2. The continuity of M_n and $(1/\mu)M_\tau$ at the surfaces of the conductors is analytically formulated by the equations

$$\begin{aligned}\frac{\partial}{\partial \tau}F &= -\frac{\partial}{\partial \tau}E_z,\\ \frac{\partial}{\partial n}F &= -\frac{\mu}{\mu_c}\frac{\partial}{\partial n}E_z,\end{aligned} \tag{28}$$

where μ is the permeability of the dielectric and μ_c that of the conductor. These relations, it will be understood, hold at the surfaces of all the conductors. F is a wave function which satisfies Laplace's equation in two dimensions *in the dielectric;* thus

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)F = 0, \tag{29}$$

[7] In the following, τ and n denote vectors tangential and normal to the conductor surface respectively.

and E_z is a wave function which *in the conductor* satisfies the two-dimensional wave equation

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)E_z = \nu^2 E_z. \tag{30}$$

In addition, E_z and F are connected with the conductor current I by the relations

$$\begin{aligned} I &= \lambda \int E_z dS, \\ 4\pi\mu i\omega \cdot I &= \oint \frac{\partial}{\partial n} F d\tau, \end{aligned} \tag{31}$$

λ here denoting the conductivity of the conductor.

It follows at once that the determination of F and E_z from (28)–(31) is a generalization of the two-dimensional potential problem involved in the determination of Φ from (24)–(26); it may be precisely stated as follows:

The function F satisfies Laplace's equation

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)F = 0 \tag{32}$$

everywhere outside the n conductors. Inside the jth conductor the electric force E_z^j satisfies the two-dimensional wave equation

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)E_z^j = \nu_j^2 E_z^j, \qquad (j = 1, 2, \cdots n) \tag{33}$$

while at the surface of the jth conductor

$$\begin{aligned} \frac{\partial}{\partial \tau_j} F &= -\frac{\partial}{\partial \tau_j} E_z^j, \\ \frac{\partial}{\partial n_j} F &= -\frac{\mu}{\mu_j}\frac{\partial}{\partial n_j} E_z^j, \qquad (j = 1, 2, \cdots n) \end{aligned} \tag{34}$$

and

$$4\pi\mu i\omega \cdot I_j = \oint \frac{\partial}{\partial n_j} F d\tau_j, \qquad (j = 1, 2, \cdots n). \tag{35}$$

Just as equations (24)–(26) uniquely determine Φ as a linear function of $Q_1 \cdots Q_n$, so equations (32)–(35) uniquely determine the potential function F in the dielectric and the electric intensities $E_z^{(1)} \cdots E_z^{(n)}$ in the n conductors as linear functions of the conductor currents; thus

$$F = f_1(x, y)I_1 + f_2(x, y)I_2 + \cdots + f_n(x, y)I_n, \tag{36}$$

and

$$E_z^j = e_{1j}(x, y)I_1 + \cdots + e_{nj}(x, y)I_n, \qquad (j = 1, 2, \cdots n) \quad (37)$$

the f and e functions depending on the geometry of the conducting system and, through the parameter ν^2, on its electrical constants. They are uniquely determined by the differential equations.

The actual solution of the differential equations (32)–(35) is essentially more difficult than the solution of (24)–(26) involved in the determination of Φ, and they have not been subjected to the exhaustive study accorded to the potential problem. On the other hand, the analogy with the potential problem suggests that extension and modifications of the general and well-known methods of solution available for that problem should be possible.

To summarize the foregoing we have succeeded in expressing the potential function Φ (and therefore E_x, E_y) *in the dielectric* as a linear function of the conductor charges, the coefficients of the conductor charges $Q_1 \cdots Q_n$ being spatial functions of x and y which we determined by the usual methods of two-dimensional potential theory.

Similarly it has been shown that E_z (and therefore the current distribution) *in the conductors*, and the potential function F (and therefore the magnetic field) *in the dielectric*, are expressible as linear functions of the conductor currents $I_1 \cdots I_n$, the determination of the coefficients depending on the solution of a generalization of the two-dimensional potential problem.

3. To complete the solution of the problem, recourse is had to the fact that E_z is continuous at the surfaces of the conductors. At the surface of each conductor we therefore equate E_z, as given by (37), in terms of the currents $I_1 \cdots I_n$ with $\gamma\Phi - F$ (see (22)), Φ being given by (27) in terms of $Q_1 \cdots Q_n$ and F by (36) in terms of $I_1 \cdots I_n$. This gives n equations of the form

$$\begin{aligned} Z_{11}I_1 + \cdots + Z_{n1}I_n &= \gamma\Phi_1 = \gamma(p_{11}Q_1 + \cdots + p_{1n}Q_n), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_{n1}I_1 + \cdots + Z_{nn}I_n &= \gamma\Phi_n = \gamma(p_{n1}Q_1 + \cdots + p_{nn}Q_n). \end{aligned} \quad (38)$$

Here $\Phi_1 \cdots \Phi_n$ are the values of Φ at the surfaces of the n conductors respectively; the p coefficients are Maxwell's potential coefficients of the system, while the Z coefficients are the self and mutual "impedances" of the conductors.

We require a further relation between I and Q; this is furnished by the well-known relation

$$\begin{aligned} i\omega Q &= \gamma I - \lambda \oint E_n ds \\ &= \gamma I + \lambda \oint \frac{\partial}{\partial n} \Phi ds \\ &= \gamma I - \frac{4\pi\lambda}{k} Q, \end{aligned} \tag{39}$$

the integration being carried around the contour of the conductor. (λ is the conductivity of the dielectric and the last term is the "leakage" current.) We have therefore, for a homogeneous dielectric,

$$\left(i\omega + \frac{4\pi\lambda}{k} \right) Q = \gamma I, \tag{40}$$

which furnishes the necessary relation.

Elimination of Q from (38) by means of (40) gives n homogeneous equations in $I_1 \cdots I_n$, the coefficients involving only one unknown quantity, the propagation constant γ. A finite solution necessitates the vanishing of the determinant of the coefficients; equating this to zero gives an nth order equation in γ^2, which determines the n possible values of γ, and therefore the n possible modes of propagation in the system. The formal solution of the problem is thus completed.

In conclusion it is worth while reviewing and summarizing the mathematical restrictions on the solution developed in the foregoing pages; restrictions which have their counterpart in the physical requirements of the system for the efficient guided transmission of electromagnetic energy. The essential restrictions are that (1) *in the conductors* γ^2 is very small compared with ν^2, and (2) *in the dielectric* the wave equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \Phi = (\nu^2 - \gamma^2)\Phi$$

may be replaced, at least in the neighborhood of the conductors, by

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \Phi = 0.$$

If the conductors are so imperfect, or the dielectric so dissipative that these approximations are not justified, the method of solution

given above breaks down, and the problem must be attacked from the rigorous equations. These have never been solved in general, in fact the only rigorous solution known to the writer is for the case of circular symmetry and even this involves the location of the roots of an extremely complicated transcendental equation. Fortunately, in view of these difficulties, the general case of quite imperfect conductors or imperfect dielectric media is of small technical importance for the reason given above.

Some General Results of Elementary Sampling Theory for Engineering Use

By PAUL P. COGGINS

EVERY day we base conclusions on the results of the process commonly known as "sampling." For example, if five times in a week a man has waited ten minutes or more for his trolley at a street corner, he may conclude that the transportation facilities are poor. Or again, if a housewife has bought ten loaves of bread at a certain store and has found five of them not as fresh as might be desired, she decides that in the future she will buy her bread elsewhere. Both of these conclusions are based on an intuitive application of sampling theory. Such examples could be multiplied indefinitely.

Similarly, in most engineering problems, observational data are involved in one way or another. In order to be able to assign the proper significance to these data, it is essential to have some idea as to their reliability, that is, to what extent they represent all the facts under consideration. First, the measurements themselves may be in error. In the second place, although the observations may have been made with perfect precision, they may be incomplete; they may constitute but a "sample" of a large group of possible observations. The problem considered in this paper is one of this second class, generally known as "sampling" problems.

Assume the existence of a total group or "universe" of N objects and that observations have been made on a certain number n of them with reference to a particular characteristic. This number n we will call the "sample." From this sample we wish to deduce some estimate concerning the probable condition of that universe with reference to the characteristic observed.

Now the characteristic observed may itself take on one of two forms. It may be either, (1) present or absent; (2) quantitative. For simplicity in discussion we may call the first, "Sampling of Attributes," and the second, "Sampling of Variables."

An example of each will be cited from the telephone field.

Example 1: Sampling of Attributes

Suppose that 4,000 relays of a particular type constitute a day's output. In order to determine roughly what proportion of these are non-operative at a current of 12 mils, a sample of 500 relays is tested and out of this sample 10 fail to operate at the required current. In

the sample, then, two per cent of the relays were defective. What, then, is the probability that the percentage of the 4,000 relays having this defect is between one and three per cent? Or what is the probability that the percentage of defectives in the universe of 4,000 does not exceed four per cent? Or again, if we wish to be practically certain that among the 4,000 relays not more than two per cent are defective in this respect, how many defectives would be allowable in a sample of 200? or a sample of 1,000? Any number of questions of this sort can be asked and may be answered on the basis of the proper assumptions by sampling theory.

Example 2: Sampling of Variables

An office serves 5,000 subscribers lines. Measurements of the insulation resistance are made on 200 of these, selected at random, and the resulting values tabulated. They vary all the way from 12,000 ohms to 200,000 ohms. What conclusions may be drawn as to the probability that more than a certain number, say 20 of the subscribers' loops out of the 5,000, have an insulation resistance of less than 18,000 ohms? What is the most probable distribution of the insulation resistances for the office as a whole? What is the probable error of the average of the observations as a measure of the average loop insulation resistance for the office?

As before, much information *regarding the universe* may be inferred from a properly chosen sample, always, however, with some degree of uncertainty. This uncertainty, so far as the sampling process is concerned, naturally decreases as the size of the sample increases, and, of course, disappears except for inaccuracies of measurement, when the sample becomes coextensive with the universe.

The respective treatments of these two types of problems differ considerably in detail. The basic principles are, however, essentially the same, and involve in each case the notions of "a posteriori" probability, as discussed in most of the standard textbooks on the theory of probability.

In both problems there are certain observations. By means of these we desire to obtain as precise information as possible concerning some one or more characteristics of the universe from which these observations or samples were drawn. The true nature of the universe is to some degree, at least, unknown. Certain hypotheses concerning it may, however, in the light of the sample be more probable than others. What we wish to estimate is the probability that either a particular hypothesis or a group of mutually exclusive hypotheses includes the true one.

This article will be devoted to the type of problem termed "Sampling of Attributes." [1] In it are included results from an extensive series of computations in the form of charts which may be of value in the solution of practical engineering problems. The nomenclature is general, so as to be applicable to a wide variety of practical problems. For convenience in discussion we shall divide the units of any sample into the two mutually exclusive classes, "defective" and "satisfactory." The following notation will be used:

N = total number of items in universe,

n = total number of items in sample,

X = number of defective items in universe (unknown),

c = number of defective items in sample (observed),

$w(X)$ = *a priori* probability that the universe will contain exactly X defectives,

$W(X_1, X_2)$ = *a posteriori* probability that the universe contains a number of defectives X such that $X_1 \leqq X \leqq X_2$.

It is of extreme importance that, at the outset, the significance of the symbol $w(X)$ in sampling problems be clearly defined. It is a measure of the probability, *before the sample* is taken, that the lot or universe in question contains X defective items and $N - X$ satisfactory items. It may be based on previous samples, or the reputation of the manufacturer producing those items, or on any one or more of a number of other pertinent data. For example, even before a sampling inspection, we should unhesitatingly say that in a lot of 1,000 relays sent out by a reputable manufacturer it is very much more likely, *a priori*, that the lot will contain less than 100 relays with a short-circuited winding than that the lot will contain more than 800 relays defective in the same respect. We should probably find ourselves in a quandary, however, if we attempted to state without a sample inspection, the relative likelihoods of 3, 4, 5, 6, $\cdots$, etc., defectives existing in the lot. $w(X)$ is a function whose numerical value is assumed to state this *a priori* probability. The extent to which we are able to make use of this function, then, depends on how precisely we are able to assign numerical values to it before we study our sample.

[1] This general type of problem has been under study within the Bell System for some time. In an article "Deviation of Random Samples from Average Conditions and Significance to Traffic Men" by E. C. Molina and R. P. Crowell which appeared in the *Bell System Technical Journal* for January, 1924, a special case of sampling theory was developed and various possible applications were suggested. In August, 1924, Molina delivered a paper entitled "A Formula for the Solution of Some Problems in Sampling" before the statistical section of the International Mathematical Congress in Toronto, Canada. This paper dealt with a somewhat more general case of the sampling problem than was discussed in the article just mentioned.

It may be helpful at this point to state and solve a simple problem, which will serve to bring out the fundamental principles involved. An urn is known to contain 10 balls, some of which are white and the others black. Five balls have been drawn and *not* replaced. Of these five, one is white and four are black. What is now the probability that the urn originally contained just one white ball and nine black? Two white and eight black?

Before we proceed to obtain a solution for this problem we have to make some assumption, based on knowledge available before the drawings were made, concerning the probability that the urn contains black and white balls in any given proportion.

Consider two such assumptions—

(*a*) All proportions are *a priori* equally likely, i.e., before the drawings it is as likely that three whites and seven blacks were put in the urn as six whites and four blacks, etc.

(*b*) The urn was filled with ten balls drawn at random from a bag containing a very large number of balls of which a quarter are white and the remainder are black.

There are, before the drawings, 11 possible hypotheses concerning the contents of the urn. They range from 0 whites and 10 blacks to 10 whites and 0 blacks, as listed in the two left-hand columns of Table I and shown in Fig. 1. The probability in favor of each of

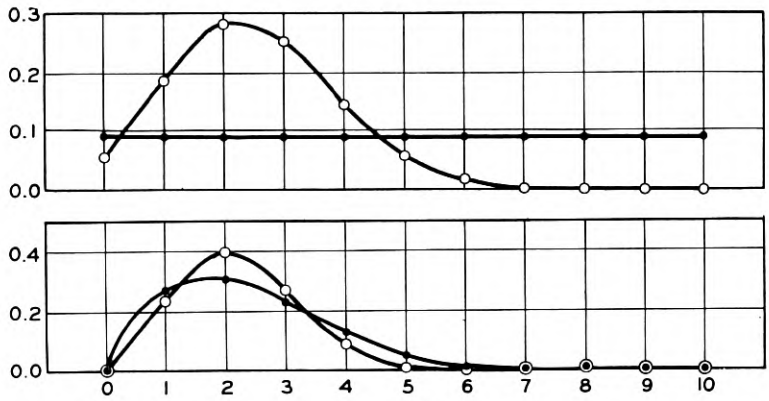


Fig. 1. The upper curve shows two different assumptions concerning the *a priori* probabilities, while the lower pair shows the *a posteriori* probabilities. In both cases the dots refer to the hypothesis of uniform *a priori* probability while the circles refer to the assumption that the urn itself is a random sample from a large stock of which one fourth of the balls are white.

these hypotheses is the "*a priori* existence probability" in favor of the hypotheses, and is represented by the symbol $w(X)$, X referring to the number of white balls assumed to be in the urn.

Under assumption "a" (Case 1) each hypothesis has a probability of 1/11 or .090909. Under assumption "b" (Case 2) the probability that the urn contains X whites and $10 - X$ blacks is the binomial term $\binom{10}{X}\left(\frac{1}{4}\right)^X\left(\frac{3}{4}\right)^{10-X}$.

TABLE I

Contents		Existence Prob. $w(X)$		Prod. Prob.	*A Posteriori* Prob. P_X	
"X" Wh.	Bl.	Case 1	Case 2	p_X	Case 1	Case 2
0	10	.090909	.056314	.000000	.000000	.000000
1	9	.090909	.187712	.500000	.272727	.237305
2	8	.090909	.281568	.555556	.303030	.395509
3	7	.090909	.250282	.416667	.227272	.263671
4	6	.090909	.145998	.238095	.129870	.087889
5	5	.090909	.058399	.099206	.054112	.014650
6	4	.090909	.016222	.023810	.012987	.000876
7	3	.090909	.003090	.000000	.000000	.000000
8	2	.090909	.000386	.000000	.000000	.000000
9	1	.090909	.000029	.000000	.000000	.000000
10	0	.090909	.000001	.000000	.000000	.000000

In the column headed p_X we give the productive probabilities for both cases. These are the probabilities that five drawings from an urn whose contents were as given by the corresponding hypothesis would yield the observed one white ball and four black. These are zero in the case of $X = 0$, 7, 8, 9 and 10 since urns so constituted could not have given the observed drawings.

For the other cases, the productive probability is the ratio

$$p_X = \frac{\binom{X}{1}\binom{10 - X}{4}}{\binom{10}{5}}.$$

In this expression the denominator is the total number of combinations of 10 balls taken five at a time, and the numerator is the number of ways of selecting one out of X white balls and four out of the remaining $10 - X$ black balls. These figures are tabulated in Table I under the heading p_X.

We now have all of the component parts of our problem under the two different assumptions "a" and "b." It only remains to apply "Bayes' Rule."[2] Now the generalized Bayes Rule tells us that the *a posteriori* probability, P_X, in favor of an hypothesis *after* the drawings

[2] This rule was first enunciated by an English cleric, Bayes by name, in a memoir in *Philosophical Transactions* for 1763. It was generalized by Laplace in 1812 to cover cases not equally likely.

have been made and taking account of the *a priori* information is given by the ratio

$$P_X = \frac{w(X)p_X}{\sum w(X)p_X},$$

the summation in the denominator being extended over all possible cases.

The numerical values of this ratio are shown in the last two columns of Table I, corresponding to the two assumptions in our problem and also by the circles in Fig. 1. That we should have a different set of results corresponding to the different assumptions is to be expected. It is interesting, however, that the difference in this case is by no means great as Fig. 1 brings out.

If after each drawing we had replaced the ball drawn, we would have used for the productive probability p_X the binomial term

$$p_X = \binom{5}{1}\left(\frac{X}{10}\right)^1\left(\frac{10 - X}{10}\right)^4$$

since the successive drawings would not have changed the relative constitution of the urn. The same would also be true if the urn contained an indefinitely large number of balls with the same relative proportions of black and white.

Now if we agree that a white ball corresponds to a defective item and a black ball to an acceptable item, we are immediately able, by the use of these fundamental principles of *a posteriori* probability, to write the general basic formal relation

$$W(X_1, X_2) = \frac{\sum_{X=X}^{X=X_2} w(X)\binom{X}{c}\binom{N-X}{n-c}}{\sum_{X=c}^{X=N-n+c} w(X)\binom{X}{c}\binom{N-X}{n-c}}. \qquad (1)^3$$

As we have just indicated, the troublesome element in this formula is the function $w(X)$ to which, in many practical problems, it is difficult to assign any particular numerical values. In order to proceed further, therefore, without detailed consideration of various specific engineering problems we are forced to make some rather general assumptions concerning the nature of the function $w(X)$.

Case I

One of the most natural assumptions to make when no knowledge exists to the contrary is that $w(X)$ is a constant within that range

[3] It should be noted that in his original treatment of this formula Molina used S instead of Σ as the symbol for summation on account of the fact that finite integration entered into his analysis. Since in this presentation we are dealing only with summation, we shall use the commoner form Σ to denote summation.

of values of X which essentially affects the value of the denominator of (1). This assumption may seem at first glance rather arbitrary and wide of the mark, especially since the range of values which essentially affects the value of the denominator in (1) depends on the value of c obtained from the sample. However, if the sample is reasonably large, consisting of 100 items or more, and the proportion of defectives observed is small, say 10 per cent or under, the probability that universes having a proportion of defectives widely different from the one observed would yield such results is so small that a wide range of assumptions concerning the *a priori* probability of such universes existing makes very little change in the final result.

Applying, then, this assumption analytically to the basic formula (1) we obtain the simpler formulæ

$$W(X_1, X_2) = \frac{\sum_{X=X_1}^{X=X_2} \binom{X}{c}\binom{N-X}{n-c}}{\sum_{X=c}^{X=N-n+c} \binom{X}{c}\binom{N-X}{n-c}} = \frac{\sum_{X=X_1}^{X=X_2} \binom{X}{c}\binom{N-X}{n-c}}{\binom{N+1}{n+1}} \tag{2}$$

and by means of a transformation outlined in the Appendix we obtain, from (2),

$$W(X_1, X_2) = \frac{\sum_{t=0}^{t=c}\left[\binom{X_1}{t}\binom{N+1-X_1}{n+1-t} - \binom{X_2+1}{t}\binom{N-X_2}{n+1-t}\right]}{\binom{N+1}{n+1}} \cdot \tag{2a}$$

Formula (2*a*) is the one embodied in the paper referred to in footnote 1. While apparently less simple than (2), it is actually easier to compute when c is less than the range $X_2 - X_1$.

When in (2*a*) we set $X_1 = c$ and $X_2 = X$ the resulting formula

$$W(c, X, n, N) = 1 - \frac{\sum_{t=0}^{t=c} \binom{X+1}{t}\binom{N-X}{n+1-t}}{\binom{N+1}{n+1}}, \tag{3}$$

which is at the basis of our computational work, shows explicitly certain properties which are not apparent in (2). Various analytical transformations and approximations based on this formula lead to several interesting extensions which are discussed in the Appendix. We shall leave these phases of the problem for the present, however, and discuss the results of the calculations which have been made as presented on the attached charts.

Charts A

Charts A have been prepared by means of exact formula (3) to show, for universes $N = 300$, 500, 700 and 900 from which samples, n, of various indicated sizes are assumed to have been drawn, the probability or "weight" $W(c, X)$ as ordinate versus X as absicissa for various values of c as indicated by the solid curves so designated. The dotted curves crossing these solid curves show the weight indicated by various values of the difference "d" between the percentage observed defective and the percentage assumed defectives in the universe.

As examples illustrating the interpretation of Charts A consider the following:

Example 1: From a universe of $N = 700$ items a random sample $n = 300$ items has shown $c = 3$ or 1 per cent defectives. What is the probability or weight to be associated with the hypothesis that the universe contains not more than $X = 14$ or two per cent defectives? From the A Chart corresponding to $N = 700$ and $n = 300$ we find the $c = 3$ curve (shown heavy because it is an even per cent of the sample $n = 300$). On this curve corresponding to an abscissa of $X = 14$ we read our desired result as the ordinate $W = .94$. We note that this is also a point on the $d = 1$ per cent dotted curve since $100(X/N - c/n)$ per cent $= 1$ per cent.

Example 2: We are going to make a sample of $n = 199$ items out of a universe of $N = 500$ items and wish the weight or probability to be .9 or better that the universe does not contain more than five per cent defective items. What is the maximum number of defective items that we may tolerate in our sample? Now five per cent of $N = 500$ is $X = 25$. Corresponding to an abscissa $X = 25$ and an ordinate $W = .9$ we locate a point which lies between the $c = 6$ and $c = 7$ curves. We could, therefore, accept the lot provided the sample showed six or less defectives, or three per cent or less defectives.

These Charts A are fundamental in nature, and involve the five variables, N, n, X, c and W. The formula by means of which they were computed is exact on the basis of the assumptions. Such errors or irregularities as may appear to exist in them are of negligible practical importance in view of the nature of the assumptions made, and are mainly due to the difficulties in drafting such a family of curves.

Naturally a function involving several variables may be represented graphically in many different ways, some of which may be more convenient than others to use in connection with various practical problems. One of the restrictions often encountered in practical

problems is that the weight W shall not be less than some specified figure which may be considered to give us the desired degree of confidence in the efficacy of our sampling procedure in weeding out defective lots. Charts B and C are drawn up on the basis of three such specified figures which are of practical interest, $W = .75$, $W = .9$, and $W = .99$. Such restrictions enable us to show, without the large amount of labor which would be required without them, the results of calculations for a wider range of the other variables.

Chart B

Chart B shows roughly for the proportion of observed defectives $c/n = .01$, .04, and .07, the proportion of defectives in the universe which we may expect not to exceed with weights $W = .75$, .9 and .99 for various values of the sample n as abscissa and for $N = 300$, 500, 700, 900 and also the limit approached as N becomes infinite. This form of presentation serves to relate the present material to the earlier charts which accompanied the earlier article already mentioned as having appeared in the *Bell System Technical Journal* for January, 1924, and shows how with a given size of sample n and a given proportion of defectives observed, the larger the value of the universe N, the larger the variation which may be expected with any given degree of probability. As would be expected, we also see that when the size of the sample approaches the size of the universe, the range of uncertainty approaches 0 and our sample inspection becomes a complete inspection.

It will be noted that, up to the present point, we have not considered cases for $N > 1{,}000$. The exact formulæ become rather troublesome to compute for these larger values of N. Fortunately, however, various approximate methods outlined in the Appendix become sufficiently accurate to be of service in these cases.

Charts C

We have, therefore, by their aid when $N > 1{,}000$, prepared the Charts C which we believe will cover a rather wide range of the variables with sufficient precision to be of considerable practical value. The points shown by dots are believed to be accurate to the degree to which they are readable on the chart. For intermediate values and for other values of the trouble limit the discrepancies are indicated on the charts. One of these charts corresponds to each of the three following weights, $W = .75$, $W = .9$, and $W = .99$. As abscissa we show the per cent sample, $100\ n/N$. The ordinate scale is proportional to the number of items n in the sample. The same

proportionality factor K enters also in the ratio X/N which we designate as the trouble limit. We shall later discuss the purpose of this factor K in more detail. The understanding of the charts will be simplified, however, if we consider the case for $K = 1$ in which the charts become direct reading for the case of a trouble limit $X/N = .01$.

The values of c, the number of defective items observed in the sample, are shown as a family of curves marked $c = 0$, $c = 1$, $c = 2$, etc., sloping downward from left to right. Any point on the $c = 5$ curve, for example, on the Chart C for weight $W = .9$ shows the corresponding values of n as ordinate and n/N as abscissa which are necessary in order that this number of defectives may be accepted with a degree of assurance [4] indicated by $W = .9$ that the true proportion of defectives in the universe N is not greater than .01.

It will be readily noted that for every value of the universe N, there may be drawn a diagonal straight line through the origin whose ordinate for an abscissa of 100 per cent sample is equal to $n = N$. Certain representative N lines are drawn in on the charts in this manner, and as many more could be inserted as desirable. Thus, for a constant value of W and a constant value of X/N we have provided on Charts C a ready means of determining the relationships which must exist between the remaining variables N, n, and c.

As an example of the use of these charts for the case where $K = 1$, i.e., for $X/N = .01$, consider the following:

Example 3: In a sample of $n = 900$ out of a universe $N = 3{,}000$, what is the maximum number of defectives c that we may accept with an assurance of $W = .9$ or better that the true proportion of defectives in the universe is not greater than .01?

Referring to the Charts C for $W = .9$ and considering $K = 1$, we locate the point corresponding to an abscissa of 100 n/N per cent $= 90{,}000/3{,}000 = 30$ per cent, and an ordinate $n = 900$. We find that this lies on the diagonal straight line marked $N = 3{,}000$ K as it should and that it also lies between the $c = 5$ and $c = 6$ curves. From this we may infer that we may accept five defectives but not six in the above case.

We shall now proceed to explain the significance of the factor K and the cross-hatched areas beneath the $c = 0$, 5, 10, 15, etc., curves. The purpose of these features is to extend the application of Charts C to values of X/N other than .01. It may be noted from the mathematical analysis or from actual plotting of charts similar to Charts C, but for different values of X/N, that the general shape and spacing of the curves remains practically unchanged for any given value of W.

[4] This statement is not strictly true when we are dealing with non-integral values of X. In such cases the weights W shown on the Charts C are slightly too high.

In other words, the value of $W(c, X)$ depends mainly on the ratio n/N, and the values of X and c, and only in a secondary way on the absolute values of n and N. This being the case, if we make a given per cent sample of two different universes N and KN, the number of defectives c which we may allow in our sample out of the first universe N in order that our weight W may have a given value, .9 say, for the true proportion of defectives in this universe to be not greater than .01 is practically the same as the value of c that we may allow in the sample out of the second universe KN for the same weight W and a proportion of defectives $.01/K$. For values of $K > 1$ there is no appreciable change introduced in the location of the c curves on Charts C. For values of $K < 1$, some error is made. The magnitude of this error is indicated by the cross-hatched bands on the $c = 0$, 5, 10, 15, etc., curves. The lower boundaries of these bands were calculated to show the magnitude of the error introduced for the corresponding values of c when $K = .1$. The upper boundaries of these areas correspond to values of $K \geqq 1$. For other values of c only the upper boundaries of the corresponding bands are shown, the lower boundaries being easily deducible by visual interpolation to a sufficient degree of approximation for most practical purposes.

As examples which may serve to illustrate this sort of application of Charts C consider the following:

Example 4: A sample of $n = 5{,}000$ items has been drawn out of a universe of $N = 20{,}000$ items and $c = 15$ defectives were observed. May we assume with a weight $W = .9$ or more that the true proportion of defectives or trouble limit X/N is .005?

Here $.01/K$ is to equal .005 for our charts to apply. Therefore, $K = 2$. Our sample $n = 500 = 2{,}500K$ and our per cent sample is 100 $n/N = 25$ per cent. Corresponding then to an abscissa of 25 per cent and an ordinate of $2{,}500K$ on the $W = .9$ chart we locate a point between the $c = 19$ and $c = 20$ curves. We could have allowed, therefore, $c = 19$ defectives at the desired weight and trouble limit. Since we observed a smaller number of defectives than was allowed, our weight W is therefore greater than .9. As a matter of fact it is practically only slightly less than .99 as appears from the $W = .99$ chart when utilized in a corresponding manner.

Example 5: As our next example we shall attempt to determine what is the trouble limit which corresponds with $W = .9$ to the results of the sample of Example 4. On the $W = .9$ chart corresponding to an abscissa of 25 per cent we read from the $c = 15$ curve an ordinate of $2{,}015K$. But this must be our sample $n = 5{,}000$. We, therefore, determine K from the equation $2{,}015K = 5{,}000$ which gives $K = 2.48$. Hence, our corresponding trouble limit is

$$\frac{.01}{K} = \frac{.01}{2.48} = .0040.$$

So far our values of K have been greater than unity, so we have not had to consider our cross-hatched bands at all. In our next example we shall remedy this defect.

Example 6: What number of defective items c may be allowed in a sample of 200 items out of a universe of 500 items so that $W \geq .9$ corresponding to a trouble limit $X/N = .08$. Here $.01/K = .08$ $\therefore K = .125$. Our ordinate, therefore, is $200 = 1{,}600K$ and our abscissa is $200/500 \times 100 = 40$ per cent sample. The point corresponding to this on the $W = .9$ chart lies just below the $c = 12$ curve indicating at first glance that we could not accept 12 defectives in such a case. However, we note that $K = .125$ should be near the lower boundary of our cross-hatched band for $c = 12$ if such a band had been drawn in. From an inspection of the widths of the bands for $c = 10$ and $c = 15$ we correctly infer that our point determined by the 40 per cent sample and $1{,}600K$ would lie well within this band, and that after all we could accept 12 defectives in the example in question.

This example has been included merely to illustrate the interpretation of the bands shown on Charts *C*. It may be anticipated that in many if not most of the practical engineering problems only the upper boundaries of the bands need be used to obtain a degree of accuracy commensurate with the precision of the results desired and the applicability of the basic assumptions concerning randomness and the form of the *a priori* existence probability $w(X)$.

If it should be desired to extend the range of these charts to cover values of W other than those shown, this may be done by means of the methods outlined in the mathematical analysis, the particular method to be used depending on the degree of precision required.

The preceding pages have contained an outline of some of the theory and results based on the assumption that, within a range at least, all possible values of X, the unknown number of defectives, were *a priori*, that is, before the sample in question was made, equally likely. This assumption we mentioned as appropriate to consider in case we have no information to the contrary. The results may be also applicable to certain cases where we do have some information of a general sort, but which it is difficult to express analytically. However, it is by no means the only reasonable assumption to make concerning the form of $w(X)$ as it enters into the basic formula (1).

Case II

Another assumption is suggested by the following considerations which enter into many of the standard works on probability theory. Assume that the lot or universe in question was itself drawn at random from an extremely large stock or major universe in which the proportion of defective items was p. Under these conditions the *a priori* probability $w(X)$ that our universe of N items would contain exactly X defective items would be given by the expression

$$w(X) = \binom{N}{X} p^X (1 - p)^{N-X}.$$

Using this expression for $w(X)$ in the fundamental formula (1), we obtain, by a process given in detail under the heading Appendix, the formula

$$W(X_1, X_2) = \sum_{X=X_1}^{X=X_2} \binom{N-n}{X-c} p^{X-c} (1 - p)^{N-n-X+c},$$

which for $X_1 = c$ and $X_2 = X$ reduces to

$$W(c, X) = \sum_{t=0}^{X-c} \binom{N-n}{t} p^t (1 - p)^{N-n-t},$$

which is precisely the expression for the *a priori* probability that the remaining $N - n$ items which we did not inspect contain not more than the $X - c$ defectives which together with the c we have observed would assure us of a satisfactory universe.

In this form $W(c, X)$ turns out to be a simple binomial which, when $N - n$ is large and p is small, may be reasonably approximated by the Poisson Exponential Binomial Limit for which extensive curves and tables already exist[5] and will, therefore, not be included in this article.

In order to make practical use of the results of this assumption, we must have some knowledge of the appropriate value of the factor p in any given case. This factor should measure the probability that an item, selected at random, will, on inspection, prove to be defective. If a large number of tests have been made in the past on similar items, prepared by essentially the same process, the ratio of the total defectives observed to total items inspected in such tests may be a reasonable figure to use for p. In the case of many manufactured articles such a ratio ought not to be very difficult to obtain. In certain cases it might be necessary to allow for such factors as

[5] See article, "Probability Curves Showing Poisson's Exponential Summation," by George A. Campbell, *Bell System Technical Journal*, January, 1923.

trend, improved process of manufacture, changes in personnel and so on. In the final sampling of such complicated equipment as is involved in a completely installed telephone central office it may be necessary to take account also of breakage and such troubles due to shipping and setting up of the equipment which may introduce marked deviations from average conditions.

Such general considerations as these should determine whether or not we can safely assume any given value for p, and if so, what value. It will be evident on a little consideration that the assumed value for p need not be extremely precise for many practical applications.

Concerning the restrictions on the function giving the *a priori* probabilities, $w(X)$, it may be well to point out that this function is only defined for the positive integral values of X such that $0 \leq X \leq N$. Moreover, since probabilities are essentially positive, it cannot be negative for any of these values of X. Also since the composition of the lot is certainly comprised in this range of values of X, one has

$$\sum_{0}^{N} w(X) = 1.$$

The questions raised concerning the form of $w(X)$ are of particular importance in connection with the economic phases of sampling, that is, the relative costs of having satisfactory lots rejected and unsatisfactory lots accepted by the sampling process. These costs are, of course, dependent on the frequency with which given proportions of defects occur in the lots in practice, and a detailed consideration of these would itself warrant a separate treatment.

It is felt that the general methods outlined in this treatment, while not sufficiently detailed for immediate practical application to many of the problems in sampling of attributes, will nevertheless serve as a satisfactory basis for further work of a more specific nature.

APPENDIX

Case I: Assuming $w(X)$ is a constant and noting that[6]

$$\sum_{X=c}^{X=N-n+c} \binom{X}{c}\binom{N-X}{n-c} = \binom{N+1}{n+1},$$

the fundamental formula

$$W(X_1, X_2) = \frac{\sum_{X=X_1}^{X=X_2} w(X)\binom{X}{c}\binom{N-X}{n-c}}{\sum_{X=c}^{X=N-n+c} w(X)\binom{X}{c}\binom{N-X}{n-c}} \tag{1}$$

[6] Netto, "Lehrbuch der Kombinatorik," p. 15, Eq. 11.

gives us, as one computing formula,

$$W(X_1, X_2) = \frac{\sum_{X=X_1}^{X=X_2} \binom{X}{c}\binom{N-X}{n-c}}{\binom{N+1}{n+1}}, \tag{2}$$

which is fairly manageable so long as the range X_1 to X_2 is not too great and we have tables for the logarithms of the factorials involved. When $X_2 - X_1$ is large compared with c, one may use the equivalent formula

$$W(X_1, X_2) = \frac{\sum_{t=0}^{t=c} \binom{X_1}{t}\binom{N+1-X_1}{n+1-t} - \binom{X_2+1}{t}\binom{N-X_2}{n+1-t}}{\binom{N+1}{n+1}}. \tag{2a}$$

This transformation may, as Molina has shown, be effected as follows:[7]

$$\sum_{X=X_1}^{X=X_2}\binom{X}{c}\binom{N-X}{n-c} = \sum_{X=X_1}^{X=N-n+c}\binom{X}{c}\binom{N-X}{n-c} - \sum_{X=X_2+1}^{X=N-n+c}\binom{X}{c}\binom{N-X}{n-c}.$$

Now

$$\begin{aligned}\sum_{X=X_2+1}^{X=N-n+c}\binom{X}{c}\binom{N-X}{n-c} &= \sum_{X=X_2+1}^{X=N-n+c}\binom{N-X}{n-c}\left(\sum_{t=0}^{t=c}\binom{X_2+1}{t}\binom{X-X_2-1}{c-t}\right)\\ &= \sum_{t=0}^{t=c}\binom{X_2+1}{t}\left(\sum_{X=X_2+1}^{N-n+c}\binom{N-X}{n-c}\binom{X-X_2-1}{c-t}\right)\\ &= \sum_{t=0}^{t=c}\binom{X_2+1}{t}\binom{N-X_2}{n+1-t}.\end{aligned}$$

Likewise

$$\sum_{X=X_1}^{X=N-n+c}\binom{X}{c}\binom{N-X}{n-c} = \sum_{t=0}^{t=c}\binom{X_1}{t}\binom{N+1-X_1}{n+1-t}.$$

If in (2a) we let $X_1 = c$ and $X_2 = X$, we obtain

$$W(c, X) = 1 - \frac{\sum_{t=0}^{t=c}\binom{X+1}{t}\binom{\overline{N+1} - \overline{X+1}}{n+1-t}}{\binom{N+1}{n+1}}, \tag{3}$$

from which we may compute $W(X_1, X_2)$ from the equation

$$W(X_1, X_2) = W(c, X_2) - W(c, X_1 - 1).$$

[7] See also Netto, "Lehrbuch der Kombinatorik," p. 12, Eq. 6; p. 15, Eq. 11.

A direct interpretation of the expression for $W(c, X)$ gives us the following interesting

Theorem A: The *a posteriori* probability that a universe of N items contains not more than X defectives when c defectives have resulted from a random sample of n items is equal to the *a priori* probability of obtaining at least $c + 1$ defectives in a random sample of $n + 1$ items from a universe of $N + 1$ items of which exactly $X + 1$ are defective. This theorem assumes that *a priori* all values of X are equally likely.

Writing $s = X + 1 - t$, we obtain from (3)

$$1 - W(c, X) = 1 - \frac{\sum_{s=0}^{s=X-c} \binom{X+1}{s}\binom{\overline{N+1} - \overline{X+1}}{N-n-s}}{\binom{N+1}{N-n}}. \quad (4)$$

The right-hand side of this equation is exactly what would have been obtained directly from (3) if we had been dealing with a sample of $N - n - 1$ instead of n and had observed $X - c$ defectives instead of c, since the particular symbol chosen for the variable of summation is immaterial.

This fact, which follows immediately from physical consideration of the equivalent *a priori* problem of Theorem A, may be stated as

Theorem B: If we calculate the probability W that a universe of N items contains not more than X defectives when a sample of n has shown exactly c defectives, then $1 - W$ is the probability that a universe of N items contains not more than X defectives when a sample of $N - n - 1$ has shown exactly $X - c$ defectives.

In making extensive calculations, this relation will serve to cut down the amount of computation considerably, as each calculated value of W may be made to do double duty. For a single calculation either (3) or (4) may be used depending on which involves the shorter summation.

Another interesting relation also appears when we note that

$$\frac{\binom{X+1}{t}\binom{N-X}{n+1-t}}{\binom{N+1}{n+1}} = \frac{\binom{n+1}{t}\binom{N-n}{X+1-t}}{\binom{N+1}{X+1}},$$

which may be proved simply by cross multiplication of the combination factors, writing them in terms of factorials. From this we see that

$$W(c, X) = 1 - \frac{\sum_{t=0}^{t=c} \binom{n+1}{t}\binom{\overline{N+1} - \overline{n+1}}{X+1-t}}{\binom{N+1}{X+1}}. \quad (5)$$

If we compare the right-hand sides of (3) and (5), we see that n and X have simply been interchanged, which proves the rather interesting

Theorem C: The probability $W(c, X)$ that a universe of N items does not contain more than X defectives when a sample of n has shown exactly c defectives is equal to the probability $W(c, n)$ that a universe of N items does not contain more than n defectives when a sample of X has shown exactly c defectives.

Thus equations (3), (4) and (5) taken together show that we may make three different interpretations of a single calculation.

Up to this point, all of the analysis has been *exact* on the basis of the fundamental assumptions. We may now proceed with advantage to consider some approximate relationships which have been for some years of service in the calculation of practical curves and tables for cases where the values of N, n, or X were too great to be handled conveniently by means of the exact formulæ.

Now consider in formula (3) a single term, π_t say, where

$$\pi_t = \binom{X+1}{t} \frac{\binom{N-X}{n+1-t}}{\binom{N+1}{n+1}}$$

$$= \binom{X+1}{t} \frac{\left(\frac{(n+1)!}{(n+1-t)!}\right)\left(\frac{(N-n)!}{(N-n-X-1+t)!}\right)}{\left(\frac{(N+1)!}{(N-X)!}\right)}$$

$$= \binom{X+1}{t}\left(\frac{n}{N}\right)^t\left(1-\frac{n}{N}\right)^{X+1-t} \cdot F(N, n, X, t), \qquad (6)$$

where the form of the function F is to be determined.

To facilitate the consideration of this function we may split it up into three similar parts as follows:

$$F(N, n, X, t) = \frac{\varphi(n+1, t-1)\cdot\chi(N-n, X-t)}{\varphi(N+1, X)},$$

where

$$\varphi(n+1, t-1) = \left(1+\frac{1}{n}\right)\left(1\right)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{t-2}{n}\right),$$

$$\varphi(N+1, X) = \left(1+\frac{1}{N}\right)\left(1\right)\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{X-1}{N}\right),$$

$$\chi(N-n, X-t) = \left(1\right)\left(1-\frac{1}{N-n}\right)\cdots\cdots\cdots\left(1-\frac{X-t}{N-n}\right).$$

Recalling that

$$\log (1 + X) = X - \frac{1}{2}X^2 + \frac{1}{3}X^3 - \frac{1}{4}X^4 + \cdots,$$

which converges for $X^2 < 1$,

we have

$$\log \varphi(n + 1, t - 1) = \log\left(1 + \frac{1}{n}\right) - \sum_{y=0}^{t-2}\sum_{r=1}^{\infty}\frac{1}{r}\left(\frac{y}{n}\right)^r,$$

$$\log \varphi(N + 1, X) = \log\left(1 + \frac{1}{N}\right) - \sum_{y=0}^{X-1}\sum_{r=1}^{\infty}\frac{1}{r}\left(\frac{y}{N}\right)^r,$$

$$\log \chi(N - n, X - t) = -\sum_{y=0}^{X-t}\sum_{r=1}^{\infty}\frac{1}{r}\left(\frac{y}{N-n}\right)^r,$$

whence

$$\log F(N, n, X, t) = \log\left(1 + \frac{1}{n}\right) - \log\left(1 + \frac{1}{N}\right) + \sum_{y=0}^{X-1}\sum_{r=1}^{\infty}\frac{1}{r}\left(\frac{y}{N}\right)^r - \sum_{y=0}^{t-2}\sum_{r=1}^{\infty}\frac{1}{r}\left(\frac{y}{n}\right)^r - \sum_{y=0}^{X-t}\sum_{r=1}^{\infty}\frac{1}{r}\left(\frac{y}{N-n}\right)^r$$

Neglecting terms of the second order in $1/n$, $1/N$ and $1/(N - n)$, we have as an approximation

$$\log F(N, n, X, t) \doteq -\frac{1}{2}\left(\frac{t^2-3t}{n} + \frac{X^2-2Xt+t^2+X-t}{N-n} - \frac{X^2-X-2}{N}\right).$$

If now we select as our value of t, $t = (n/N)X$, we have $\log F(N, n, X, t) \doteq (X - 2)/2N$ which is > 0; if $X > 2$,

$$F(N, n, X, t) \doteq e^{(X-2)/2N},$$

which gives us as an approximate value for the maximum term π_t, where $t = (n/N)X$,

$$\pi_t' = \left(\frac{X+1}{t}\right)\left(\frac{n}{N}\right)^t\left(1 - \frac{n}{N}\right)^{X+1-t} e^{(X-2)/2N}. \tag{7}$$

Having this term, it is a simple matter to calculate the other terms necessary for evaluating $W(c, X)$ by means of the exact equations

$$\pi_{t+1} = \pi_t\left(\frac{n-t+1}{t+1}\cdot\frac{X-t+1}{N-n-X+t}\right), \tag{8}$$

$$\pi_{t-1} = \pi_t\left(\frac{t}{n-t+2}\cdot\frac{N-n-X+t-1}{X-t+2}\right). \tag{9}$$

Due to the reciprocal relationship between n and X, we may obtain in a similar manner

$$\binom{n+1}{t}\frac{\binom{N-n}{x+1-t}}{\binom{N+1}{X+1}}=\binom{n+1}{t}\left(\frac{X}{N}\right)^{t}\left(1-\frac{X}{N}\right)^{n+1-t}e^{(n-2)/2N}, \quad (10)$$

when $t = (X/N)n$.

It is by means of these relationships that we have calculated the cases for $N > 1{,}000$ as shown on Charts C and feel that the precision obtained is rather better than would have resulted from using formula (6) for all values of t and assuming $F(N, n, X, t) = 1$. However, for suitable ranges of the variables involved, the formula resulting from this procedure

$$W(c, X) \doteq 1 - \sum_{t=0}^{t=c}\binom{X+1}{t}\left(\frac{n}{N}\right)^{t}\left(1-\frac{n}{N}\right)^{X+1-t} \quad (11)$$

would be a fairly good approximation. This is simply part of the well-known binomial expansion and is far simpler to compute than the more precise formulæ, although by no means easy at that.

We may draw several interesting practical conclusions, however, from formula (11). For instance, we may note that as n/N approaches 0 and X becomes infinite in such a way that the product $(X + 1)(n/N)$ remains constant and equal to the average a, we have the familiar Poisson Exponential Binomial Limit

$$W(c, X) = 1 - \sum_{t=0}^{c}\frac{a^{t}e^{-a}}{t!},$$

where $a = (X + 1)(n/N)$.

In addition we note from formula (11) that, for small values of X/N, the variable N enters into the formula only in the ratio n/N. From this we deduce the fact, borne out by independent calculations, that by means of the proper use of a proportionality factor K applied directly to n and N and inversely to X/N we may extend the Charts C to care for values of $X/N \leq .1$ to a very good degree of accuracy and with considerable saving in space and computational labor.

By the reciprocal relationship between X and n as shown in exact formulæ (3) and (5), we obtain

$$W(c, X) \doteq 1 - \sum_{t=0}^{t=c}\binom{n+1}{t}\left(\frac{X}{N}\right)^{t}\left(1-\frac{X}{N}\right)^{n+1-t}, \quad (12)$$

which differs only in form from equation (3) of Molina's paper [8] on

[8] Footnote 1.

the infinite universe case. Formula (12) does not give the same results as (11) as it is most exact when n/N is small and becomes absolutely exact in the limiting case of an infinite universe where $n/N \doteq 0$. This formula also approaches the Poisson Limit, in this case as X/N approaches 0 and $n + 1$ becomes infinite in such a way that the product $(n + 1)(X/N)$ remains constant and equal to α, say.

The Poisson Limit, for the case of an infinite universe, was given by Molina in the Appendix to the article in the *Bell System Technical Journal* of January, 1924, already mentioned in this memorandum.

Another point of interest is brought out when we note that in the limiting form of (12) the Poisson gives us

$$W(c, X) \doteq \sum_{t=c+1}^{\infty} \frac{a^t e^{-a}}{t!}, \quad a = \frac{X \cdot n}{N},$$

and for another pair of values of W and X

$$W_1(c, X_1) = \sum_{t=c+1}^{\infty} \frac{a_1{}^t e^{-a_1}}{t!}, \quad a_1 = \frac{X_1 \cdot n}{N}.$$

Thus from properly chosen Poisson curves or tables we may obtain the ratio $X_1/X \doteq a_1/a$ which corresponds to the observed value of c and the desired values of W and W_1. This ratio in exact formulæ is a function of N, n, and X also, but for many problems involving small values of n/N and X/N the degree of approximation furnished by this limiting form is fairly satisfactory and still further reduces the amount of labor necessary in extending approximate results to practice.

The sort of procedure we have just been discussing may be facilitated by means of a chart on which we show as abscissæ values of c and as ordinates values of the ratio of X_1/X which corresponds to various values of W as shown by various curves and a specified value of W_1, say $W_1 = .9$. Such a chart would enable us to interpret roughly a given Chart C for $W = .9$ in terms of other values of W. For precise work this procedure is not to be recommended, and, therefore, no charts of the nature just described are included herein.

Approximations to the binomial other than the Poisson have been discussed in many of the texts. In particular, for values of p in the neighborhood of $\frac{1}{2}$, the well-known Laplace-Bernoulli integral

$$\frac{1}{\sqrt{\pi}} \int_a^b e^{-t^2} dt$$

will serve as an approximate value for W_1 where the limits a and b

are functions of N, n, X, and c. This approximation is not so suitable, however, for most telephone sampling problems in which the proportion of defectives may be assumed in general to be far smaller than $\frac{1}{2}$.

We shall now proceed to discuss a few points concerning the analysis of Case II in which instead of assuming $w(X)$ constant we assumed it to be of the form

$$w(X) = \binom{N}{X} p^X (1 - p)^{N-X}.$$

Combining this expression for $w(X)$ with the term $\binom{X}{c}\binom{N-X}{n-c}$ which appears in the basic formula (1), we have

$$\frac{X!}{c!(X-c)!} \cdot \frac{(N-X)!}{(n-c)!(N-X-n+c)!} \frac{N!}{X!(N-X)!} p^X (1-p)^{N-X}$$
$$= \binom{N}{n} \cdot \binom{n}{c} p^c (1-p)^{n-c} \cdot \left(\binom{N-n}{X-c} p^{X-c} (1-p)^{N-n-X+c} \right).$$

Since only the factors in brackets involve the variable of summation X, the remainder of this expression will cancel out in numerator and denominator, leaving us with

$$W(X_1, X_2) = \frac{\sum_{X=X_1}^{X=X_2} \binom{N-n}{X-c} p^{X-c} (1-p)^{N-n-X+c}}{\sum_{X=c}^{X=N-n+c} \binom{N-n}{X-c} p^{X-c} (1-p)^{N-n-X+c}}$$

as the resulting form for (1) with this assumption for $w(X)$.

It may be noted that the summation in the denominator above is a complete binomial $(p + q)^{N-n}$ and as such equals unity, so

$$W(X_1, X_2) = \sum_{X=X_1}^{X=X_2} \binom{N-n}{X-c} p^{X-c} (1-p)^{N-n-X+c},$$

where p is assumed to be the *a priori* probability of a defective item as determined from reliable information concerning conditions under which the items are prepared.

As before when $X_1 = c$ and $X_2 = X$ we have

$$W(c, X) = \sum_{t=0}^{t=X-c} \binom{N-n}{t} p^t (1-p)^{N-n-t}.$$

We may be willing in certain cases to admit the binomial form for

$w(X)$ without being able or willing to assign any single value to p. In such cases we may, however, proceed to make assumptions concerning the probability that p has a given value. Let

$$s(X, p) = f(p)\binom{N}{X} p^X(1 - p)^{N-X};$$

then

$$w(X) = \int_0^1 s(x, p)dp = \binom{N}{X}\int_0^1 f(p)p^X(1 - p)^{N-X}dp,$$

where

$$\int_0^1 f(p)dp = 1 \quad \text{and} \quad \sum_{X=0}^{N} w(X) = 1.$$

Suppose we assume $f(p)$ constant for all values between 0 and 1; we have

$$w(X) = \binom{N}{X}\int_0^1 p^X(1 - p)^{N-X}dp = \frac{1}{N + 1},$$

which we note to be a constant which assigns to all of the $N + 1$ possible *a priori* hypotheses concerning X an equal weight. This pair of assumptions in Case II amounts, therefore, to the same thing analytically as the assumption of Case I.

Any number of possible hypotheses concerning $f(p)$ might be made. Some of these would complicate the analysis considerably, others might be carried through fairly simply. One of these hypotheses might fit one class of physical problems, another some other class. To consider these all in detail in this paper would be outside of the scope of a general treatment. The methods outlined here would, however, hold for such extensions. Such difficulties as might be encountered would be of an analytical rather than a logical nature.

In closing, the author wishes to express his appreciation to his numerous friends and associates in the Bell System, whose suggestions and cooperation have been of material assistance in the preparation of this work, and particularly the work of Miss Nelliemae Z. Pearson of the Department of Development and Research, under whose direction most of the computations were carried out and who has checked through the various proofs.

Key to the Charts

The charts present various graphical representations of the function $W(c, X, n, N)$, equation 3. This function gives the probability, W, that the number of defectives in a lot of N is equal to *or less than* X, after a sample of n units has shown c defectives, *assuming that each of the possible values of* X *between* 0 *and* N *were equally likely a priori.*

Charts A: Separate pages refer to different values of n and N as labelled.

Ordinates, W; abscissas, X.

Solid curves, c; dotted curves $(c/n - X/N)$ expressed as per cent.

Charts B: Separate groups of curves refer to different values of W as labelled.

Ordinates, X/N; abscissas, n.

Separate sets of curves in each group refer to different values of c/n as labelled.

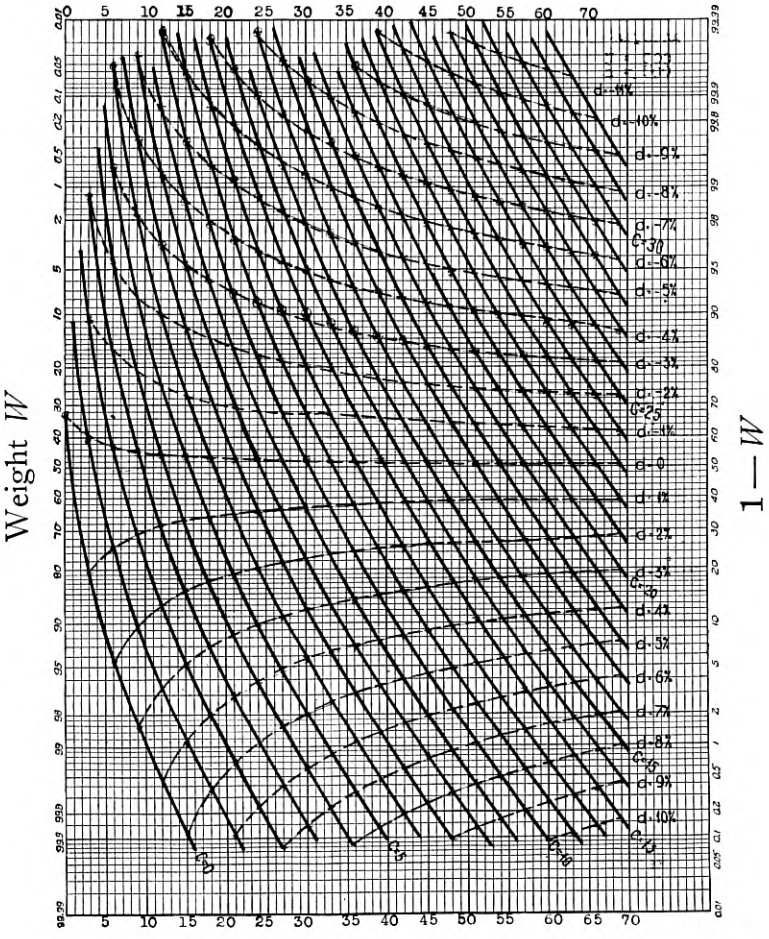
Individual curves are for different values of N.

Charts C: Separate pages refer to different values of W.

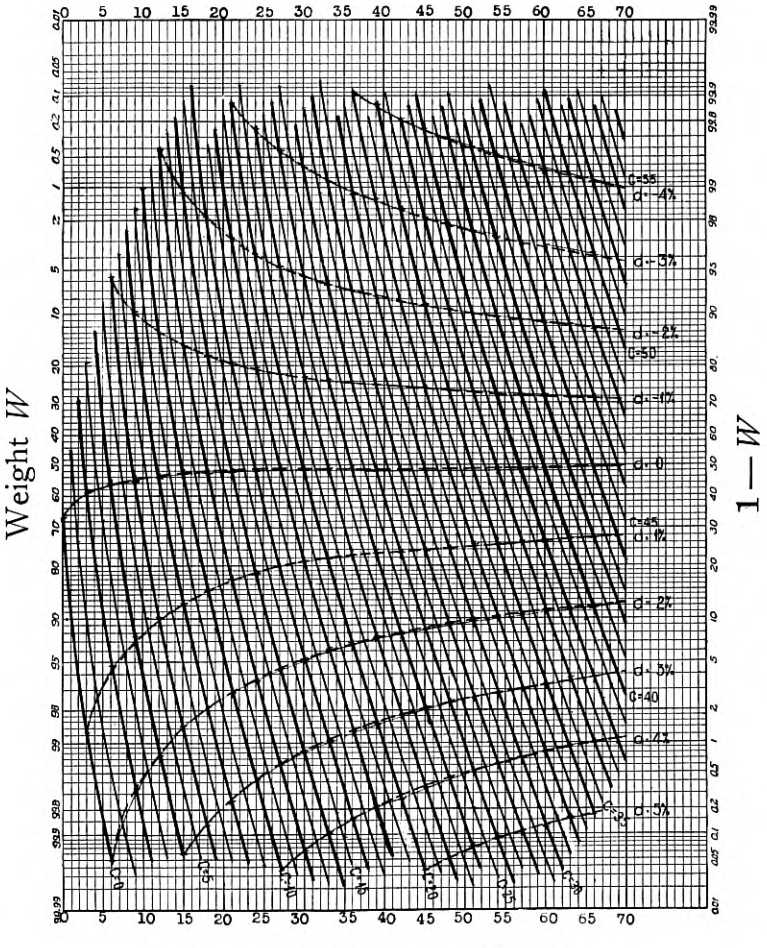
Ordinates, n; abscissas, n/N.

Separate curves for different values of c.

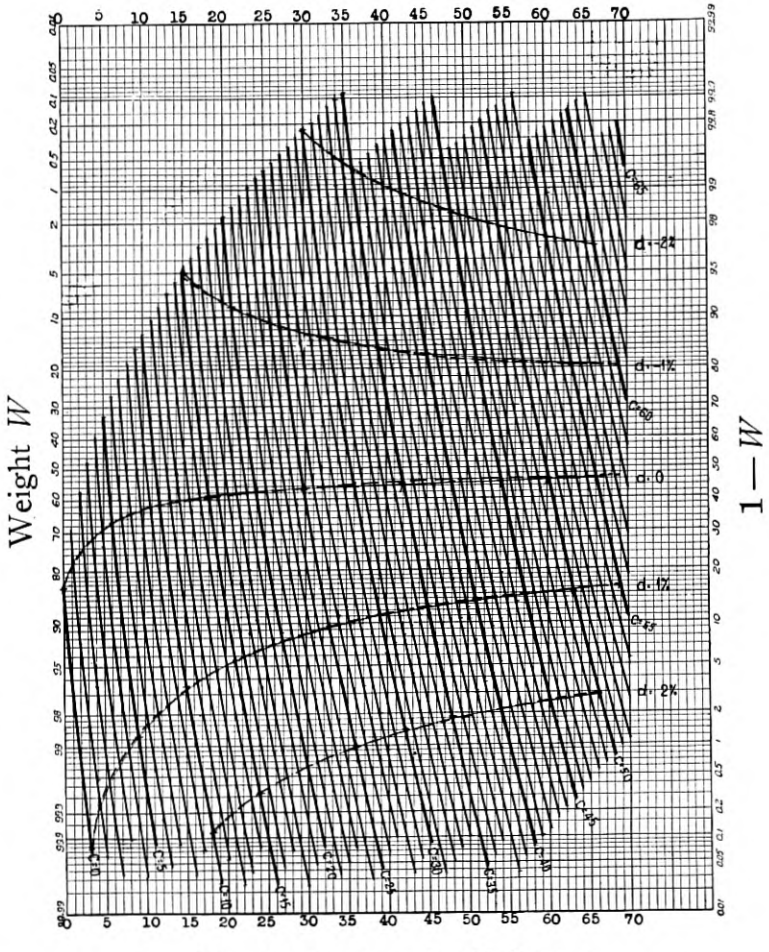
Cross-hatching indicates amount of dependence on X/N. For fuller explanation see pages 34–37 incl.



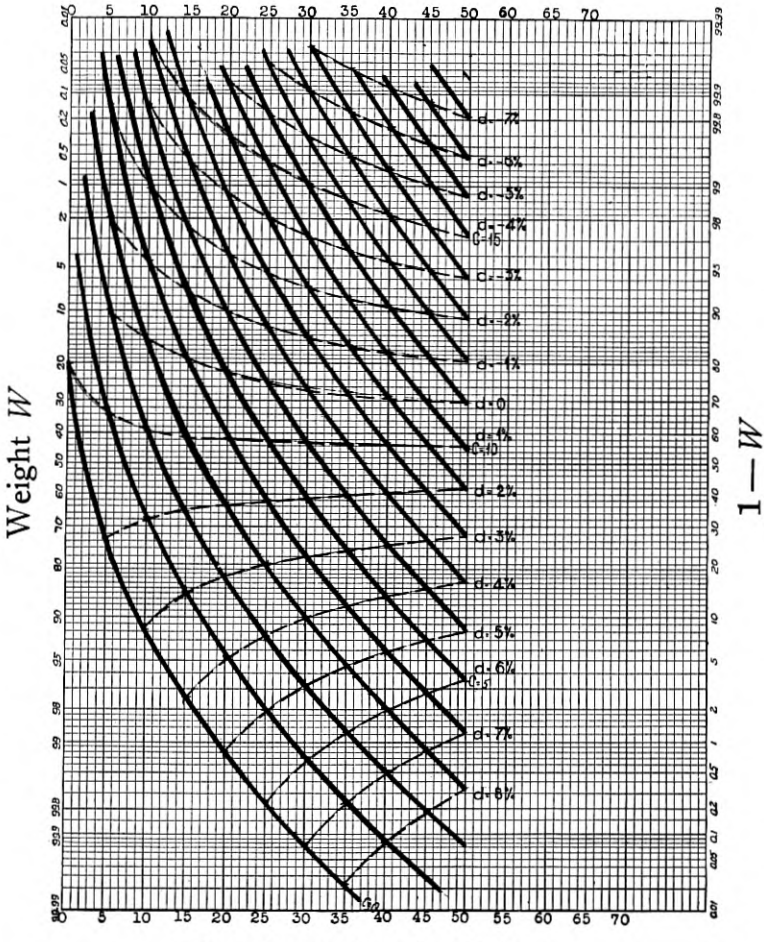
$X =$ Defectives in Universe
$N = 300,\ n = 100$
CHARTS A



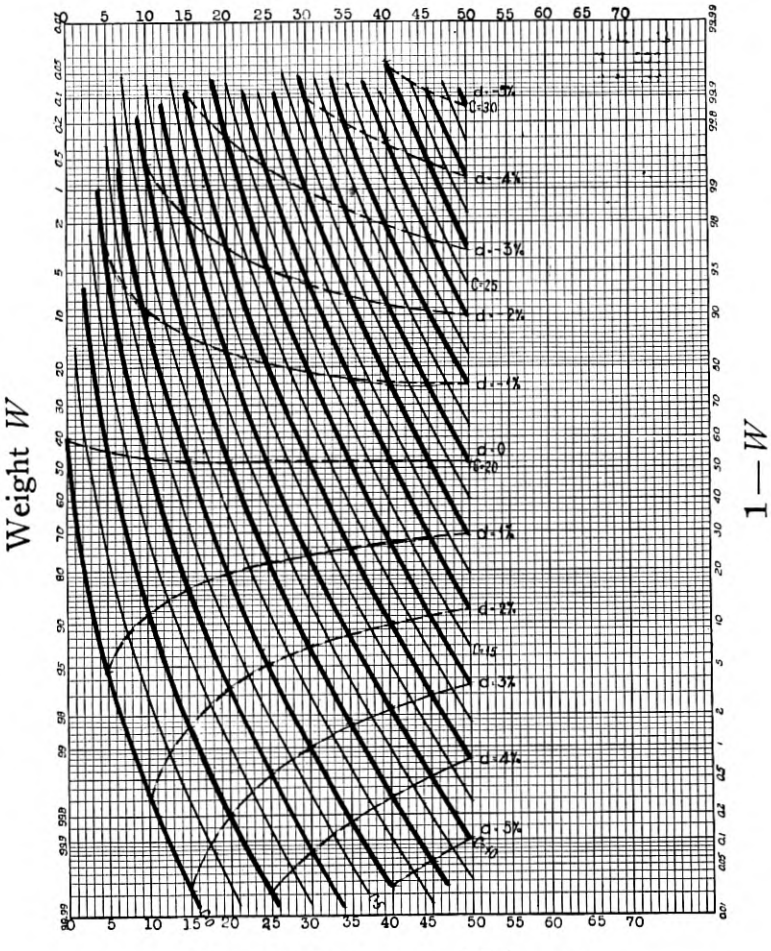
$X =$ Defectives in Universe
$N = 300,\ n = 199$
CHARTS A



$X =$ Defectives in Universe
$N = 300,\ n = 249$
CHARTS A



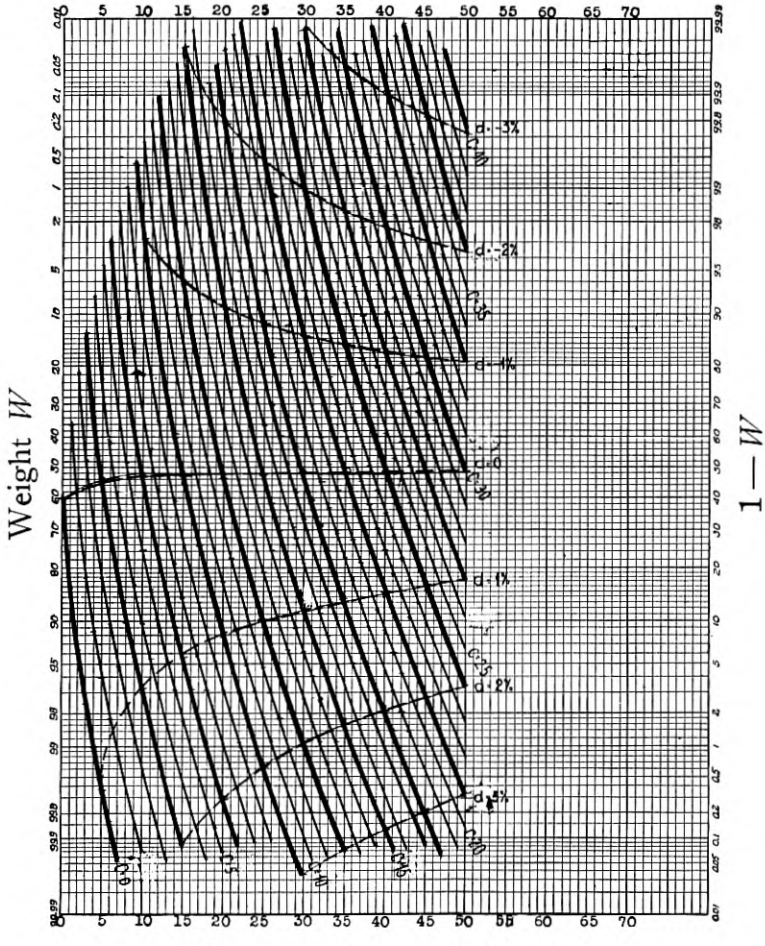
$X =$ Defectives in Universe
$N = 500,\ n = 99$
CHARTS A



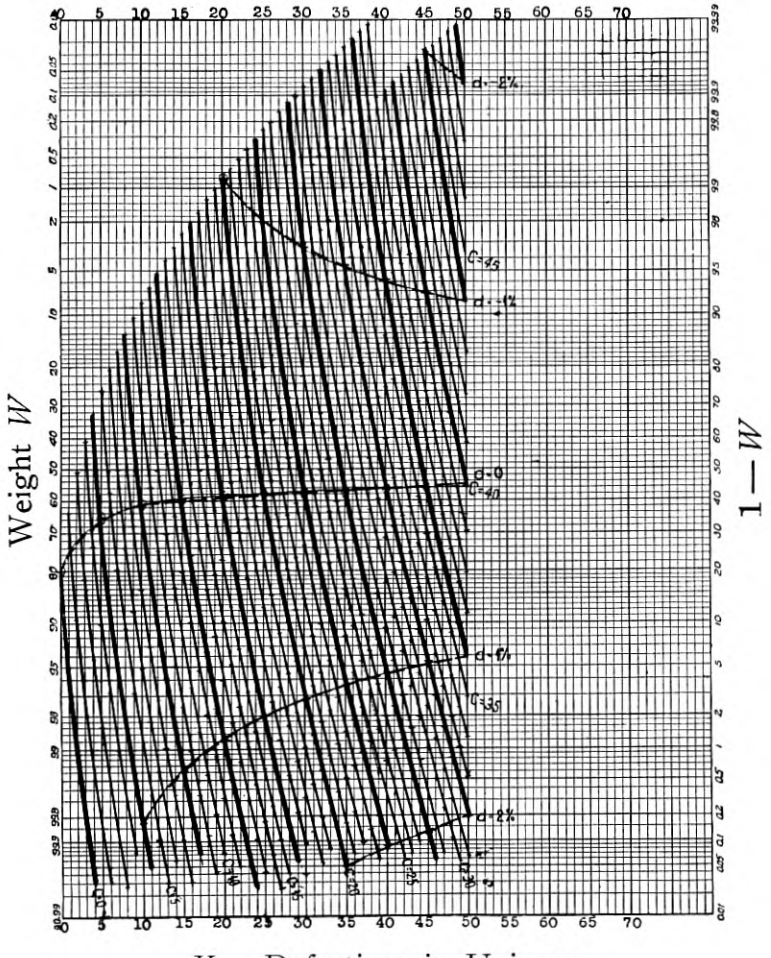
$X =$ Defectives in Universe

$N = 500,\ n = 199$

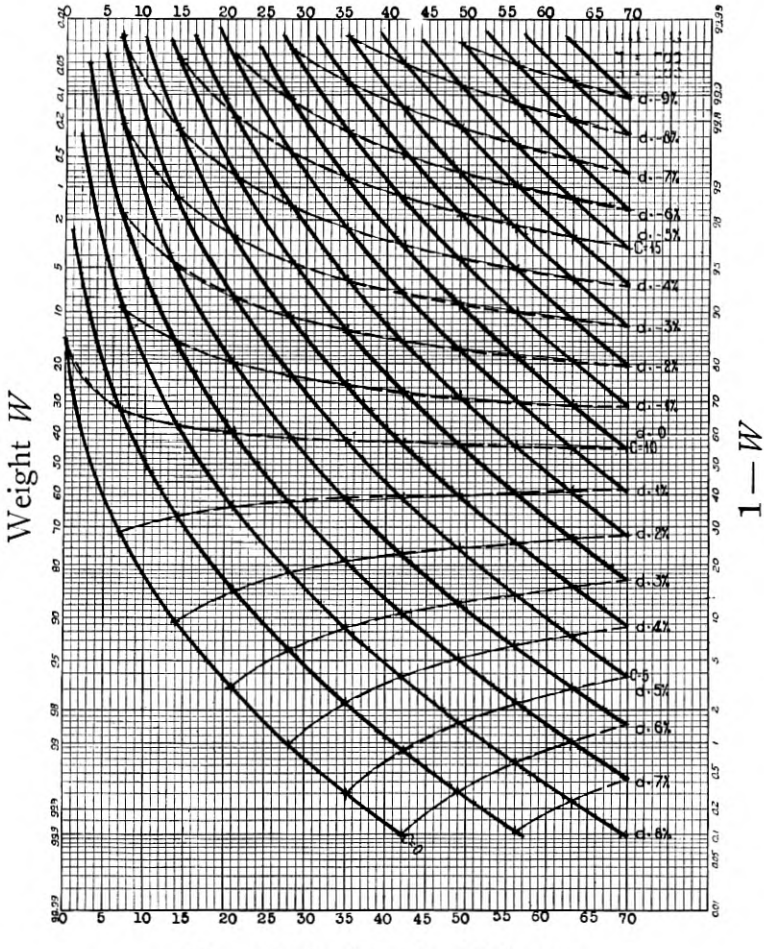
CHARTS A



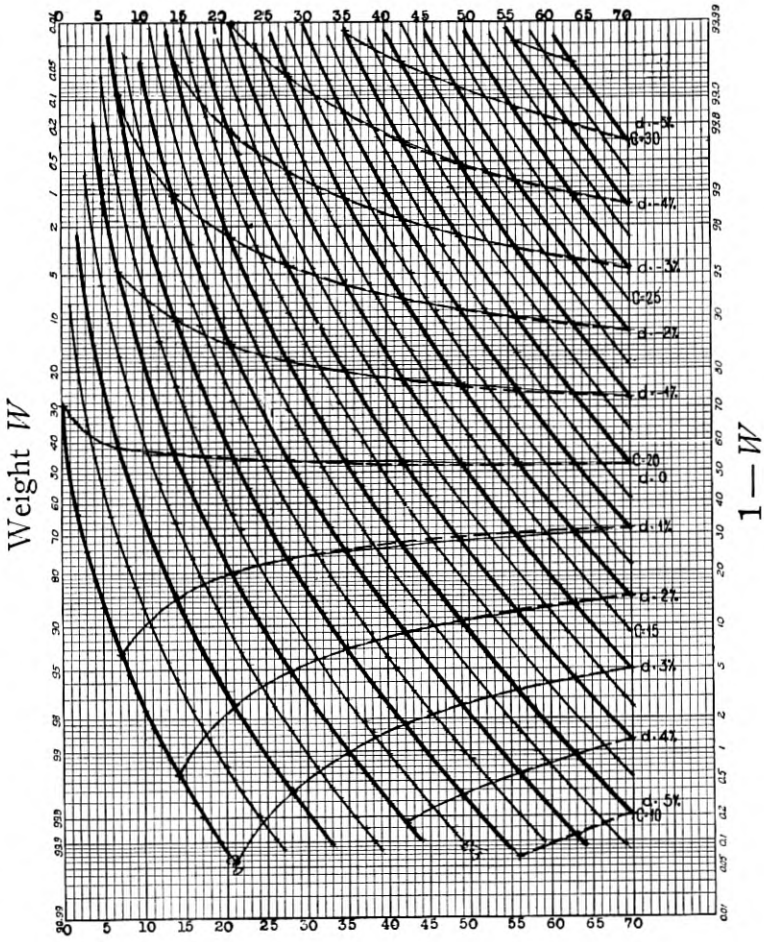
$X =$ Defectives in Universe
$N = 500,\ n = 300$
CHARTS A



CHARTS A



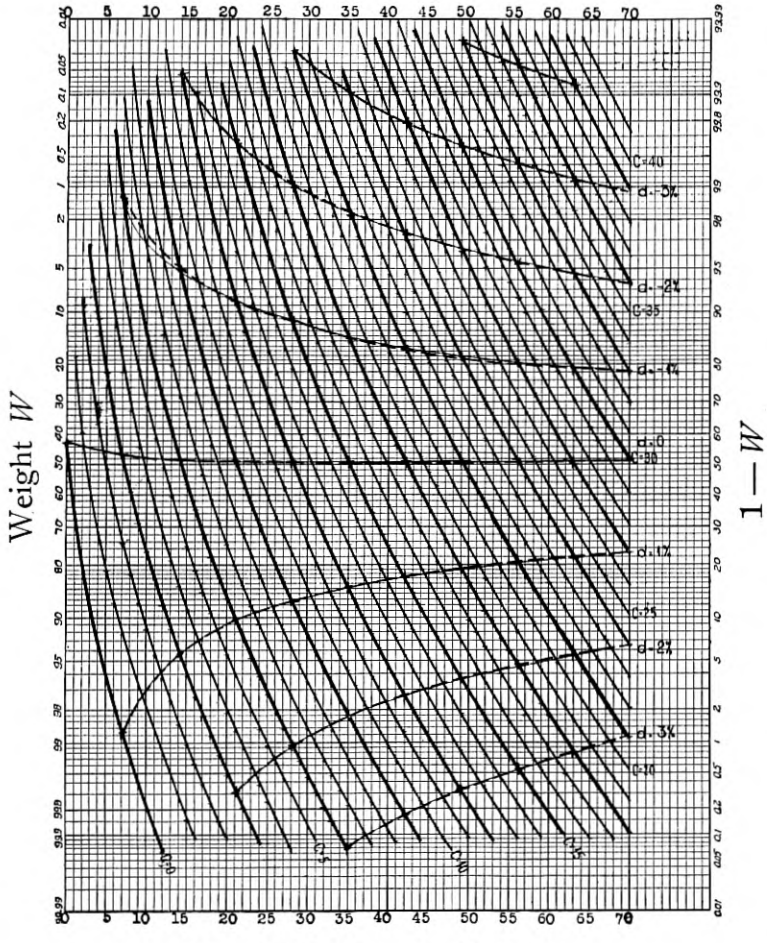
$X =$ Defectives in Universe
$N = 700,\ n = 100$
CHARTS A



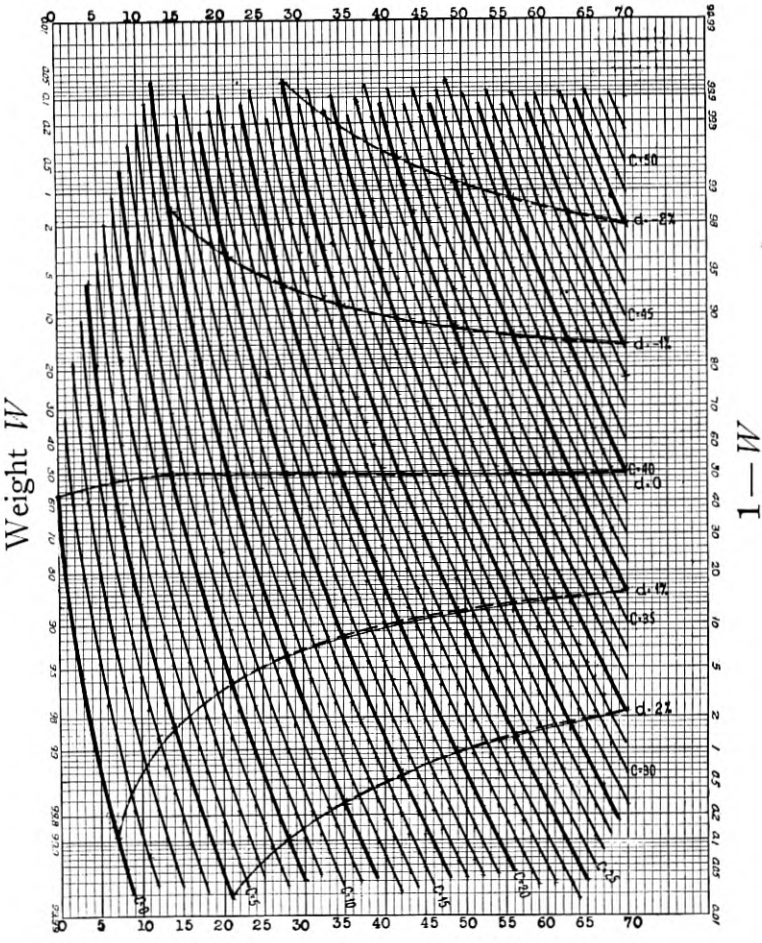
$X =$ Defectives in Universe

$N = 700,\ n = 200$

CHARTS A



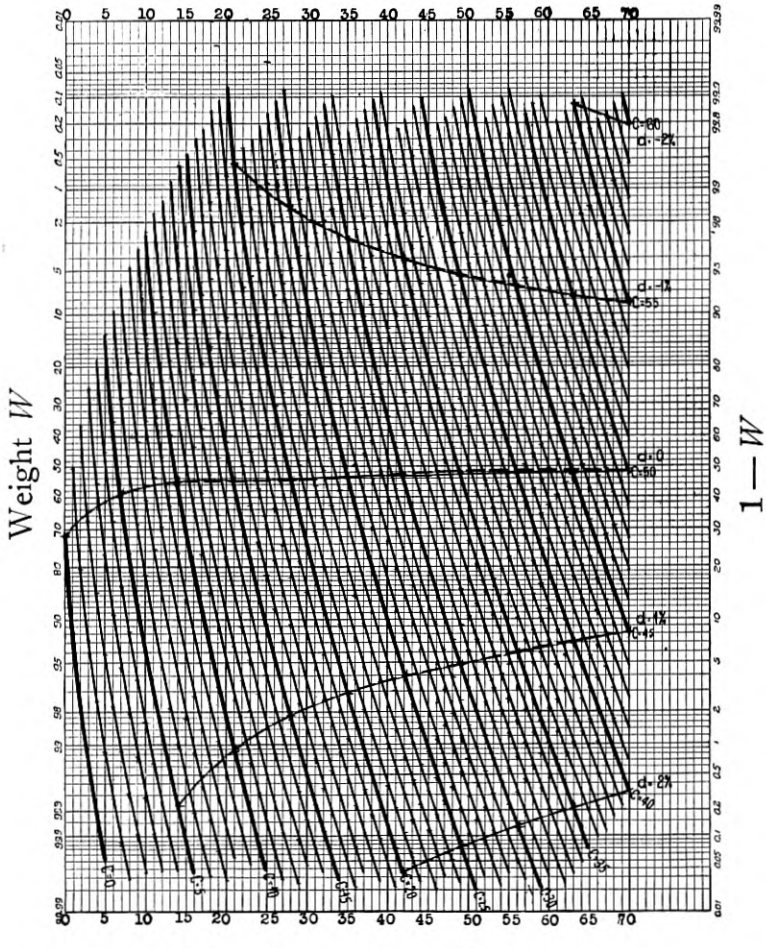
$X =$ Defectives in Universe
$N = 700,\ n = 300$
CHARTS A



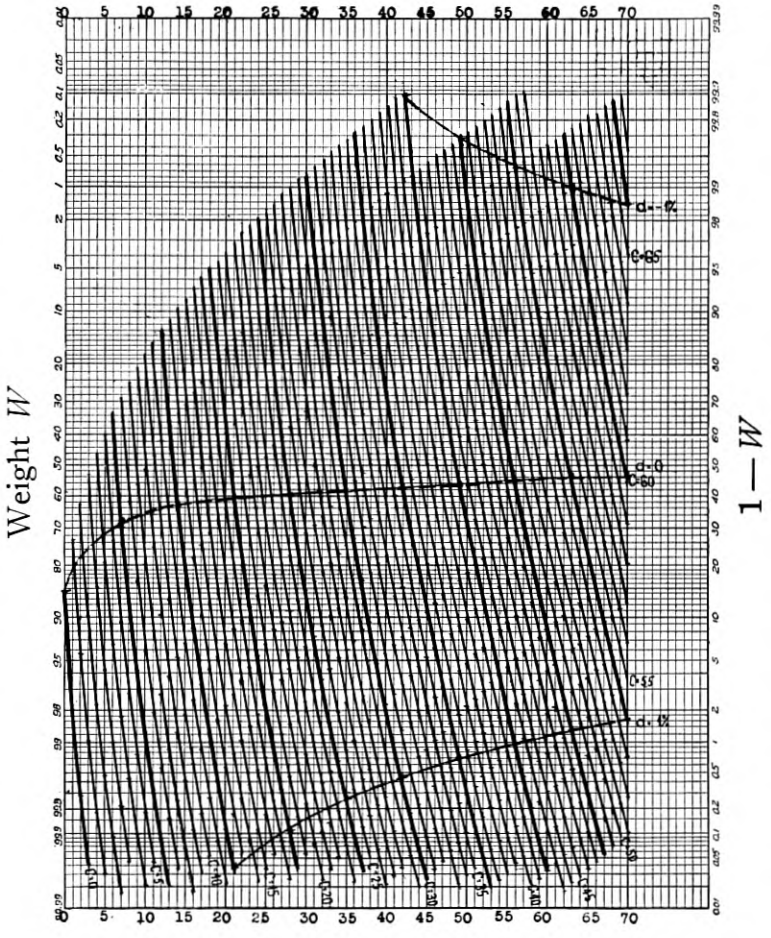
$X =$ Defectives in Universe

$N = 700,\ n = 399$

CHARTS A



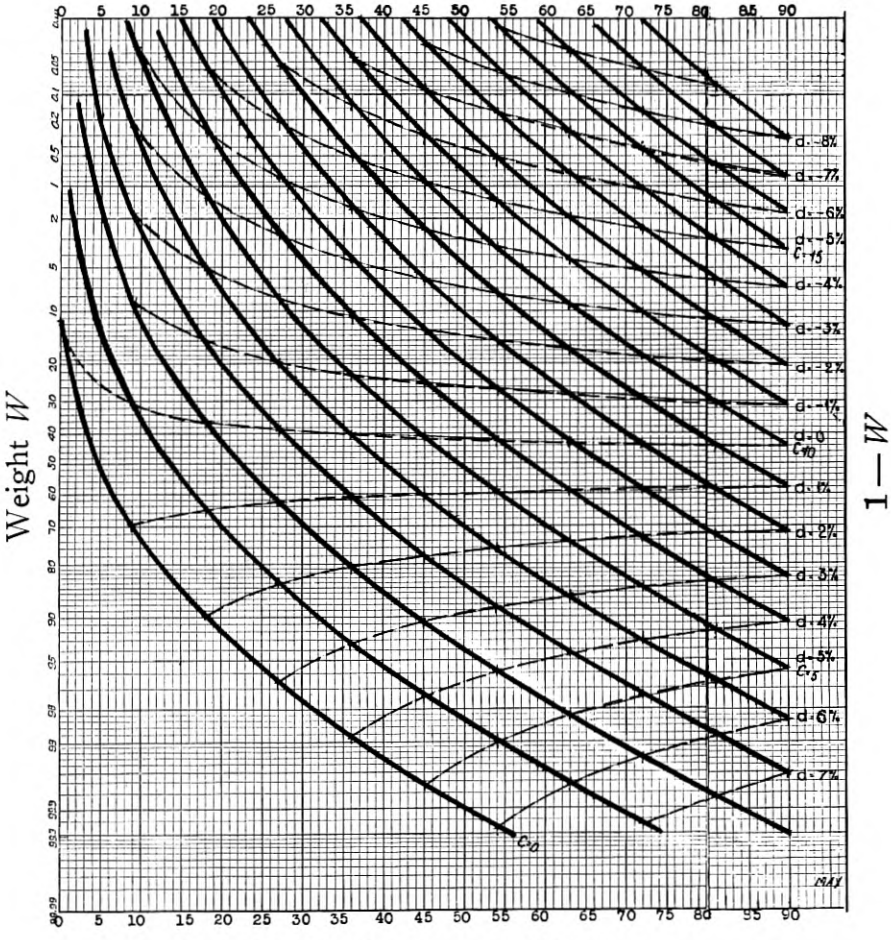
$X =$ Defectives in Universe
$N = 700,\ n = 499$
CHARTS A



$X =$ Defectives in Universe

$N = 700,\ n = 599$

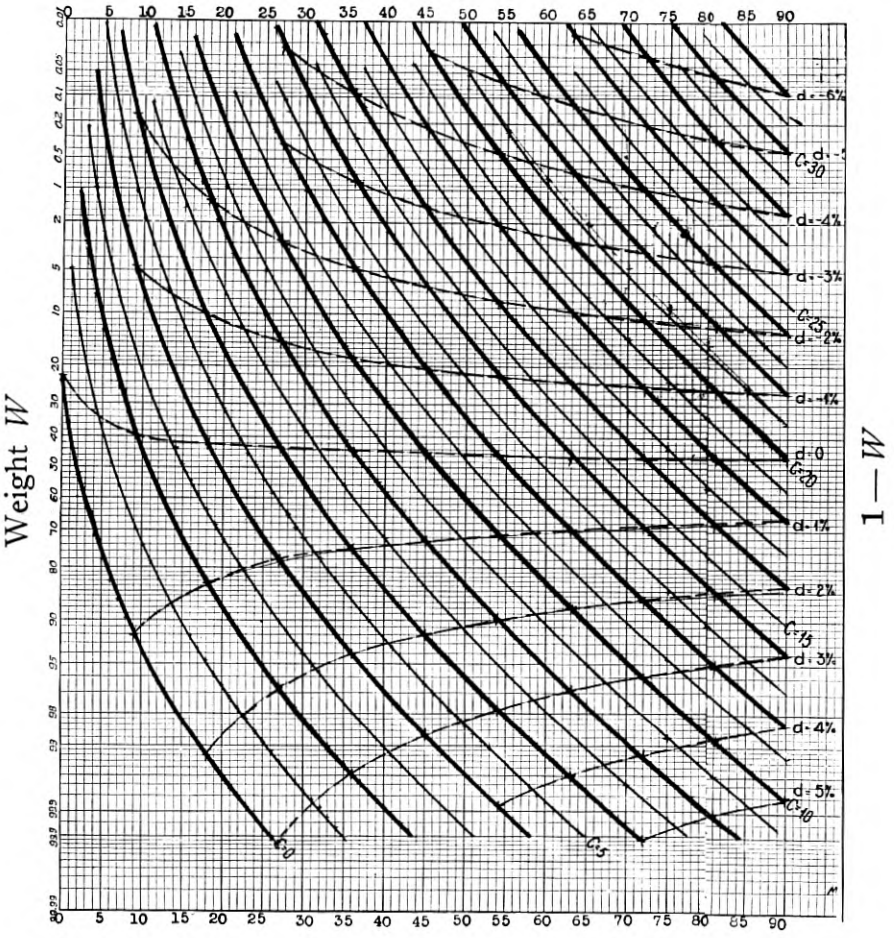
CHARTS A



$X =$ Defectives in Universe

$N = 900,\ n = 100$

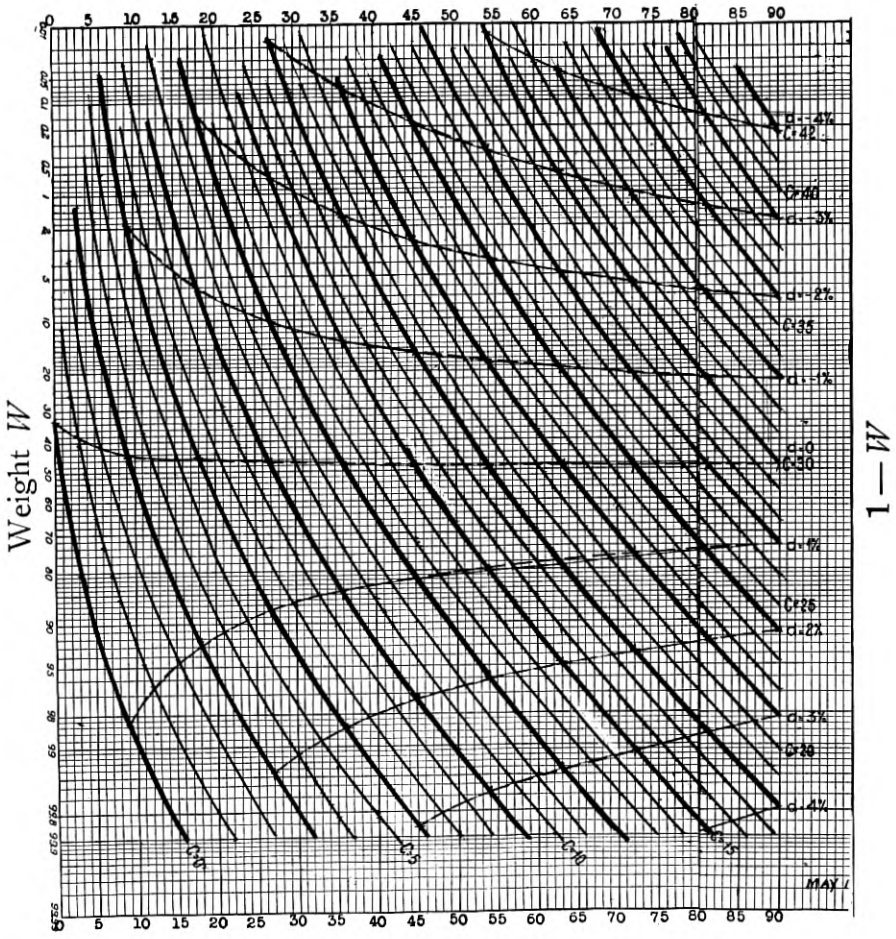
CHARTS A



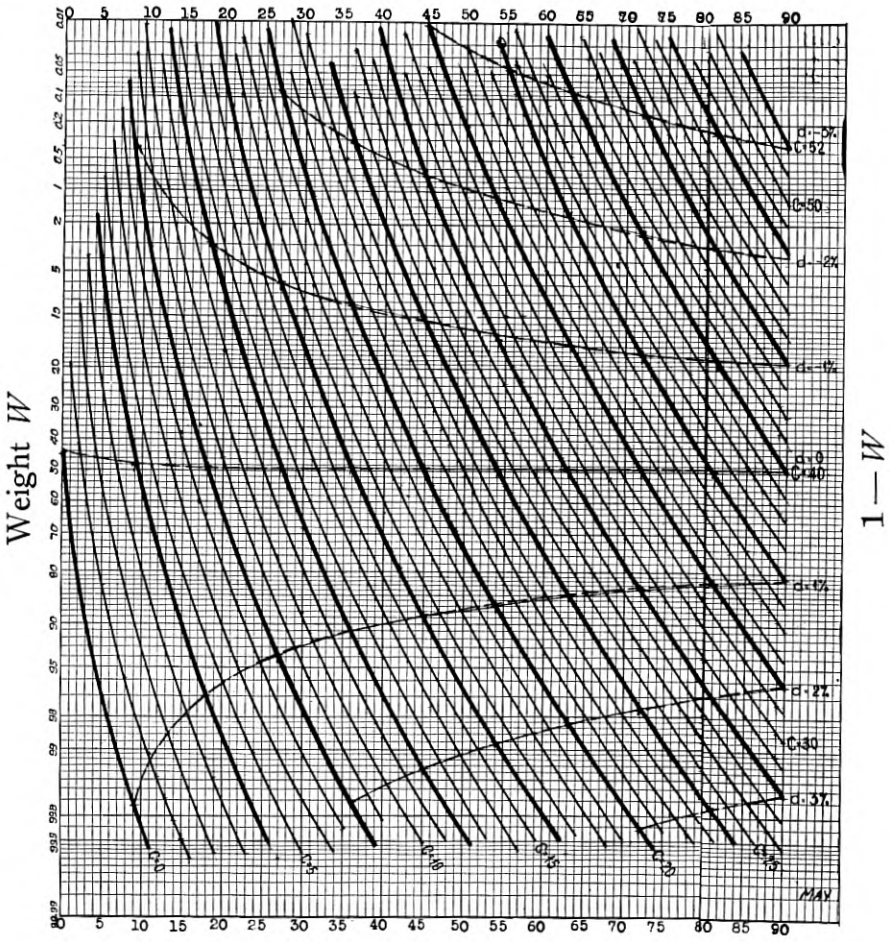
$X =$ Defectives in Universe

$N = 900,\ n = 200$

CHARTS A



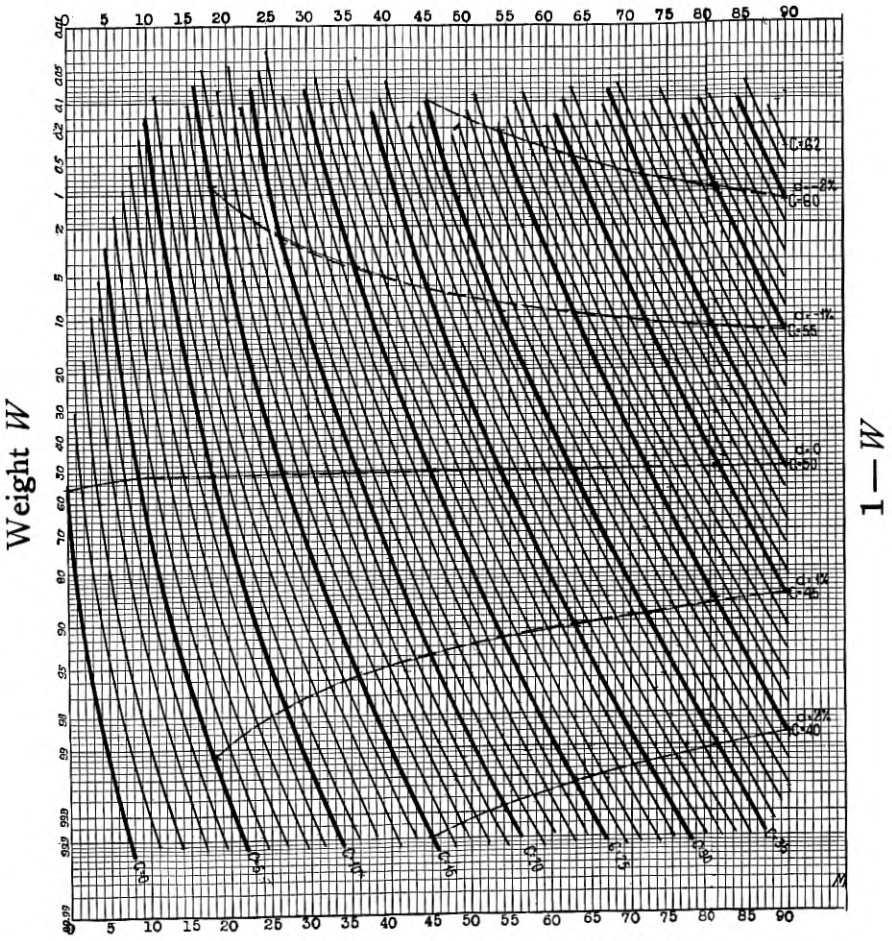
$X =$ Defectives in Universe
$N = 900,\ n = 300$
CHARTS A



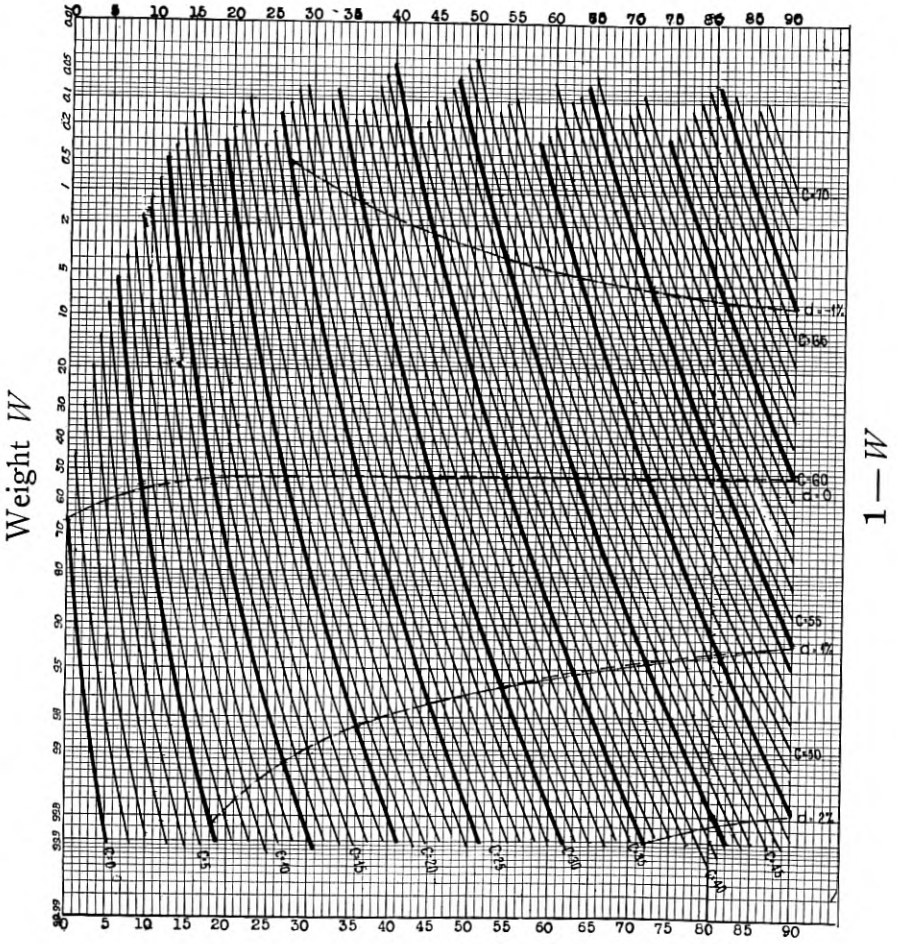
$X =$ Defectives in Universe

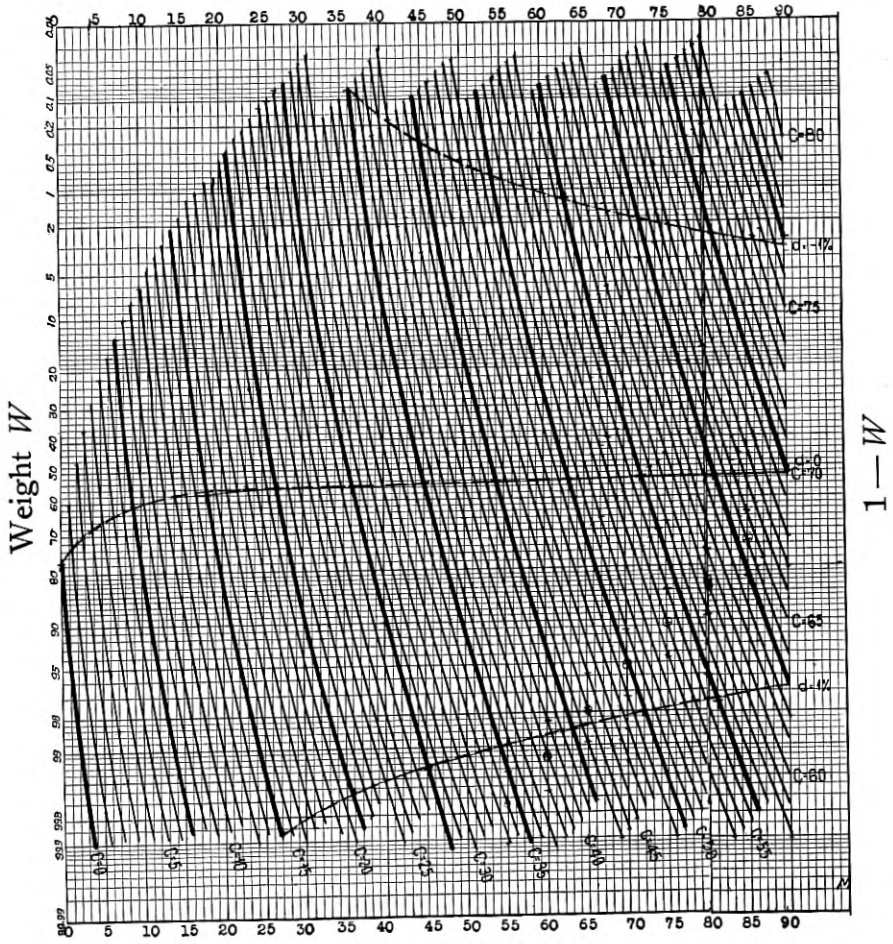
$N = 900,\ n = 400$

CHARTS A



$X =$ Defectives in Universe
$N = 900,\ n = 499$
CHARTS A

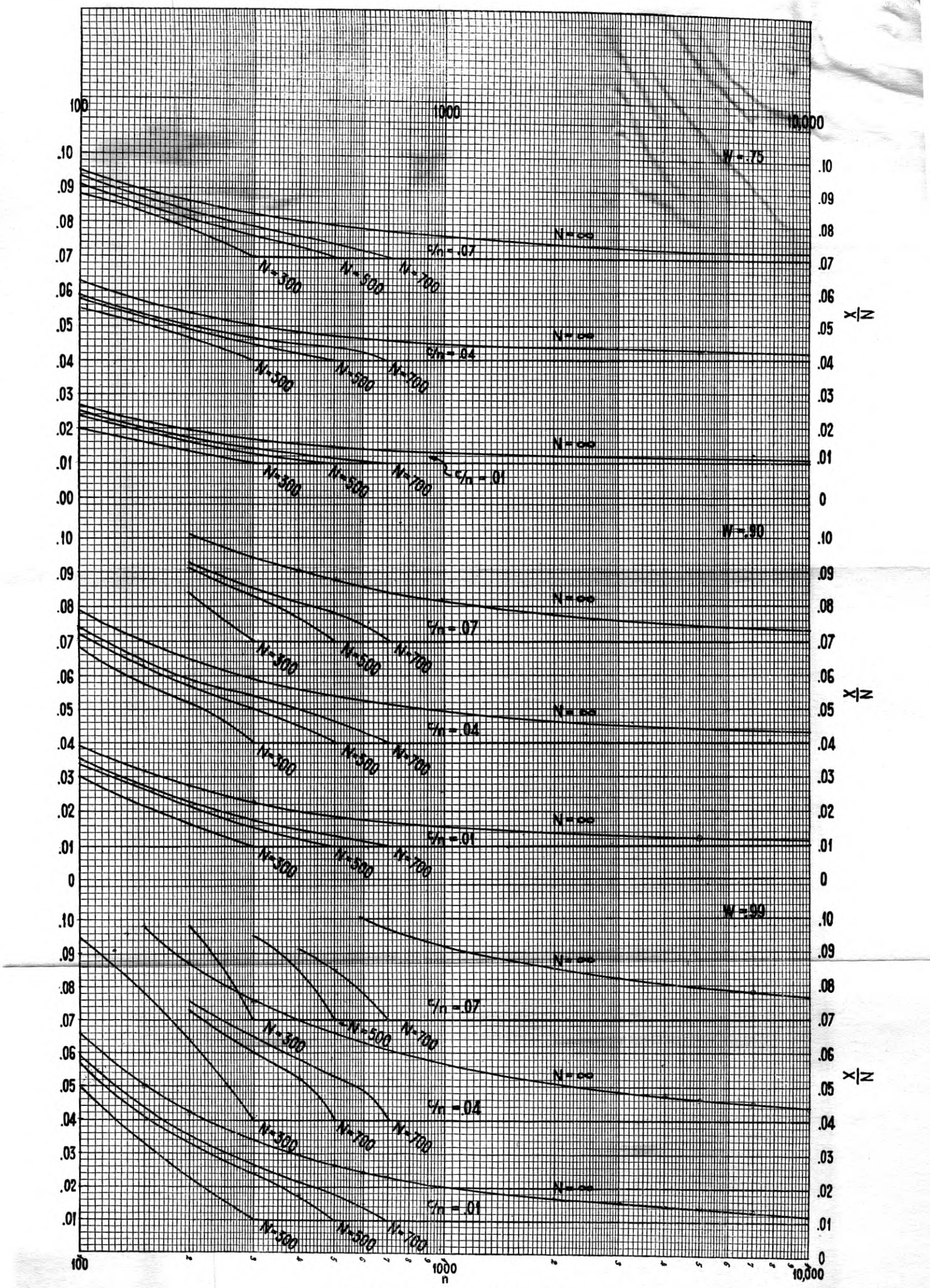

Weight W
$1-W$
X = Defectives in Universe
$N = 900,\ n = 599$
CHARTS A



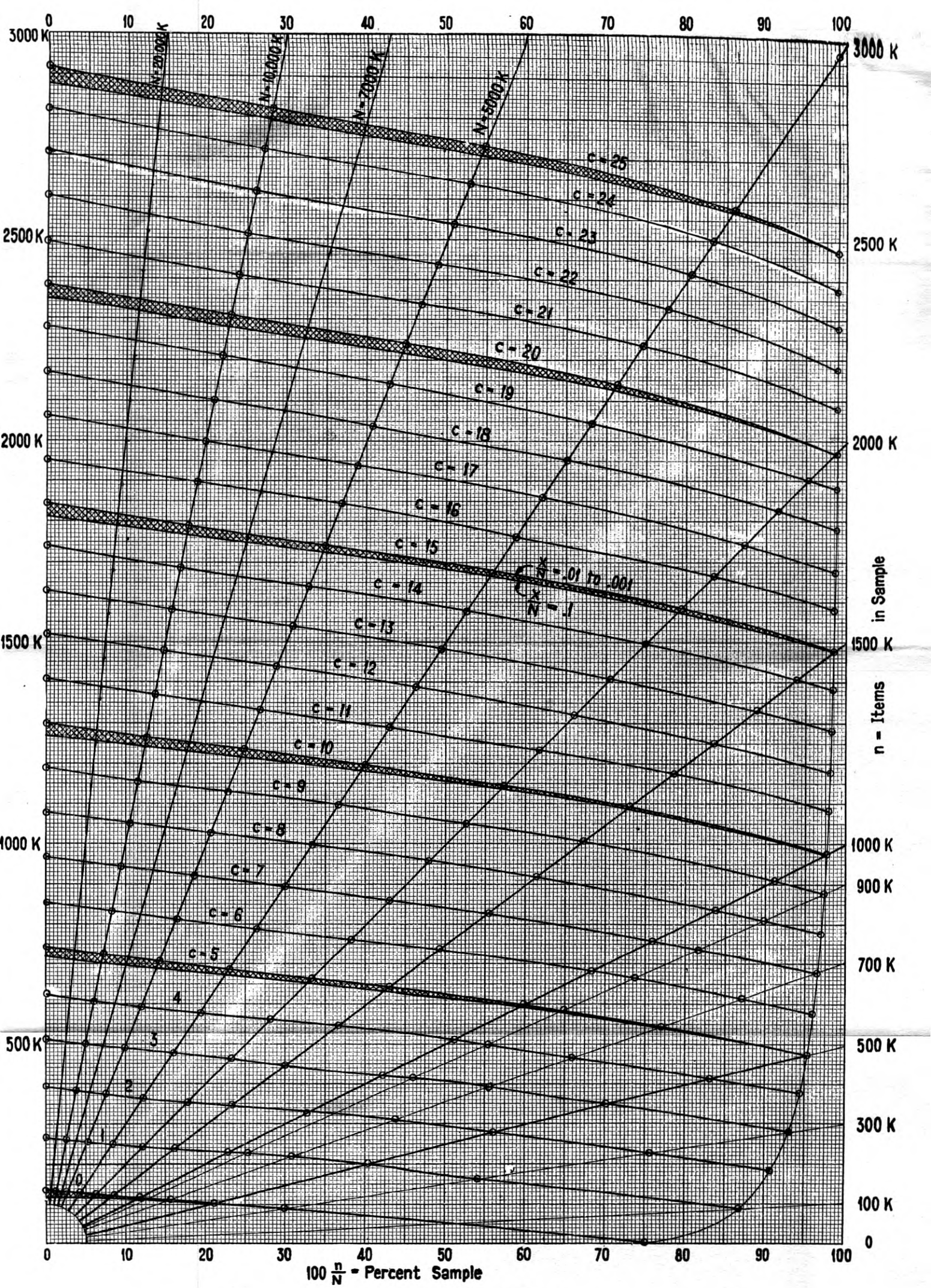
$X =$ Defectives in Universe

$N = 900,\ n = 699$

CHARTS A

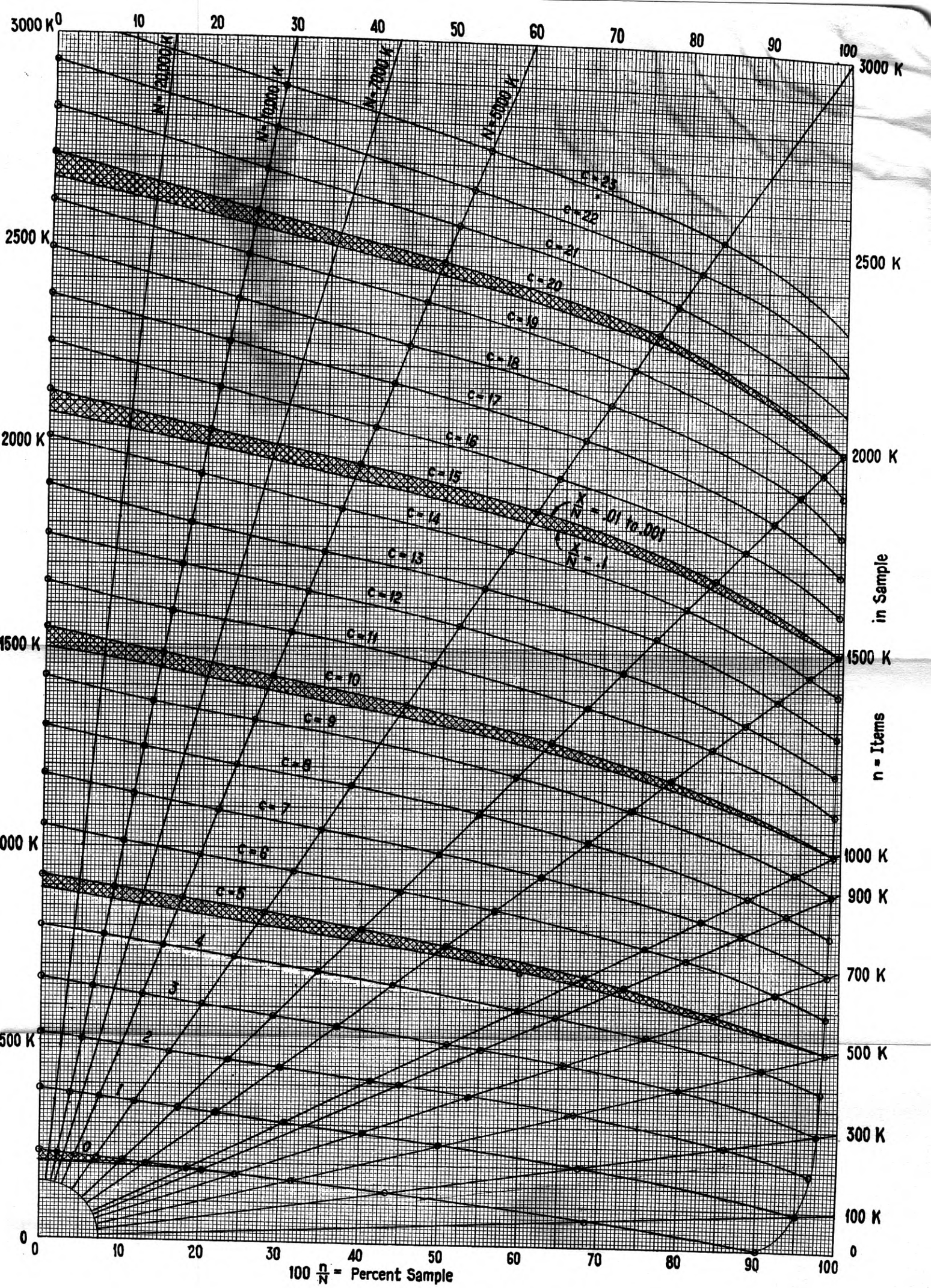


SAMPLING CHART B



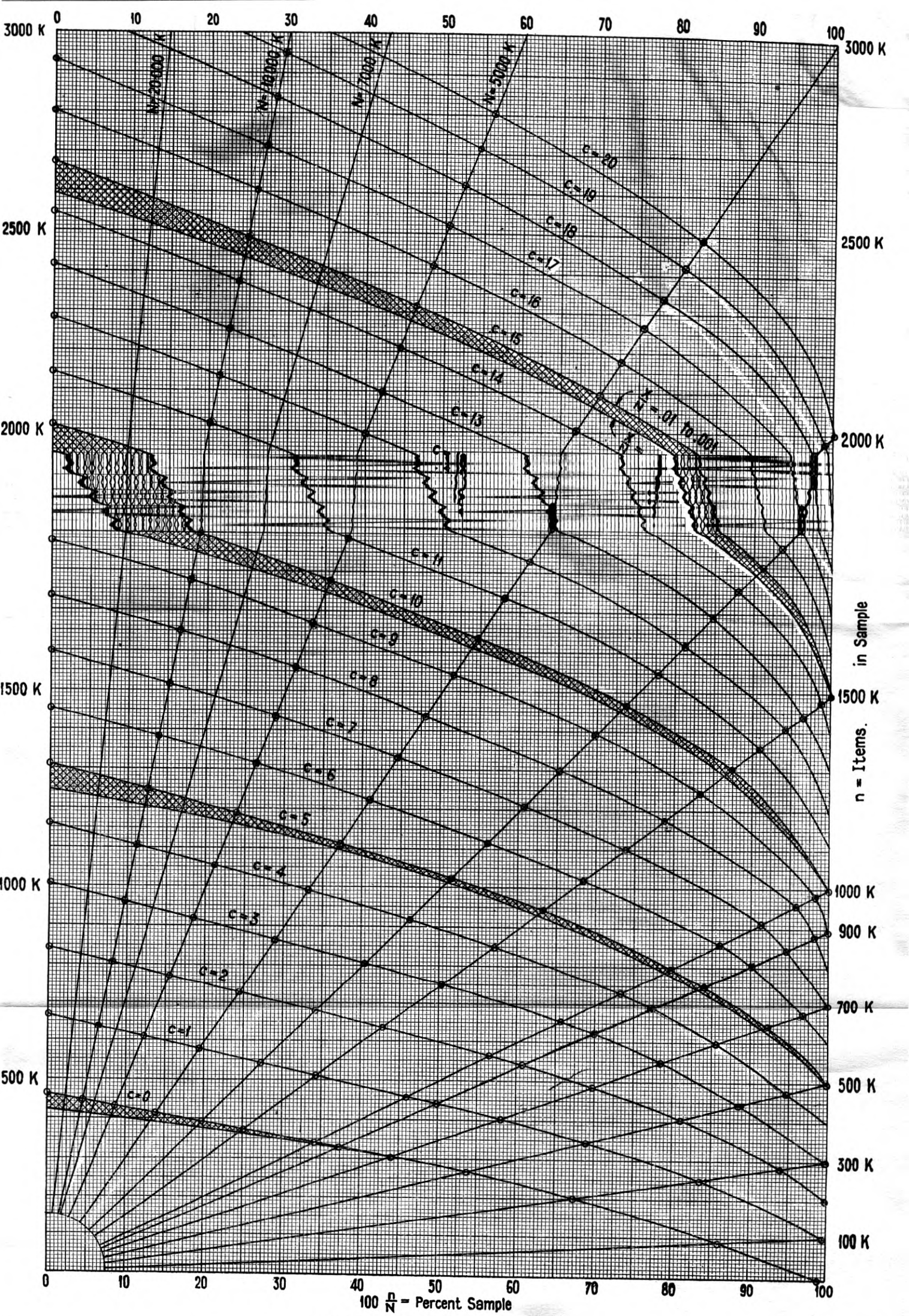
SAMPLING CHARTS C

$$\frac{X}{N} = \frac{.01}{K} \qquad W = .75$$



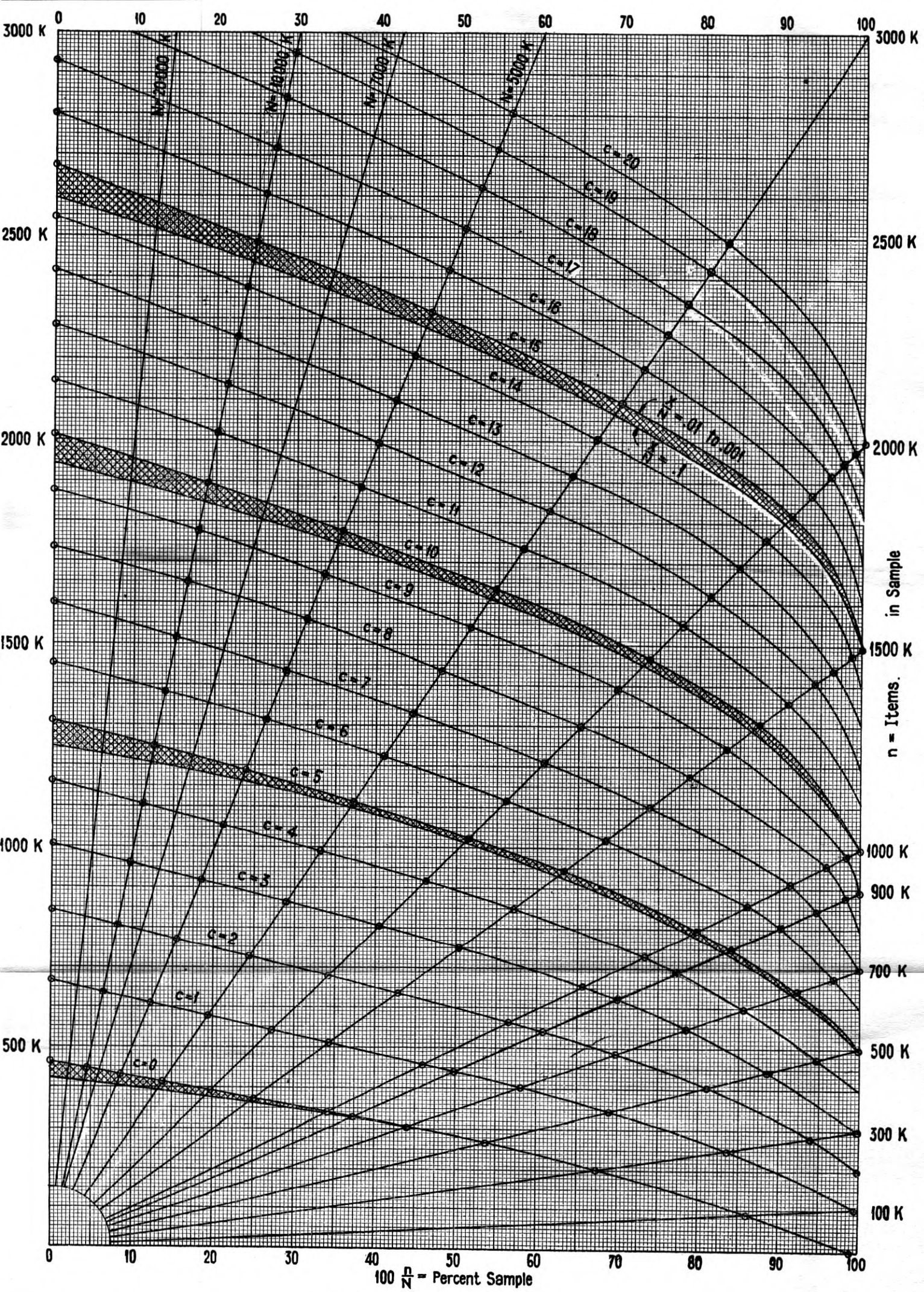
SAMPLING CHARTS C

$$\frac{X}{N} = \frac{.01}{K} \qquad W = .9$$



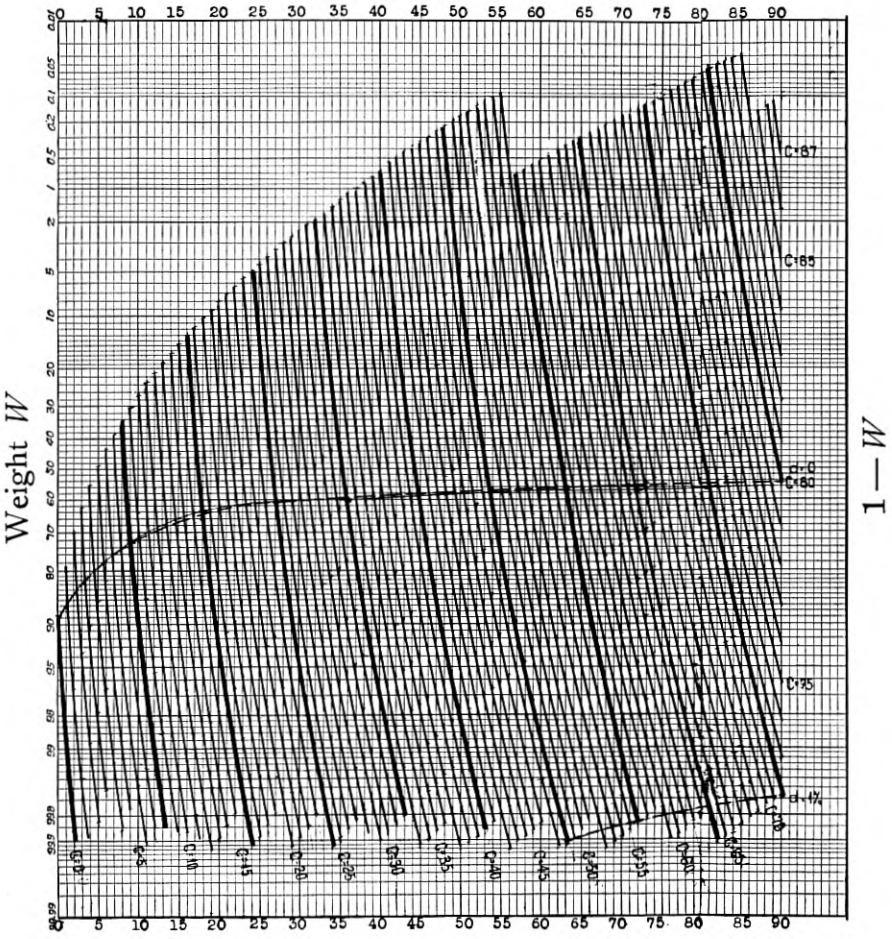
SAMPLING CHARTS C

$$\frac{X}{N} = \frac{.01}{K} \qquad W = .99$$



SAMPLING CHARTS C

$$\frac{X}{N} = \frac{.01}{K} \qquad W = .99$$



CHARTS A

Electrical Measurement of Communication Apparatus[1]

By W. J. SHACKELTON and J. G. FERGUSON

SYNOPSIS: This paper describes precision high-frequency measurements of a fundamental type, special emphasis being placed on the measuring circuits rather than on the types of apparatus measured. Standards of frequency, resistance, capacitance, and inductance are discussed briefly. Bridge measurements are described for the measurement of frequency, inductance, effective resistance, capacitance, dielectric loss, capacitance balance and inductance balance. Circuits for the measurement of other high-frequency characteristics such as attenuation, gain, and cross-talk are included.

INTRODUCTION

LONG DISTANCE electrical communication is now being effected by means of frequencies embracing the audible range and extending from there to the so-called short wave-lengths employed in radio transmission. According to the field of usefulness, this whole range has been subdivided into the audio, the carrier, and the radio ranges. From the viewpoint of the power engineer, all of the frequencies embraced in these ranges are high frequencies, but to the communication engineer, only those frequencies in the upper regions are considered high.

This paper discusses methods of measurement and measuring instruments adapted to the measurement of communication apparatus over this complete range. Most of the measuring apparatus described is designed particularly for use at audio and carrier frequencies. The measuring methods which are discussed are intended primarily for laboratory use in connection with the development and inspection of telephone apparatus prior to its application in the field.

Many of the transmission problems in the communication field involve the impedance characteristics of apparatus and circuits. In the manufacture of apparatus, impedance limits are used to a very great extent in inspection tests. Consequently, quantities of prime importance are those defining impedance characteristics; that is, inductance, capacitance and resistance at specified conditions, of course, such as temperature, frequency, and current or voltage. Other characteristics, of a less fundamental nature but nevertheless of considerable importance, are attenuation, gain, inductance and capaci-

[1] Presented at the Regional Meeting of District No. 1 of the A. I. E. E., Pittsfield, Mass., May 25–28, 1927.

tance balance, cross-talk, flutter and modulation. Since the three impedance components mentioned above, together with frequency, are probably of more general interest, this paper will be devoted largely to a discussion of their measurement, only brief reference being made to the methods used for the measurement of the latter group of characteristics.

As in all measurement work, standards representing the quantity are required, and these are of two classes, prime standards and secondary or working standards. In our case, the prime standards are resistance and frequency. From these we derive inductance and capacitance. Working standards are stable types of inductance coils, air and mica condensers, adjustable resistances, and for frequency, resonance type meters and highly stable oscillators.

Prime Standards

Frequency. The standard of frequency used is that described by Horton, Ricker and Marrison.[2]

Briefly, it comprises a special self-driven fork held at constant temperature and having all other conditions of operation so thoroughly controlled that a high degree of frequency stability is obtained. The exact frequency is measured by driving synchronously a phonic wheel for determining the number of cycles occurring in a given time interval. This time interval is usually a period of 24 hr. as measured by time signals received from Arlington. The average frequency of this fork is capable of being held constant and measured in this way with an accuracy of about 0.001 per cent.

The frequency of 100 cycles obtained from this fork is used to drive a 1000-cycle slave fork from which an equally constant 1000-cycle frequency is obtained. Having these frequencies, all other frequency measurements may be made with as high an accuracy as desired by direct comparison, using the cathode-ray tube as described in detail by Rasmussen.[3]

Resistance. Resistance standards specially designed for use with direct currents and having a very high degree of stability may be readily purchased or constructed and calibrations to a high degree of accuracy may be obtained from the Bureau of Standards. These resistance standards are not suitable for precision measurements at high frequencies, usually being wound on metal spools, and the value of the phase angle receiving only secondary consideration. It is

[2] Horton, Ricker and Marrison, "Frequency Measurement in Electrical Communication," *A. I. E. E. Transactions*, 1923.

[3] F. J. Rasmussen, "Frequency Measurements with the Cathode Ray Oscillograph," *A. I. E. E. Journal*, January, 1927.

necessary, therefore, to use resistance standards of special construction, depending upon the particular application to be made. In all cases, constancy of resistance with variations in atmospheric conditions, frequency and time is imperative. Generally as small a phase angle as possible is also highly desirable, although for some uses a suitable degree of constancy may be sufficient provided that the angle is known, and not large enough to affect appreciably the magnitude of the impedance of the resistance over the frequency range used.

To obtain the highest degree of stability of both resistance and phase angle, it has been found desirable to wind the wire on a spool made of a material not affected appreciably by atmospheric conditions, for example, phenol fiber, and to immerse the complete resistance in a sufficient amount of a suitable sealing compound to exclude all moisture. Resistances meeting all of the requirements outlined have been constructed as described in a recent paper by one of the authors.[4] Coils such as described there, having a resistance of approximately 1000 ohms, may be constructed to have an effective inductance of less than five microhenrys, and this inductance is practically independent of frequency up to at least 100 kc. Coils having lower values down to about 10 ohms can be made with equally small phase angles. Below this value of resistance, it is more difficult to hold a low phase angle.

Coils constructed as described may be considered to have so small a change in resistance with frequency that a calibration with direct current may be used without appreciable error for all frequencies at which they are used. Both the variation in resistance with frequency and the phase angle may be most readily measured by comparison with some simple type of resistance of such geometrical form that the phase angle may be readily computed. Satisfactory resistances for this purpose are short lengths of fine wire of definite shape, sputtered metal films on glass or other insulating material, and carbon in the form of rod or film.

Secondary Standards

Capacitance. The value of our capacitance standards is determined in terms of the prime standards of frequency and resistance. This determination may be made in several ways, the following bridge method being a simple and accurate one. The circuit, as shown in Fig. 1, consists of two equal resistance ratio arms, a resistance and capacitance in parallel in the third arm and a resistance and capacitance in series in the fourth arm. When this bridge is balanced at any particular frequency, the relations between the impedance arms

[4] W. J. Shackelton, "A Shielded A-C. Inductance Bridge," *A. I. E. E. Journal*, February, 1927.

of the bridge are such that the value of each capacitance may be determined in terms of the frequency and the two resistances.

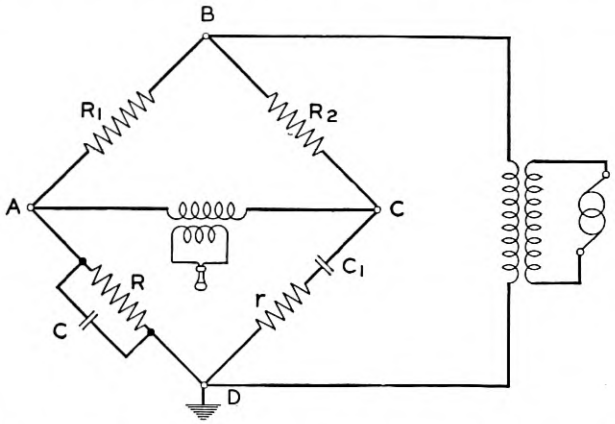


Fig. 1—Bridge circuit for measuring capacitance in terms of resistance and frequency

The requirements for a capacitance standard are high constancy with variations in frequency, time, voltage, and atmospheric conditions, and a small phase difference. Mica has been found to be the best solid dielectric, used either alone or impregnated with a high quality wax such as paraffin. If mica alone is used, the condenser must be sealed to prevent the entrance of moisture.

Good mica condensers can be obtained with a temperature coefficient below 0.005 per cent per deg. cent., and having a variation of less than 0.1 per cent over a frequency range from 500 cycles to 100 kc. Variations in capacitance with voltage are also negligible provided voltages below 100 volts are used. It has been our experience that the paraffin-impregnated condensers generally have a negative change of capacitance with temperature. This change is smaller than that of the unimpregnated type which has a positive change with temperature. The paraffin-impregnated condensers, however, usually change more with time than the unimpregnated condensers.

Air condensers may be used as standards in small sizes. For the larger values, the air condensers become large and cumbersome and are not as stable as the mica condensers. Even in the smaller sizes, very special precautions must be taken to obtain air condensers which have appreciably smaller phase differences than the mica condensers, which may be made with phase differences considerably less than one minute.

Inductance. Requirements for inductance standards are high con-

stancy with variations in time, current or saturation, atmospheric conditions, and frequency. It is also desirable that they be made with a small external field. Otherwise, very great care must be taken to avoid errors due to this cause.

In order to obtain stability with variations in saturation, it is usual to make inductance standards with air cores. This requires standards of large physical size if a time constant as large as the average iron core coil is desirable. This large size results in large capacitance distributed in the coil itself and from the coil to ground. These capacitances cause large variations in inductance with frequency and with the position of the coil with respect to ground. On account of this difficulty with air core coils, permalloy [5] as core material has been used with considerable success as described by one of the authors. [4]

The calibration of these inductance standards may be made by comparison with any two of the quantities, capacitance, resistance and frequency. Comparison with frequency and resistance may be made in a bridge circuit exactly similar to the one used for capacitance determination, substituting inductances for capacitances. A comparison with frequency and capacitance may be made by means of a resonant method, and comparison with capacitance and resistance may be made by means of the Owen bridge.[6] The resonant method is used generally except for those cases requiring large capacitance, in which cases the Owen bridge is used.

Frequency. As a secondary standard of frequency for use with th cathode ray tube, where practically only one standard frequency is required, a special 1000-cycle oscillator is used, designed particularly for high stability of frequency with ordinary variations in external conditions. This oscillator is shown in Fig. 2. It allows the use of a cathode ray tube for frequency measurements with a high degree of accuracy under conditions where the prime standard of frequency is not accessible.

Where a portable frequency standard is desirable, for instance, as a means of shop frequency checks, a resonance type of meter is used. This is shown in Fig. 3. It is essentially a resonance bridge circuit consisting of two equal resistance ratio arms, a third arm containing a resonant circuit, and a variable resistance as the fourth arm. The capacitance and resistance are variable over wide ranges by means of decade switches, and the capacitance is capable of fine variations by the use of a form of precision variable air condenser having provision for fine control. There are four air-core inductance coils which give,

[5] H. D. Arnold and G. W. Elmen, *Franklin Institute Journal*, Vol. 195, 1923.

[6] D. Owen, "A Bridge for the Measurement of Self-Inductance," *Proceedings of the Physical Society of London*, October, 1914.

in conjunction with the variable capacitance, a frequency range of about 100 cycles to 150 kc.

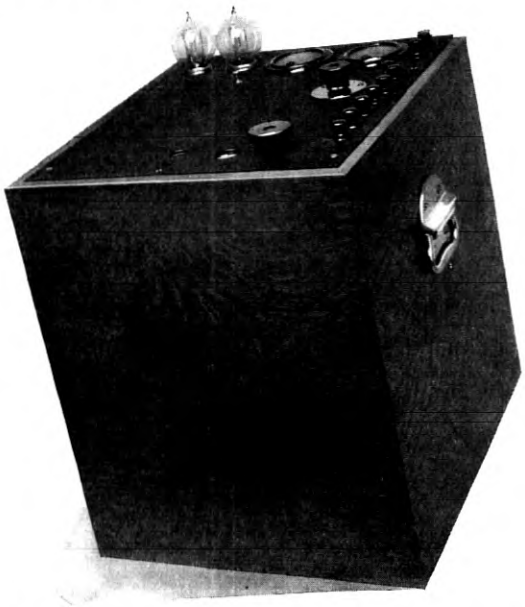


Fig. 2—Single-frequency vacuum tube oscillator used as secondary standard of frequency

The meter is calibrated by balancing the circuit by means of the variable resistance and capacitance with a known frequency input, and recording the coil and condenser settings. It is used for checking frequencies by reversing the process, that is, connecting the source of unknown frequency to the bridge, balancing as before, and determining the frequency by reference to the calibration. There are no input or output transformers connected to this circuit and on this account certain precautions must be taken in connecting the output and input circuits to it; but it is a relatively low impedance circuit, and troubles due to this cause have not been found serious.

Resistance. A convenient secondary standard of resistance is a dial box having the resistance units designed to meet the same requirements as the prime standards. Commercial dial boxes are available, having satisfactory stability with variations in frequency and atmospheric conditions, and having sufficiently small phase angles for all frequencies but the highest radio frequencies.

A dial box, requiring, as it does, a certain amount of wiring between dials, and having all of the dials connected permanently whether they

are used or not, always has more capacitance and inductance associated with it than a single resistance of the same value. A certain amount of compensation between the capacitance and inductance may be effected by proper design, but it may be generally accepted that the inductance of the wiring makes the phase angle of the low

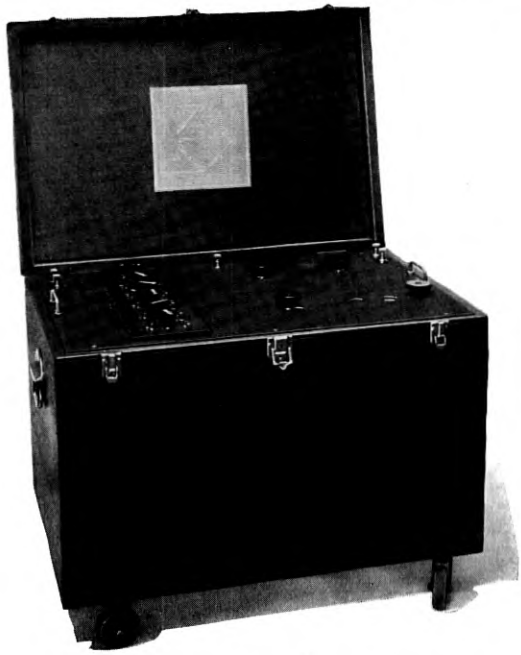


Fig. 3—Resonance-type frequency meter

resistance values comparatively high and the capacitance between dials and between units of each dial makes the phase angle of the high values comparatively high. This effect can only be overcome by a compact design using coils of small physical size. This sets a limitation on coils for use in dial boxes which is not present to such an extent in the case of single resistance units or single value prime standards.

Methods of Measurement

We have discussed already measurements of frequency and resistance in connection with the description of standards, and we will not discuss them further here. We are particularly concerned with the measurement of impedance of all types, it being understood that any resistance having a phase angle which is not negligible or which is of special interest is to be considered a special type of impedance.

In measuring impedances, we have found that those methods which determine the unknown in terms of circuit constants are superior to those requiring the measurement of current and voltage. Accordingly, bridge methods are used almost exclusively and, furthermore, the bridge type which is used wherever possible is the equal ratio arm bridge in which a direct comparison is made of the unknown impedance with a known impedance adjusted to that same value. This type of measurement has the disadvantage of requiring standards of the same value as the quantity measured over the whole range of impedances used, but it has the compensating advantages that, having standards whose value is known, this circuit is extremely simple, very easy to check at any time, and may be made extremely accurate.

Auxiliary Apparatus. Without going into details regarding the auxiliary apparatus used in connection with bridge measurements, we may state briefly that vacuum tube oscillators are used almost exclusively for furnishing all frequencies, and that the telephone receiver is used almost exclusively as a detector, due to its simplicity and the rapidity with which it may be used. For frequencies below 200 cycles, it is used with a chopper to give a tone of about 1000 cycles, and above 3000 cycles, it is used with a heterodyne detector to give a beat note of about 1000 cycles. In the audio frequency range, it is used alone or with an amplifier, if necessary.

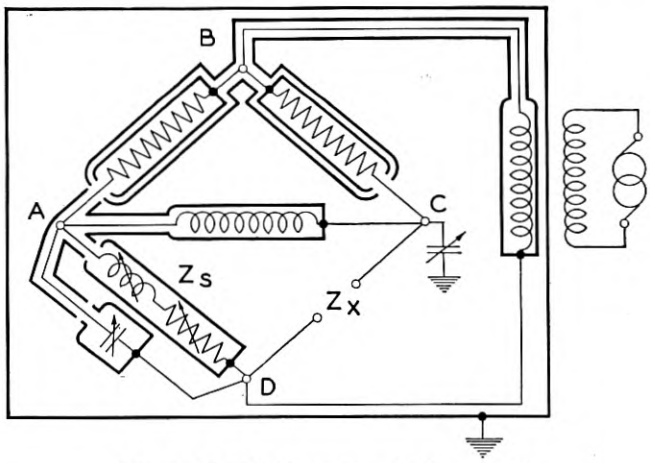


Fig. 4—Shielded impedance bridge circuit

While it is impossible to draw a distinct line between the methods of measurement of different types of impedances, certain bridge circuits have been designed primarily for certain types of measure-

ments, and we will therefore classify them in this way, although in general they have a considerably wider sphere of usefulness than indicated.

Inductance. A simple shielded bridge for the measurement of inductance and resistance has been described by one of the authors [4] and is shown in schematic form in Fig. 4. It comprises two equal resistance ratio arms, an adjustable standard of self-inductance, an adjustable resistance standard, a thermocouple milliammeter, two reversing switches, two transformers, and two air condensers. This apparatus is grouped into three separate units, as shown in Fig. 5, one comprising the

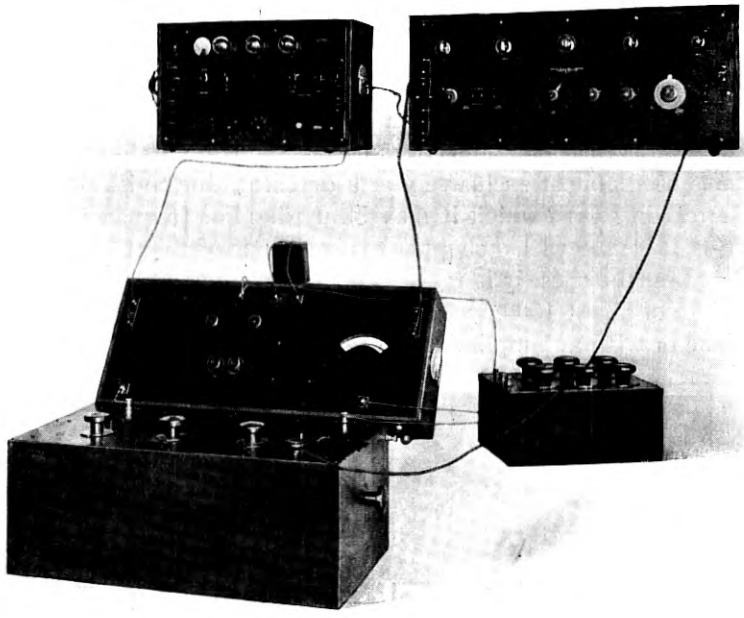


Fig. 5—Shielded impedance bridge and standards, connected to vacuum tube oscillator and heterodyne detector

standards of inductance, one the resistance standard, and one the remaining parts of the circuit. Each of these units is shielded electrostatically. The last assembly constitutes the balance element of the system, by means of which the unknown and standard impedances are compared. This unit may be used alone for the comparison of two impedances of any type since the only condition for balance is the exact equality of impedances in the two arms. Using in addition the standard inductance and resistance shown, it is adapted particularly for measuring inductance and effective resistance. The inductance

standard may be made with a range of 10 henrys to a minimum of two millihenrys, using an inductometer having a minimum scale division of 0.1 millihenry, or the range may be any simple multiple of this. Values as low as one microhenry at frequencies as high as 150 kc. are measured in this way.

By connecting the resistance in one arm of the bridge and a capacitance in series with an inductance in the other arm, we may use it to indicate resonance, and if we measure the frequency we may use this method for the comparison of capacitance with inductance. This is the method actually used for the calibration of the inductance standard used with the bridge. The bridge may be used for the comparison of capacitance. The bridge described later for the measurement of capacitance, however, has certain special features which make it peculiarly adapted to the measurement of capacitance and conductance.

Inductance with Superposed Direct Current. In telephone work, it is often of value to know the performance of apparatus, particularly of iron core impedances, when used at telephone frequencies while at the same time carrying direct current. The bridge shown in schematic form in Fig. 6 will measure the inductance of the coil at audio frequency

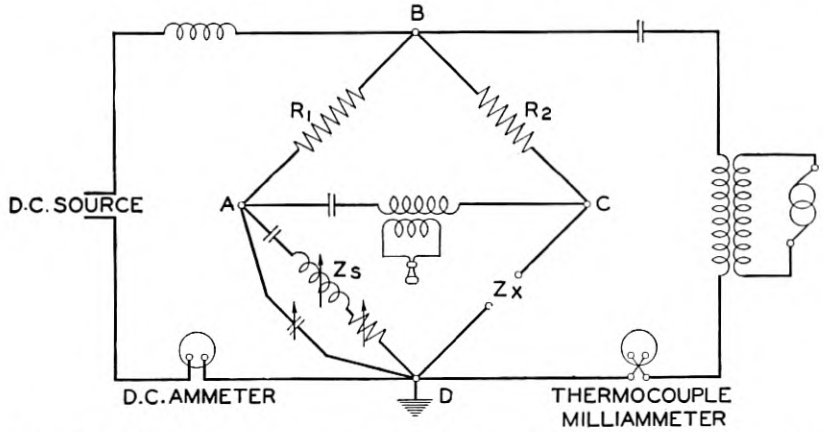


Fig. 6—Bridge circuit for measuring impedances with superposed direct current

with a direct current flowing through it. As shown in the figure, the direct current is kept out of all of the arms of the bridge except one ratio arm and the test arm, by means of condensers, and the alternating measuring current is separated from the direct current by means of a choke coil. None of these added features affect the bridge balance except the capacitance in the standard arm, and this is made large

enough (26 μ f.) to have an impedance small compared with the impedance measured. In any case, a correction may be made by taking first a zero reading which will be slightly positive due to the inductance necessary to compensate for the capacitance in this circuit. This correction will vary with frequency but at 1800 cycles, for instance, with 26-μ f. capacitance, the correction is only about 0.3 millihenry and the inductances measured are usually considerably larger than this.

The circuit is extremely simple and convenient to use. The values of alternating current and direct current can each be measured separately outside of the bridge circuit and the inductance standards do not need to be constructed to carry the direct current. The only part of the bridge required to carry the direct current is one ratio arm and, in consequence, it is a comparatively simple matter to construct such a bridge to carry several amperes of direct current. Where very high direct currents are required, the ratio arms may be reactances wound on a single core, instead of resistances, thus reducing the loss due to the passage of the direct current.

Flutter. In telephone circuits used for joint telephone and telegraph service, it is desirable to know the effect of the telegraph impulse on the telephone frequency inductance and effective resistance of the loading coils used on the lines. This effect, known as "flutter," with a method of measuring it, is described in detail by Fondiller and Martin.[7] The measuring circuit consists of a double bridge, the inner one consisting of two similar loading coils on which the flutter effect is to be measured and two other coils of comparatively high impedance approximately equal in value and which have negligible flutter effects, the four coils being connected to form a balanced bridge. The low frequency corresponding to the telegraph impulse is introduced at two diagonal corners and the other two corners, which are at a common potential with respect to the low frequency, are connected to the usual test terminals of an impedance bridge of the type already described. With no low-frequency current passing through the coils, a continuous balance may be obtained on the main or high-frequency bridge using an audio frequency input. From this, the normal effective resistance and inductance of the coils may be obtained.

When the low-frequency current passes through the coils, the inductance and effective resistance are different for every point of the low-frequency cycle. Thus, only an instantaneous balance of the outer bridge is possible. This instantaneous balance for any particular point in the low-frequency cycle may be made by the use of an electro-

[7] W. Fondiller and W. H. Martin, *Transactions of the A. I. E. E.*, 1921, Vol. 40, p. 553.

magnetic oscillograph. By this means as described in the paper already mentioned, it is possible to obtain the curve of variation of inductance and effective resistance of the coil over one low-frequency cycle.

Another method used at the present time employs the same bridge circuit but an entirely different method of detecting the cyclic variation in the balance. This method of detection uses the cathode-ray oscillograph and is as follows. The low-frequency source is connected across a high resistance and condenser in series, the two having equal impedances. The potentials across the condenser and resistance are then placed respectively across the horizontal and vertical plates of the oscillograph. These two potentials, being equal in magnitude but 90 deg. apart in phase, give a circle on the screen. The output of the main bridge is now connected through a transformer whose secondary is connected in series with the oscillograph cathode potential. Due to the fact that the sensitivity of the tube to deflections by the plate potentials varies with the cathode potential, the radius of this circle produced by the low frequency is a function of the telephone frequency input from the bridge, and instead of a circle we get a band, the width of which is a measure of the degree of unbalance of the bridge. The point in the cycle at which the bridge is balanced, is indicated on the screen as the point where this band diminishes to a line, and the angular position of this point in the band determines the phase position of this balance with respect to the low-frequency cycle. It is possible in this way to balance the bridge for any angular position corresponding to any point in the low-frequency cycle, and by taking sufficient points, to obtain a curve of variation of the coil constants over a complete cycle. This method is found to be simpler and faster than the method using the mechanical oscillograph.

Inductance Balance. A simple form of bridge for measuring inductance balance of the two windings of a transformer or other coil uses the two windings of the transformer for two arms, the other two arms being resistances, one of which at least is variable. The balance is made by means of the variable resistance, the ratio of the two resistances at balance then giving the unbalance of the transformer. If one of these resistances is made 100 ohms, the variation of the other from 100 ohms at balance gives directly the percentage unbalance. Any unbalance in resistance is usually comparatively small and may be taken care of by low resistances in series with the transformer windings.

Ratio of Transformation. A similar bridge may be used for the measurement of ratio of transformation. There are many cases where

the secondary of a step-up transformer has an inductance which is inconveniently large to measure directly, and the ratio of transformation circuit eliminates this necessity. The circuit used is practically the same as that already described for measuring inductance balance, the ratio of transformation being equal to the ratio of the resistance arms of the bridge at balance.

Capacitance. The direct comparison of capacitance is made in a special bridge known as the Campbell [8]-Colpitts [9] capacitance and conductance bridge. The ratio arms, input and output circuits, and the shielding are similar to the impedance bridge already described. The unique feature of this bridge is the method of connecting the standard air condenser to eliminate the dielectric loss in the measurement of capacitance. The schematic diagram of the bridge is shown in Fig. 7. Instead of connecting the standard condensers in the

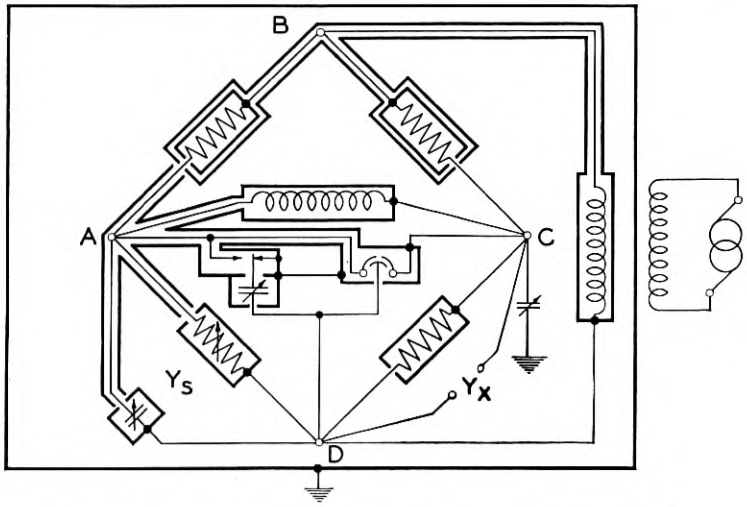


Fig. 7—Schematic circuit of capacitance and conductance bridge

arm AD as in the case of the impedance bridge already described, a special switch is used to switch these condensers from AD to CD, and in the case of the continuously variable condenser, the three-plate construction is used, causing a decrease in the capacitance in CD as the capacitance in AD is increased.

The method of construction of the unit air condensers is shown in

[8] G. A. Campbell, "The Shielded Balance," *Electrical World and Engineer*, April 2, 1904, p. 647.

[9] G. A. Campbell, "Measurement of Direct Capacities," *Bell System Technical Journal*, July, 1922, p. 18.

Fig. 8. It may be seen from this figure that all capacitances which include dielectric material are permanently connected across CD or AC and so are not changed when the condenser is switched, or else

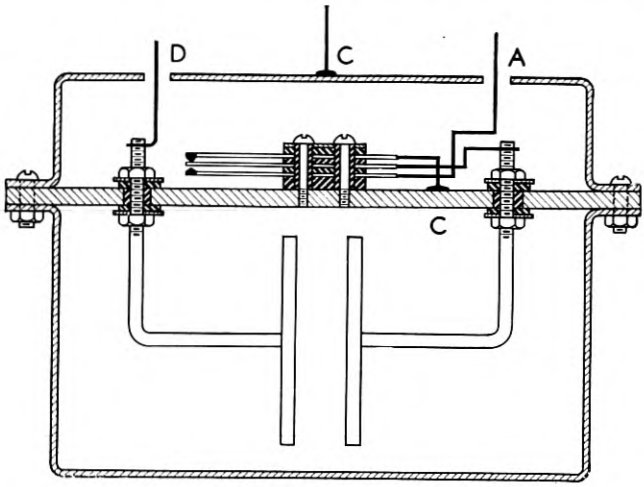


Fig. 8—Air-condenser construction employed in the capacitance and conductance bridge

they are switched so that capacitances across AC, which do not enter into the bridge balance, are short-circuited on switching. This scheme eliminates all dielectric loss in the standards when measuring condensers by comparison with them. It has the additional advantage that the capacitances in the bridge have twice the effect they would have if simply switched in and out of the circuit.

By the use of this bridge, it is possible to measure capacitances up to the maximum limit of the range of the air condensers with a negligible loss in the standard condensers. This capacitance range is usually up to 0.01 μ f. and for condensers above this value the conductance is measured by comparison with that of the maximum value of the air condenser, assuming it to have negligible conductance. Of course this method of eliminating dielectric loss is not applicable to the use of mica condenser standards and if a range greater than 0.01 μ f. is desired, the mica condensers are simply connected in the usual way across AD.

Another feature of this bridge is the method of measuring conductance. The connection of a variable resistance, either in series or in shunt, with the standard condenser for the measurement of loss in the test condenser has objections due to the wide range of resistance

values required to cover the possible variations in losses. A compromise is effected in this bridge by connecting a 10,000-ohm shunt across each of the arms *CD* and *AD*. A slight difference in the losses in these two arms can then be measured by varying one of these resistances slightly. Since the standard condenser practically always will have lower losses than the condenser tested, it is usual to place a fixed 10,000-ohm resistance across *CD* and a resistance across *AD* variable in 0.01-ohm steps to 10,000 ohms. A change of one ohm in this resistance, when balancing a condenser, is equivalent to shunting it with a resistance of 100 megohms or 0.01 micromho. Accordingly, the conductance of a condenser may be measured in micromhos by simply dividing the resistance change in ohms by 100. This, of course, is only approximate in the case of large conductances, but is correct to 1 per cent for values up to one micromho.

Due to the condensers forming such an integral part of the bridge circuit, they are all built into the bridge. The complete bridge is shown in Figs. 9 and 10. Fig. 9 is a top view showing the capacitance and

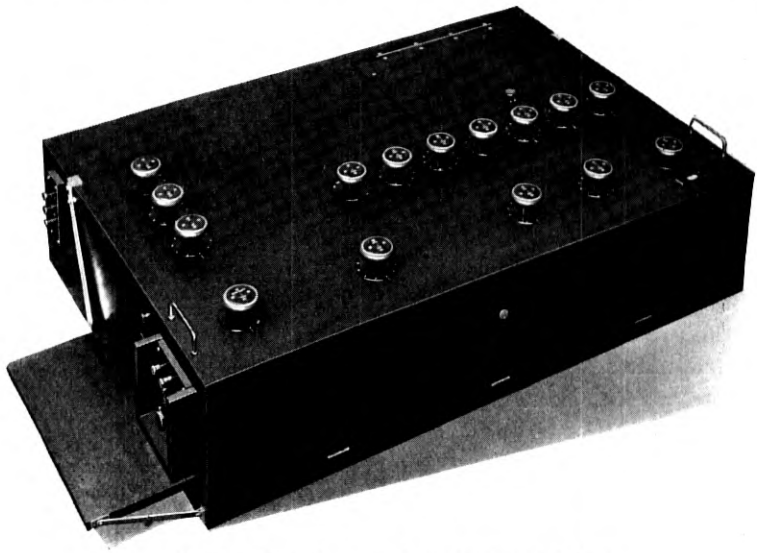


Fig. 9—Capacitance and conductance bridge

resistance dials for effecting a balance, and Fig. 10 is a view with the cover removed, showing the method of shielding the individual parts. The range of capacitance is from 0.1 $\mu\mu$ f. up to three μ f., and the frequency range is from about 10 cycles up to about 150 kc., the only modifications required in the bridges to cover this whole frequency

range being a change in input and output transformers, as it is not found practicable to design these transformers to give efficient operation over such a wide frequency range.

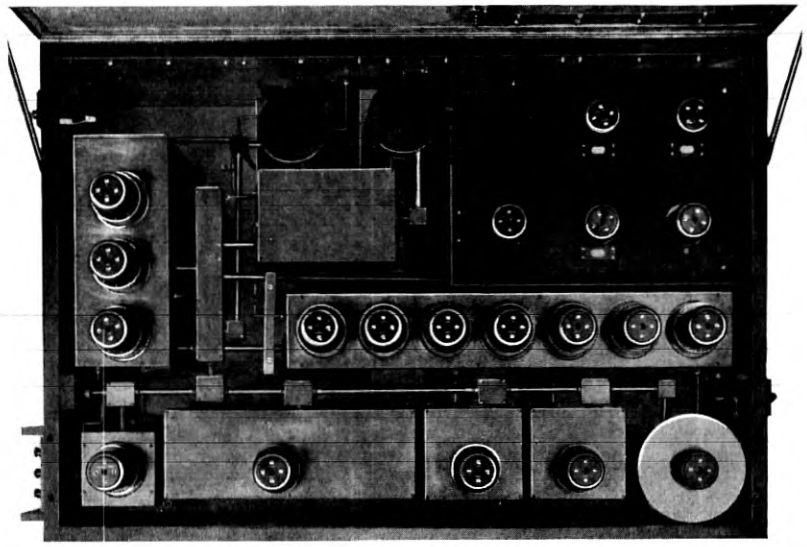


Fig. 10—Capacitance and conductance bridge with cover removed, showing method of assembly and shielding

A comparison of this bridge with the impedance bridge already mentioned shows it to be essentially the same circuit, the capacitance bridge having conductance shunts not included in the impedance bridge which allow a conductance balance to be made more readily. It is obvious that any two impedances can be compared on this bridge. Inductances may be measured by parallel resonance by simply placing them in the *AD* arm in parallel with the standard condenser and effecting a balance with it. This method is used to some extent for the measurement of large inductances.

Capacitance Unbalance. In order to keep cross-talk low in long cable circuits, it is necessary to have a high degree of capacitance balance between the various conductors in the cable, more particularly between the four conductors of a phantom group. The unbalances of interest are the phantom to each side circuit and the side-to-side unbalances. These may be measured on a capacitance bridge by measuring all of the direct capacitances [9] associated with the group and computing the unbalances required. A special circuit, however, is generally used which measures directly the particular un-

balances in which we are interested. It consists of an input and an output transformer, two equal resistance ratio arms, a variable air condenser of the three-plate type, four binding posts for connecting the four conductors of the quad, and switches for making the various connections. By means of the switches, the cable conductors are connected to the circuit in such a way that the reading of the air condenser when a balance is obtained indicates directly the unbalance, either side-to-side or phantom-to-side, according to the switch positions. This circuit when used as a laboratory instrument is capable of measuring capacitance unbalance as low as 1 $\mu\mu$ f.

Attenuation and Gain. So far, we have discussed the measurement of the fundamental impedance characteristics of apparatus. When the component parts have been found to meet their individual impedance requirements and are assembled to form the completed apparatus, it is desirable to have tests made of the over-all performance of this apparatus. In a large number of cases, the requirement of greatest importance is the attenuation frequency characteristic. It is fairly obvious that this characteristic, of all apparatus used in telephone lines, is of interest, and this is particularly true of all types of filter circuits which are designed primarily for the purpose of furnishing definite attenuation frequency characteristics. These measurements are particularly required on apparatus used in carrier-current telephony and telegraphy.

From the very nature of the measurements, it is difficult to obtain a null method of measuring attenuation. The most direct method is to measure the input and the output of the apparatus under test simultaneously, from which the attenuation may be computed. The practical difficulty in doing this is to measure the extremely small outputs which are obtained from apparatus having high attenuations, where the characteristic must be obtained with the normal input, which is usually low. In general, it has been found necessary to use some form of amplifying device in the output circuit and it has not been found desirable to rely on the constancy of amplification of this device. Accordingly, the usual method used for the measurement of attenuation is a substitution one. The circuit is shown in Fig. 11A. There are two branches in this circuit, one of which includes the apparatus under test and the other, a variable standard attenuator. The output of each branch is arranged to connect either to a detector of impedance Z_1 equal to the impedance of the standard attenuator or to a fixed impedance of the same value. If the apparatus under test has the same impedance as the standard attenuator, the input impedances Z_1 and Z_3 are made equal and the matching impedance Z_2

is omitted. Then the two branches of the circuit will be identical, provided the attenuation of the standard attenuator is equal to that of the apparatus under test. Accordingly, the method of measurement is to switch the detector first to one and then to the other branch,

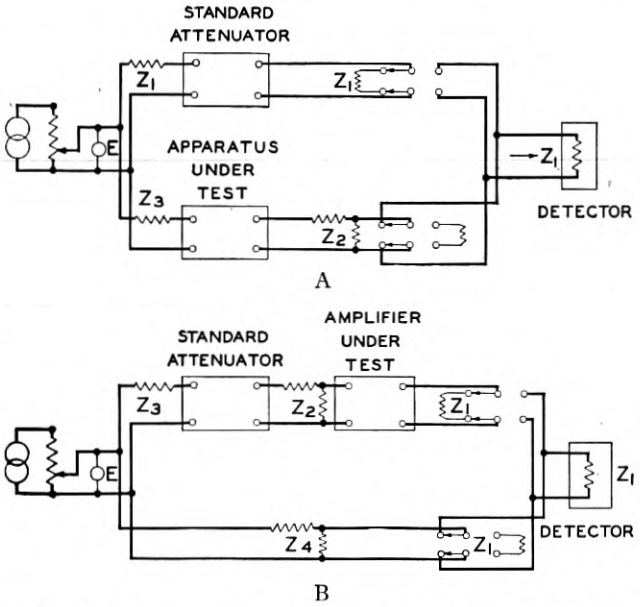


Fig. 11—Circuits for measuring attenuation and gain. A. Arrangement for measuring loss. B. Arrangement for measuring gain

adjusting the standard attenuator until an equal output is obtained for either switch position. The attenuator then reads directly the loss in the apparatus. The total input of the circuit is independent of the switch position, since the impedance conditions remain unchanged in switching.

If the apparatus under test has not the same impedance as the standard attenuator, the input impedance Z_3 and the matching network Z_2 are adjusted so that the circuit still reads directly.

The standard attenuator is a resistance network capable of variation in small steps, each step consisting of a network of the L, T or H type, the resistance values being such as to give the desired attenuation between the output and input terminals. It is usually calibrated in 0.1-T.U. steps and may read as high as 100 T.U. corresponding to a ratio of power output to power input of ten billion to one or, if the impedances are the same, which is usually the case, corresponding to a current or voltage ratio of 100,000 to 1.

The calibration of these attenuators is based on the measurement of the individual resistances. Of course, sufficient measurements are made to determine that any capacitances which enter do not affect appreciably the accuracy of the attenuator at the maximum frequency used, which may be as high as 150 kc.

By modifying the circuit of Fig. 11A, we may use it to measure gain as shown in Fig. 11B. In this arrangement, the lower branch contains an impedance Z_4 that is adjusted to introduce a loss equal to that of the matching impedance Z_2 in the upper branch. In other words, with the amplifier under test out of the circuit and the standard attenuator set at zero, the detector will read the same for either position of the output switch. Then when the amplifier is introduced into the circuit, the attenuator is adjusted until the detector reads the same for either switch position, which means that the gain of the amplifier is just neutralized by the attenuator and the setting of the latter is read as gain.

This circuit is used principally for the measurement of gain of audio frequency amplifiers, and is capable of measuring gain as high as 120 T.U. corresponding to a power output of 1,000,000,000,000 times the power input.

Cross-Talk. When there is an appreciable amount of coupling between two telephone circuits, any mutual interference which results is known as cross-talk. It is measured in cross-talk units, a cross-talk unit being defined as the relation existing between the two circuits when the current in the disturbed circuit is one millionth of the current in the disturbing circuit, the impedances of the two circuits being the same. Under these conditions, one cross-talk unit may be assumed the same as 120 T.U. An interesting form of cross-talk is that due to loading coils and is of a complex type, produced by a combination of capacitance, inductance and resistance unbalances in the windings. Since the actual cross-talk caused by an unbalance in the coil is dependent upon all of the conditions of the circuit, it is necessary that any measurement of cross-talk made on the individual coils be made in a circuit as nearly as possible the equivalent of the line in which the coil is to be used. Consequently, all cross-talk circuits for the measurement of loading coil cross-talk consist of networks simulating the impedance of an ideal line of the type for which the loading coil is designed. The principle of the method is to apply to the disturbing circuit a definite input of a single frequency, usually 900 cycles, and to measure the cross-talk in the disturbed circuit at the desired point in it by comparing the tone heard in the telephone receiver connected at this point with the tone obtained from a cross-talk meter which is simply a device

for obtaining a definite part of the input, and having a scale reading in millionths, that is, in cross-talk units. The measurement is made by switching from the cross-talk meter to the disturbed line and adjusting the cross-talk meter until the tone heard in each case is the same. The method is therefore not a null method and depends to some extent on the judgment of the operator, but results accurate to one or two cross-talk units may be obtained by this method. The coils as commercially produced after adjustment for this requirement are usually within 10 cross-talk units, representing an unbalance in the circuit due to the coil unbalance of less than one part in 100,000.

Conclusion

We have described in this paper a number of the more important high-frequency methods of measurement and measuring circuits. It has been impossible to cover all of the different methods and circuits used, but we believe that the information given will be of value to those interested in this field of work.

We have not been able, in a paper of this type, to go into details concerning any specific circuits used, but we have referred to papers which describe in greater detail some of these methods and circuits, and it is expected that other papers will be published in the future covering other circuits which have received only brief mention here.

The Diffraction of Electrons by a Crystal of Nickel

By C. J. DAVISSON

This article is taken from the manuscript prepared by the author for his address at the joint meeting of Section B of the American Association for the Advancement of Science and the American Physical Society on December 28, 1927, at Nashville, Tennessee. An account of this work giving fuller experimental details is given by Davisson and Germer in the December, 1927, issue of the *Physical Review*.

These experiments are fundamental to some of the newer theories in physics. Until they were performed, it could be said that all experimental facts about the electron could be explained by regarding it as a particle of negative electricity. It now appears that in some way a "wave-length" is connected with the electron's behavior. The work thus shows an interesting contrast with the discovery of A. H. Compton that a ray of light (a light pulse) suffers a change of wave-length upon impact with an electron, the change of wave-length corresponding exactly to the momentum gained by the electron. Until Compton's work, all the known facts about light could be explained by thinking of light as a wave motion. The Compton effect seems to prove the existence of particles of light.

Physics is thus faced with a double duality. Compton showed that light is in some sense *both* a wave motion and a stream of particles. Davisson and Germer have now shown that a beam of electrons is in some sense *both* a stream of particles and a wave motion.

At the same time, theoretical advances have been made which seem to pave the way for an understanding of this curious situation. A general account of these new developments was given by K. K. Darrow in his series "Contemporary Advances in Physics" in the *Bell System Technical Journal* for October, 1927. Some remarks on the relation of the Davisson and Germer experiments to the new mechanics were given in this article, p. 692 *et seq.*—EDITOR.

THE experiments which I have been asked to describe are the most recent of an investigation of the scattering of electrons by metals on which we have been engaged in the Bell Telephone Laboratories for the last seven or eight years.

The investigation had its inception in a simple but significant observation. We observed some time in the year 1919 that when a beam of electrons is directed against a metal target, electrons having the same speed as those in the incident beam stream out in all directions from the bombarded area. It seemed to us at the time that these could be no other than particular electrons from the incident beam that had suffered large deflections in simple elastic encounters with single atoms of the target. The mechanism of scattering, as we pictured it, was similar to that of alpha ray scattering. There was a certain probability that an incident electron would be caught in the field of an atom, turned through a large angle, and sent on its way without loss of energy. If this were the nature of electron scattering it would be possible, we thought, to deduce from a statistical study of the deflections some information in regard to the field of the

deflecting atom. It was with these ideas in mind that the investigation was begun. What we were attempting, it will be seen, were atomic explorations similar to those of Sir Ernest Rutherford and his collaborators but explorations in which the probe should be an electron instead of an alpha particle. I shall not stop to recount the earlier experiments of this investigation, but shall pass at once to the most recent ones—those in which Dr. Germer and I have studied the scattering of electrons by a single crystal of nickel.

The unusual interest that attaches to these experiments is due to their revealing the phenomenon of electron scattering in a new and, I may say, fashionable rôle. Electron scattering is not, it would seem, the mildly interesting matter of flying particles and central fields that we supposed, but is instead a much more interesting phenomenon in which electrons exhibit the properties of waves. The experiments reveal that the way in which electrons are scattered by a crystal is very similar to the way in which x-rays are scattered by a crystal. The analogy is not so much with the alpha ray experiments of Sir Ernest Rutherford, as with the x-ray diffraction experiments of Professor von Laue.

My task of describing these experiments is much simplified by the fact that the experiments of Professor von Laue are so well known and so thoroughly comprehended. I remind you very briefly that in the original Laue experiment a beam of x-rays was directed against a crystal of zincblende, that about the transmitted beam was found an array of regularly disposed subsidiary beams proceeding outward from the irradiated portion of the crystal, and that these subsidiary beams could be interpreted completely and precisely in terms of the then already popular wave theory of x-radiation. They could indeed be explained as diffraction beams that resulted from the superposition of secondary wave trains expanding from the regularly arranged atoms of the crystal lattice.

There are two features of the Laue experiment which we shall need particularly to remember. The first is that diffraction beams issue not only from the far side of the crystal along with the transmitted beam, but also from the near or incidence side of the crystal—these latter being disposed in a regular array about the incident beam. The second is that each diffraction beam is characterized by a particular wave-length, and that a given beam appears in the diffraction pattern if the incident beam contains radiation of its characteristic wave-length, or of some submultiple value of this wave-length, but not otherwise. If the incident beam is monochromatic, no diffraction beams appear at all unless the wave-length of the incident beam

happens to coincide with a wave-length of one or more of the diffraction beams. In that case the favored beams appear but no others.

With this picture of x-ray scattering in mind one sees at once the significance of the main results of the present experiments. A homogeneous beam of electrons is directed against a crystal of nickel, and at certain critical speeds of bombardment full speed scattered electrons issue from the incidence side of the crystal in sharply defined beams —a few beams at each of the critical speeds—the totality of such beams making up a regularly disposed array similar to the array of Laue beams that would issue from the same side of the same crystal if the incident beam were a beam of x-rays.

The electron beams are not identical in disposition with the Laue beams, and yet it is possible to treat them as diffraction beams, and from their position and from the geometry and scale of the crystal to calculate "wave-lengths" of the incident beam—just as we might do if we were dealing with x-rays or with any other wave radiation. When this is done we arrive at a definite and simple relation between the speed of the electron beam and its apparent wave-length—the wave-length is inversely proportional to the speed.

Surprising as it is to find a beam of electrons exhibiting thus the properties of a beam of waves, the phenomenon is less surprising today than it would have been a few years ago. We have been prepared, to a certain extent, by recent developments in the theory of mechanics for surprises of just this sort—for the discovery of circumstances in which particles exhibit the properties of waves. We have witnessed, during the last three years, the inception and development of the idea that all mechanical phenomena are in some sense wave phenomena—that the rigorous solution of every problem in mechanics must concern itself with the propagation and interference of waves. The wave nature of mechanical phenomena is not ordinarily apparent, we are told, because the length of the waves involved is ordinarily small compared to the dimensions of the system. It is only in such small scale phenomena as the intimate reactions between atoms and electrons that the wave-lengths are comparable with the dimensions of the system. Here only are we to expect notable departures from classical mechanics, and here only are we to find evidence of a more comprehensive wave mechanics.*

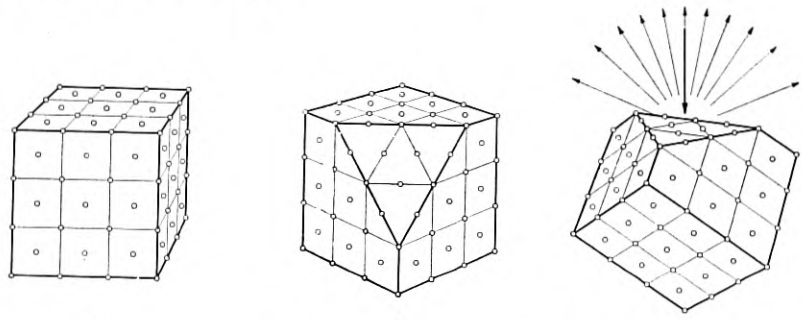
The success of this new theory has been confined, up to the present time, to explanations of certain of the data of spectroscopy. In this field the theory has appealed very strongly to all of us because of the

* It was predicted by W. Elsasser in 1925 (Naturwiss., *13*, 711 (1925)) that evidence for the wave mechanics would be found in the interaction between a beam of electrons and a crystal.

elegance of its methods and because of its remarkable facility in accounting for various of the inhibitions with which the radiating atom is afflicted. We have been prepared by these successes to view with not too great surprise—or alarm—evidence for the wave nature of phenomena involving freely moving electrons. And any reluctance we may feel in treating electron scattering as a wave phenomenon is apt to be dispelled when we find that the value calculated for the wave-length of the equivalent radiation is in acceptable agreement with that which L. de Broglie assigned to the waves which he associated with a freely moving particle—that is to say, the value h/mv (Planck's constant divided by the momentum of the particle).

In this account of the experiments I will describe the general method of the measurements and the general character of the results rather than attempt to go into these matters in detail.

Nickel forms crystals of the face centered cubic type. In Fig. 1 (*a*) the crystal which we had at our disposal is represented by a block of unit cubes of this type.



a *b* *c*

Fig. 1—Diagrams of nickel lattice, of cut lattice, and of lattice with incident and scattered beams

Our first step in preparing the crystal for bombardment was to cut through this structure at right angles to one of the cube diagonals. The appearance of the crystal after the cut was made, and the corner of the cube removed, is indicated in Fig. 1 (*b*). It is this newly formed triangular surface that was exposed to electron bombardment. The bombardment was at normal incidence as indicated in Fig. 1 (*c*). We are to think of electrons raining down normally upon this triangular surface, and of some of these emerging from the crystal without loss of energy, and proceeding from it in various directions.

What is measured is the current density of these full speed scattered electrons as a function of direction and of bombarding potential. The way in which the measurements are made is illustrated in Fig. 2. The electrons proceeding in a given direction from the crystal

enter the inner box of a double Faraday collector and a galvanometer of high sensitivity is used to measure the current to which they give rise. An appropriate retarding potential between the parts of the collector excludes from the inner box all but full speed electrons.

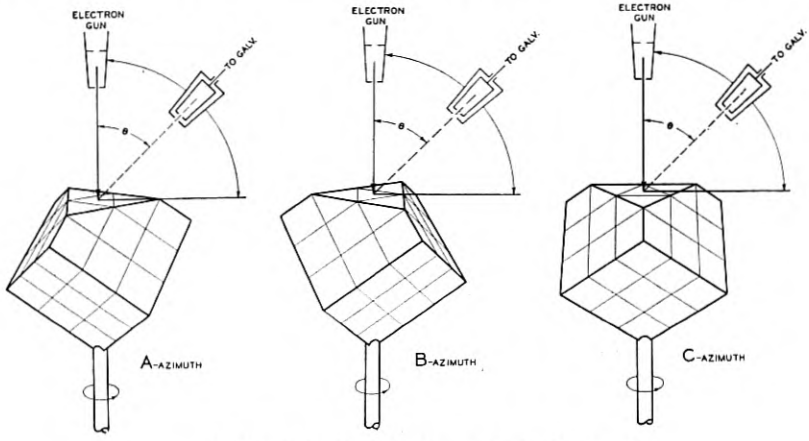


Fig. 2—Showing the three principal azimuths

The collector may be moved over an arc of a circle in the plane of the drawing as indicated, and the crystal may be rotated about an axis which coincides with the axis of the incident beam of electrons. Thus the collector may be set for measuring the intensity of scattering in any direction relative to the crystal—by turning the crystal to the desired azimuth, and moving the collector to the desired colatitude. The whole solid angle in front of the crystal may be thus explored with the exception of the region within twenty degrees of the incident beam.

Certain of the azimuths related most simply to the crystal structure we shall refer to as "principal azimuths." Thus there are the three azimuths that include the apexes of the triangle. If we find the intensity of scattering depending on colatitude in a certain way in one of these azimuths, we expect, of course, to find it depending upon colatitude in the same way in each of the other two. We shall call these the "A-azimuths." On the left in Fig. 2 the crystal has been turned to bring one of the A-azimuths into the plane of rotation of the collector.

Another triad of principal azimuths consists of the three which include the mid-points of the sides of the triangle. These we shall call the "B-azimuths." The next most important family of azimuths comprises those which are parallel to the sides of the triangle; of these there are six, the "C-azimuths."

If we turn the crystal to any arbitrarily chosen azimuth, set the bombarding potential at any arbitrarily chosen value, and measure the intensity of scattering as a function of colatitude, what we find ordinarily is the type of relation represented by the curve on the left in Fig. 3.

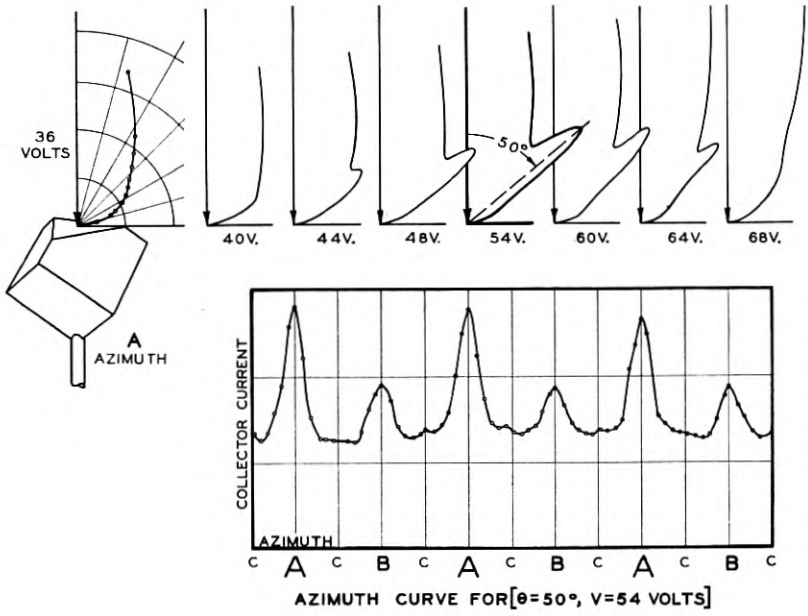


Fig. 3—Curves showing development of diffraction beam in the A-azimuth . . . and variation of intensity with the azimuth at colat. 50° for which beam is strongest in the A-azimuth

This curve is actually one found for scattering in the A-azimuth when the bombarding potential is 36 volts. It is typical, however, of the curves that are obtained when no diffraction beam is showing. The intensity of scattering in a given direction is indicated by the length of the vector from the point of bombardment to the curve. The intensity is zero in the plane of the crystal surface, and increases regularly as the colatitude angle is decreased. This type of scattering forms a background upon which the diffraction beams are superposed.

The occurrence of a diffraction beam is illustrated in the series of curves to the right in Fig. 3. When the bombarding potential is increased from 36 to 40 volts, the curve is characterized by a slight hump at colatitude 60 degrees. With further increase in bombarding potential this hump moves upward, and at the same time develops

into a strong spur. The spur reaches its maximum development at 54 volts in colatitude 50 degrees, then decreases in intensity, and finally vanishes from the curve at about 70 volts in colatitude 40 degrees.

We next make an exploration in azimuth through this spur at its maximum; we adjust the bombarding potential to 54 volts, set the collector in colatitude 50 degrees, and make measurements of the intensity of scattering as the crystal is rotated. The results of this exploration are exhibited by the curve at the bottom of Fig. 3, in which current to the collector is plotted against azimuth. We find that the spur is sharp in azimuth as well as in latitude and that it is one of a set of three spurs as required by the symmetry of the crystal.

We observe also that there are small spurs showing in the B-azimuths. We turn the crystal to bring the B-azimuth under observation, and again make explorations in latitude for various speeds of bombardment. We find that the spur in the B-azimuth is similar to the "54 volt" spur in the A-azimuth, but that it attains its maximum development at a higher voltage and at a higher angle. Curves exhibiting its growth and decay are shown in Fig. 4. Maximum

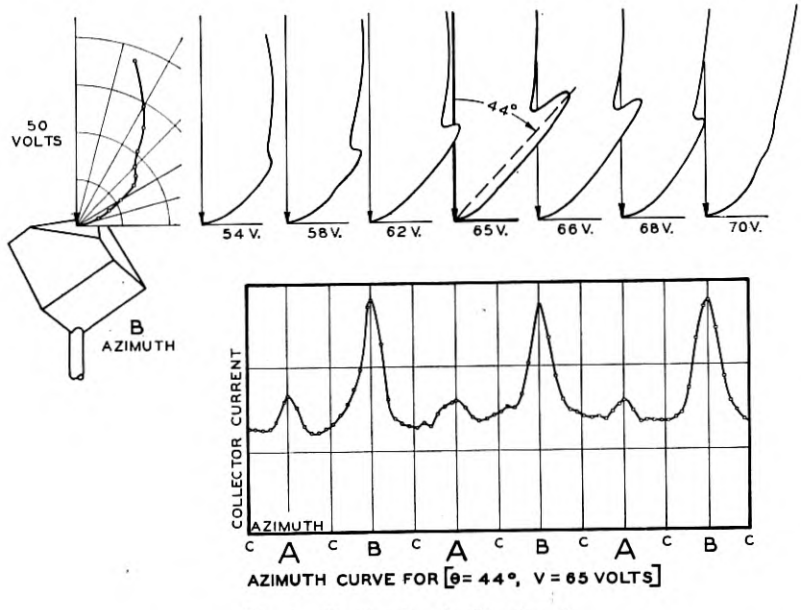


Fig. 4—Similar for the B-azimuth

development is attained at 65 volts in colatitude 44 degrees. At the bottom of the figure we show the intensity-azimuth curve through

this spur at its maximum. The small maxima in the A-azimuths represent the remnants of the "54-volt" spurs.

We have thus a set of spurs at colatitude 50 degrees in the A-azimuths when the bombarding potential is 54 volts and a set of 44 degrees in the B-azimuths when the bombarding potential is 65 volts. These spurs are due to beams of full speed scattered electrons which are comparable in sharpness and definition with the beam of incident electrons. This is inferred from the widths of the spurs and the resolving power of the apparatus.

It is hardly necessary to point out that these sharply defined beams of scattered electrons are similar in their behavior to x-ray diffraction beams. If the incident beam were a beam of monochromatic x-rays of adjustable wave-length instead of a homogeneous beam of electrons of adjustable speed, quite similar effects could be produced. If the wave-length of the x-ray beam were varied, critical values would be found at which intense diffraction beams would issue from the crystal in its A-azimuths and others at which such beams would issue in the B-azimuths. The x-ray diffraction beams would indeed be more sharply defined in wave-length than the electron beams defined in voltage. No diffraction beam would be observed until the wave-length of the incident x-rays were very close indeed to its critical value, and the beam would disappear again when the wave-length had passed only very slightly beyond the critical value. This "wave-length sharpness" or "wave-length resolving power" is dependent, however, upon the number and disposition of the atoms involved in the diffraction. If the crystal were only a few atom layers in thickness, or if the x-rays were extinguished on penetrating through only a few atom layers of the crystal, then the x-ray diffraction beams would be much less sharply defined in wave-length; they would behave more like the electron beams. We may say then that the electron beams exhibit the general behavior of diffraction beams resulting from the scattering of a beam of very soft wave radiation—radiation that is very rapidly extinguished in the crystal.

Let us try now to forget that what we are measuring in these experiments is a current of discrete electrons arriving one by one at our collector. Let us imagine that what we are dealing with is indeed a monochromatic wave radiation, and that our Faraday box and galvanometer are instruments suitable for measuring the intensity of this radiation. We are to think of the incident electron beam as a beam of monochromatic waves, and of the "54-volt beam" in the A-azimuth and the "65-volt beam" in the B-azimuth as diffraction beams that owe their intensities, in the usual way, to constructive

interference among elements of the incident beam scattered by the atoms of the crystal. With this picture in mind we try next to calculate wave-lengths of this electron radiation from the data of these beams and from the geometry and scale of the crystal.

To begin with, we shall need to look more closely into our crystal. The atoms in the triangular face of the crystal may be regarded as arranged in lines or files at right angles to the plane of the A- and B-azimuths. If a beam of radiation were scattered by this single layer of atoms, these lines of atoms would function as the lines of an ordinary line grating. In particular, if the beam met the plane of atoms at normal incidence, diffraction beams would appear in the A- and B-azimuths, and the wave-lengths and inclinations of these beams would be related to one another and to the grating constant d by the well-known formula, $n\lambda = d \sin \theta$, as illustrated at the top of the figure.

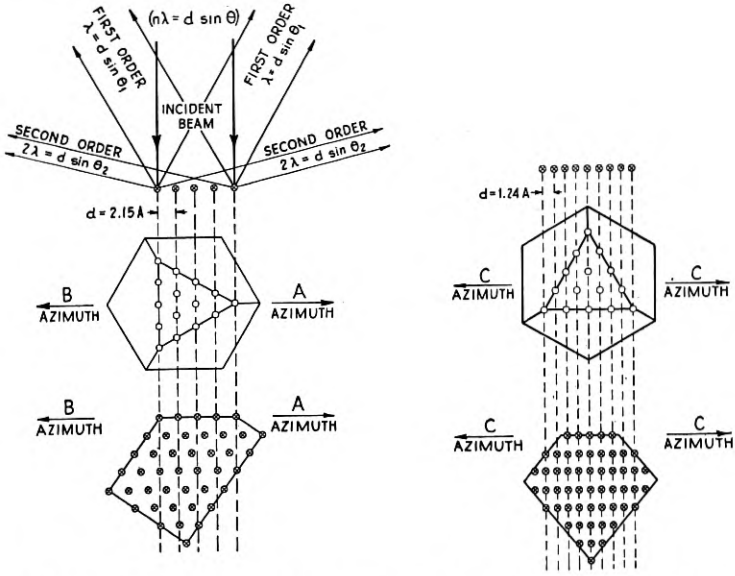


Fig. 5—Showing $n\lambda = d \sin \theta$ relation in the A-, B- and C-azimuths

In the actual experiments the diffracting system is not quite so simple. It comprises not a single layer of atoms, but many layers; it is equivalent not to a single line grating, but to many line gratings piled one above the other, as shown graphically at the bottom of the figure. What diffraction beams will issue from this pile of similar and similarly oriented plane gratings?

The answer to this question is twofold. In respect of position all the beams which appear will coincide with beams which would issue from a single grating. We get no additional beams by adding extra layers to the lattice. In respect of intensity, however, the results are greatly changed. A given beam may be accentuated or it may be diminished, both absolutely and relatively to the other beams; it may in fact be blotted out completely, or reduced to such an extent that it can no longer be perceived. These are effects of interference among the similar beams proceeding from the various plane gratings that make up the pile. Later we shall consider under what conditions these component beams combine to produce a resultant beam of maximum intensity; for the present, however, I wish only to stress the fact that whenever and wherever a space lattice beam appears its wave-length and colatitude angle θ will be related to the constant d of the plane grating through the ordinary plane grating formula. We therefore apply this formula to the 54- and 65-volt beams that have been described. The grating constant d has the value 2.15 Å., the 54-volt beam occurs at $\theta = 50°$ so that $n\lambda$ for this beam should have the value $2.15 \times \sin 50°$, or 1.65 Å. For the 65-volt beam we obtain for $n\lambda$ the value 1.50 Å.

We now compare these wave-lengths with the wave-lengths associated with freely moving electrons of these speeds in the theory of wave mechanics. Translated into bombarding potentials, de Broglie's relation

$$\lambda = h/mv \text{ becomes } \lambda = \sqrt{\frac{150}{V}} \text{ Å.},$$

where V represents the bombarding potential in volts. The length of the phase wave of a "54-volt electron" is $(150/54)^{1/2} = 1.67$ Å., and for a 65-volt electron 1.52 Å. The 54- and 65-volt electron beams do very well indeed as first order phase wave diffraction beams.

It may be mentioned that beams occur at different voltages in the A- and B-azimuths because the plane gratings that make up the crystal are not piled one immediately above the other. There is a lateral shift from one grating to the next amounting to one third of the grating constant. Because of this shift the phase relation among the elementary beams emerging in the A-azimuth is not the same as that among those emerging in the B-azimuth—and coincidence of phase among these beams occurs at different voltages, or at different wave-lengths, in the two azimuths.

We next make similar calculations for a beam occurring in the C-azimuth. One such beam attains its maximum development in

colatitude 56° when the bombarding potential is 143 volts. For diffraction into the C-azimuth we must regard the atoms in the surface layer as arranged in lines normal to the plane of this azimuth as illustrated in Fig. 5. The grating constant is 1.24 Å., and the similar gratings that make up the whole crystal are piled up without lateral shift. For this reason the C-azimuth is six-fold instead of only three-fold. For a beam occurring in this azimuth in colatitude 56°, $n\lambda$ should be equal to 1.24 $\times$ sin 56° or 1.03 Å. The value of h/mv for electrons that have been accelerated from rest through 143 volts is $(150/143)^{1/2}$ or 1.025 Å. Again the beam does very well as a first order diffraction beam.

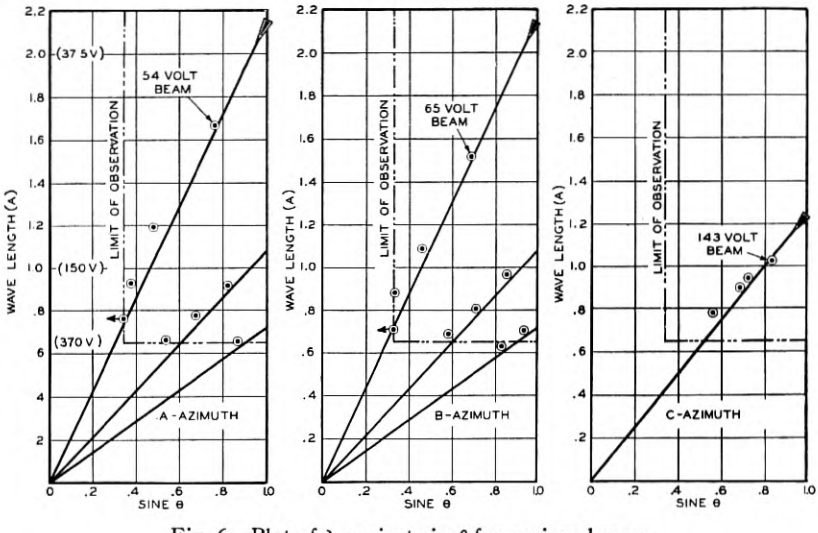


Fig. 6—Plot of λ against sin θ for various beams

The total number of such beams which we have observed in all azimuths in explorations up to 370 volts is twenty-four—nine in the A-azimuth, ten in the B-azimuth, and five in the C-azimuth. It would be possible to calculate an observed wave-length for each of these beams from $n\lambda = d \sin \theta$, and to compare this in each case with the theoretical wave-length calculated from $\lambda = h/mv$, just as we have done already for three of the beams. We have chosen, however, to display the results graphically rather than numerically.

The data for the twenty-four beams are exhibited in diagrams in Fig. 6, in which wave-length λ is plotted against the sine of the co-

latitude, θ. There is a separate diagram for each azimuth and in each the straight lines passing through the origin represent the plane grating formula $n\lambda = d \sin \theta$ in its different orders. Each of the twenty-four beams is represented by a point or by a wedge-shaped symbol in one of these diagrams. The quantities coordinated in each case are the wave-length of the incident beam as calculated from $\lambda = h/mv = (150/V)^{1/2}$ and the sine of the colatitude angle of the diffraction beam as observed. There are no points to the left of the line $\theta = 20°$ as this represents the lower limit of our colatitude range of observation, and none below the line $\lambda = 0.637$ Å. as this corresponds to the upper limit of our voltage range, 370 volts. The bombarding potentials corresponding to various wave-lengths are shown by figures enclosed in brackets.

When the data are exhibited in this fashion the question as to whether or not the observed wave-length of a beam agrees with its theoretical wave-length is answered by whether or not the point representing the beam falls on one or another of the lines representing the plane grating formula. If there were perfect agreement in all cases, each of the points would lie on some one of these lines.

It will be seen that the points all lie close to the lines, though not as a rule exactly on them. It is of course very important to decide whether the departures of the points from the lines are or are not too great to be attributed to uncertainties of measurements. It is our belief that they are in fact due to experimental error in the determination of the colatitude angles. If we accept the theoretical values of the wave-lengths as correct, and calculate the values of θ which we should have observed, we find that in no case do they deviate by more than 4 degrees from the values of θ actually set down. Corrections of this magnitude do not seem excessive when it is considered that we are making measurements with what amounts to a rather crude spectrometer, that the arm of the spectrometer is but 11 mm. in length, that the opening in the collector is 5 degrees in width, and that the spectrometer itself is sealed into a glass bulb. We therefore assume that in every case the value of the wave-length assigned by de Broglie is the correct one.

I now direct your attention to a particular group of these beams—the group comprising the beam of greatest wave-length in each of the three azimuths, which are represented in the figure by wedge-shaped symbols. The interpretation of these three is quite simple.

The radiation to which our electron beam is equivalent is extremely soft as already noted. Its intensity suffers a considerable decrement when the beam passes normally through only a single layer of atoms. This characteristic is inferred from the low resolving power of the crystal, and is consistent with what we know of the penetrating power of low speed electrons. When the beam passes through a layer of atoms at other than normal incidence the decrement in its intensity is greater still—and in the limit as the angle of incidence approaches grazing to the atom layer the intensity of the transmitted beam will approach zero. Thus we may expect that when a diffraction beam leaves the crystal at near grazing emergence the contributions to the resultant beam which come from the second and lower layers of atoms will be much less important than when the beam emerges from the crystal at a higher angle. Near grazing the radiation proceeding from the second and lower layers will be heavily absorbed in its passage through the overlying layers. Within a limited angular range near grazing the diffraction beam will be made up almost entirely of radiation scattered by the uppermost layer of atoms. The diffracting system becomes essentially a single plane grating and what we should observe is ordinary plane grating diffraction.

The first order diffraction beam from a line grating appears at grazing emergence when the wave-length of the incident radiation is equal to the grating constant. The grating constant for diffraction into the A- and B-azimuths is 2.15 Å. and grazing beams should appear in both azimuths when the wave-length of the incident electron beam has this value. The bombarding potential corresponding to wave-length 2.15 Å. is 32.5 volts, and at just 32.5 volts diffraction beams appear at grazing in both these azimuths. As the bombarding potential is increased the beams move up from the surface to satisfy the relation $\lambda = d \sin \theta$. Ten or fifteen degrees above the surface radiation from the second and lower layers escapes in sufficient amounts to reduce the intensity of the resultant beam through interference, and at a somewhat higher angle the beam disappears.

An exactly similar beam is found at grazing in the C-azimuth. The grating constant here is 1.24 Å. and the bombarding potential corresponding to wave-length 1.24 Å. is 97.5 volts. The beam appears at grazing at just this voltage. These three beams occurring and behaving exactly as required by the theory constitute the strongest evidence we have in favor of the wave interpretation of electron scattering.

We have been less successful in trying to account for the occurrences of the remaining 21 sets of beams. We do not know why they occur

where they do. The most we have been able to do is to relate their occurrences with those of the Laue beams that would issue from the same crystal if the incident beam were a beam of x-rays.

In Fig. 7 we indicate by crossed circles in a (λ, sin θ) diagram the x-ray diffraction beams that would be observed in the B-azimuth. We show also again the electron beams as actually observed. It is obvious that the law of occurrence of electron beams is not the same as the law of occurrence of Laue beams, and yet we see that the occurrences of the two sets of beams have certain features in common. The dots representing electron beams occur along the plane grating lines at about the same intervals as the crossed circles representing the Laue beams. Other points of similarity are found with further study of the data and one is led finally to the conviction that each electron beam is the analogue of a particular Laue beam. The electron beam represented by a given dot appears to be the analogue of the Laue beam of the same order represented by the crossed circle occurring next above it in the diagram. This association of beams is indicated in the figure.

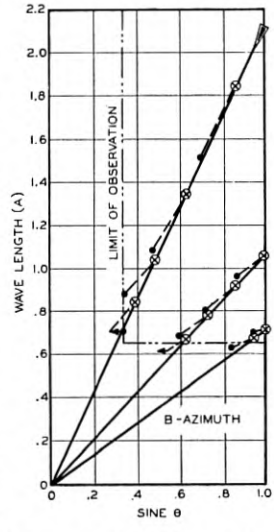


Fig. 7—λ sin θ diagram for B-azimuth

The occurrences of the Laue beams are determined in part by the separation between the atomic plane gratings that make up the crystal. If the separation between adjacent planes were increased the crossed circles representing the Laue beams would be moved upward along the plane grating lines; if the separation were decreased the crossed circles would be moved downward. Merely as a mode of description, then, we may say that a given electron beam has the wave-length and position that its Laue beam analogue would have if the separation between planes were decreased by a certain factor.

We have calculated this spacing factor for each of the 21 beams and the values found are plotted in the upper part of Fig. 8 against the voltages of the beams. The points form a very bad curve. They do indicate, however, that the factor increases with the speed of the electron, and there is the suggestion that it approaches unity as a limiting value. There is the suggestion, that is, that at high voltages the law of occurrence of electron beams is the same as the law of occurrence of Laue beams.

It has been pointed out by Eckart that if the index of refraction of the crystal for the electron radiation is other than unity diffraction beams will occur as if the separation between atom planes were other than normal. We have computed the indices of refraction that would

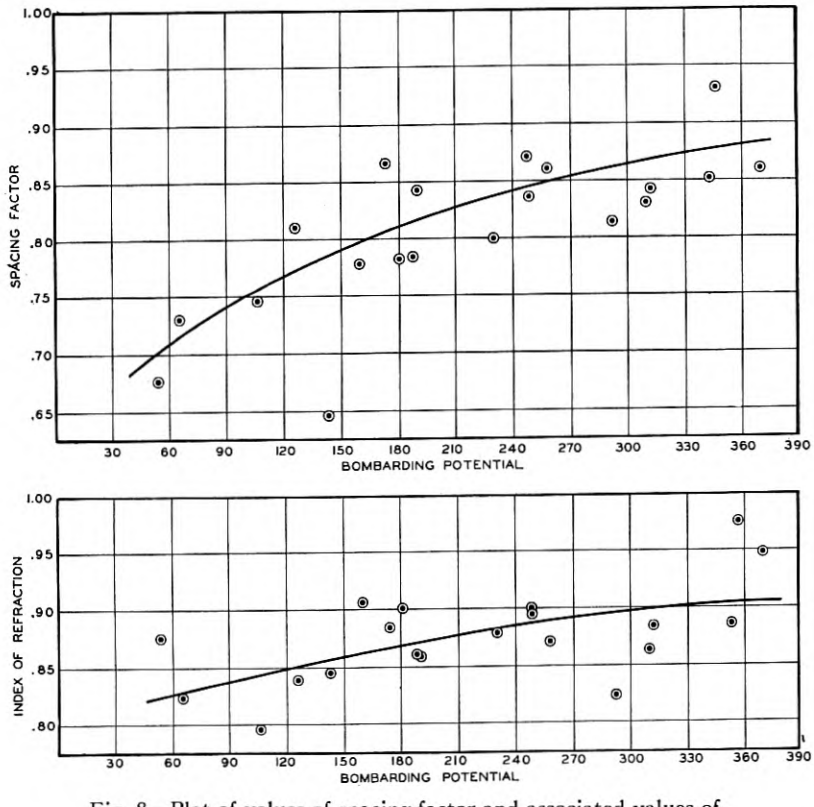


Fig. 8—Plot of values of spacing factor and associated values of refractive index for twenty-one beams

give rise to the observed occurrence of beams and these are plotted in the lower part of the diagram against bombarding potential. Again the points fall very irregularly. While it cannot be said that there is at present a satisfactory explanation of the peculiar occurrence of the space lattice electron diffraction beams, it should be clearly understood that this deficiency in no way affects either the wave-length measurements of these beams or the agreement of these wave-lengths with the values of h/mv.

The electron diffraction beams which I have described are the only ones observed when the surface of the crystal is free from gas. When the surface is not free from gas still other beams appear. These

beams are due to the scattering of electrons by the adsorbed gas and therefore we shall not consider them at this time.

In closing I should like to say a few words about the conceptual difficulty in which these experiments involve us. When Laue and his collaborators investigated the scattering of x-rays by crystals the results of their observations were accepted at once as establishing the wave theory of x-rays. It was a very simple matter for W. H. Bragg and others to give up the corpuscular theory because of the hypothetical nature of the x-ray corpuscle. It was only necessary to recognize that Laue's results were contrary to hypothesis and the corpuscle disappeared.

If the electron were not the well-authenticated particle we know it to be, it is possible that the experiment I have described would cause it to vanish in like manner. We do not, however, anticipate any such event. The electron as a particle is too well established to be discredited by a few experiments with a nickel crystal. The most we are apt to allow is that there are circumstances in which it is more convenient to regard electrons as waves than as particles. We will allow perhaps that electrons have a dual nature—when they produce tracks in a C. T. R. Wilson cloud experiment they are particles, but when they are scattered by a crystal they are waves.

A quite similar situation exists, of course, in the case of x-rays. It has been evident for some years that the adherents of the corpuscular theory of x-rays were too enthusiastic in their recantations. X-rays also exhibit a dual nature—when they give rise to diffraction patterns they are waves, but when they exhibit the Compton effect or cause the emission of electrons from atoms they are particles—quanta or photons.

This state of affairs is one that should appeal to us as intolerable. There must, it would seem, be comprehensive modes of description applicable to all electron and x-ray phenomena, but what these are we do not yet know. We do not know whether we shall eventually believe with de Broglie and Schroedinger that electrons and x-rays are waves that sometimes masquerade as particles, with Duane that electrons and x-rays are particles that sometimes masquerade as waves, or whether eventually we shall believe with Born that we are dealing in both cases with actual particles and phantom waves.

I believe, however, that for the present and for a long time to come we shall, in describing experiments, worry but little about ultimate realities and logical consistency. We will describe each phenomenon in whatever terms we find most convenient.

Grid Current Modulation

By EUGENE PETERSON and CLYDE R. KEITH

SYNOPSIS: The term grid current modulator is used to describe those vacuum tube circuits in which modulation is initially produced in the grid circuit of a three-electrode vacuum tube due to the non-linear grid current-grid voltage relation. Comparison with a representative plate current modulator using the same tubes and the same plate potential shows that by modulating at maximum efficiency in the grid circuit and using the plate circuit solely for amplification, the maximum power output is increased about eight times, the power efficiency is increased about five times and the ratio of sideband output to signal input is increased approximately three times. Under these conditions more carrier input power is needed for the grid than for the plate modulator. This improved performance has been made possible by a detailed study of the fundamental processes involved and by a design of the tubes and associated equipment, such as transformers and filters, to permit these fundamental processes to operate to their best advantage.

Normally modulation is also produced in the plate circuit which is shown to be out of phase with that produced in the grid circuit. By inserting high impedances to the input frequencies in the plate circuit, plate circuit modulation is prevented, and the reduction of grid circuit sideband is likewise avoided. By including in the grid circuit an impedance which is high to the desired sideband frequencies, the maximum grid sideband voltage is obtained. In this way the power and modulating efficiencies of the tube circuit are made maximum.

Where modulation occurs only in the plate circuit of a tube, the sideband amplitude is proportional to the product of the amplitudes of the input frequencies when these amplitudes are small. In the present type of grid current modulator the sideband amplitude is proportional to the smaller of the two input amplitudes provided the ratio between these is greater than about 3/2. This feature makes the modulator particularly valuable in communication systems.

The article concludes with an application of the fundamental principles involved to an experimental carrier telephone system in which the operating features of tubes, filters, and transformers are discussed.

INTRODUCTION

BECAUSE of the extensive application of vacuum tube modulators in systems of carrier communication, they constitute an important tool in the hands of telephone engineers. As such they have justified extensive laboratory investigation. The purpose of this paper is to discuss some of the properties of a type of modulator utilizing the non-linear relation existing between grid voltage and grid current and the advantages which recent laboratory investigations indicate that it may possess. Further studies are in progress to determine the conditions under which it can be employed practically.

There are two distinct classes of vacuum tube modulators which may be designated for convenience as grid and plate types, according to the circuit in which the modulation is initially produced, although some modulators may involve both circuits. As an example of the plate

type we might mention the coupled-plate or Heising modulator which has found extensive application in radio transmitters, in which the plate circuit of an oscillating tube is coupled to the plate of another tube through which the signal is introduced, modulation taking place ordinarily in the plate circuit of the oscillating tube. A carrier frequency amplifier is sometimes used in place of the oscillator. Another type of plate modulator, due in principle to van der Bijl, which has found extensive application in carrier telephone systems, applies the signal and carrier to the grid of the modulator, so that the two components are amplified in common before being modulated. As examples of the grid modulator there are the grid leak and condenser type used almost universally for radio reception, together with the type which forms the subject of the present paper, employing a generalized impedance in the grid circuit. The three last-mentioned modulators may incidentally produce modulation in both plate and grid circuits. This ordinarily acts to reduce the overall efficiency as well as to introduce other undesirable features of operation, so that in the design of the grid current modulator we have been led to minimize modulation in the plate circuit, operating the grid circuit as a modulator and the plate circuit purely as an amplifier.

The criteria of usefulness of modulators include some usually placed upon vacuum tube apparatus in general, together with those peculiar to frequency change; some of most importance are modulating gain and level, plate power efficiency, quality, stability, input and output impedances, and carrier suppression. These will be taken as a basis for discussing the operation of the grid current modulator and comparing it with that of the other types mentioned above. The modulating gain usually expressed in transmission units (T. U.) represents the ratio of the power output of a single sideband to the power input of the signal which produces it. It is a function of both carrier and signal amplitudes, usually decreasing at high amplitudes. This decrease in gain should not be too rapid or the modulated output power at sufficiently large signal amplitudes may actually decrease as the signal is increased, and lead to prohibitive distortion. Another aspect of the question relates to the maximum modulated power attainable. The signal amplitude fluctuates within wide limits in course of operation and it becomes desirable to limit its effects, so that the resultant modulated potentials may not disturb the operation of associated equipment. This may be accomplished in modulators in which the sideband output approaches an asymptotic maximum as the signal is increased, better than in those which pass through a maximum in the operating range. A knowledge of the signal amplitude and the

modulating gain corresponding to it is sufficient for a determination of the output amplitude or output level which is of prime importance in matters relating to noise and interference. Other significant factors from the standpoint of noise and interference are the closeness with which line and connecting apparatus impedances are matched since this determines the amount of reflection of an incident wave,[1] and the extent to which carrier current is transmitted to the line (carrier leak) in carrier suppression systems.

It is highly desirable of course to have the efficiency of energy conversion from the plate battery to the sideband output power as high as possible, since the amount of power supplied to the plate circuit is thereby minimized, and the necessary power capacity of the plate supply is reduced. Another kind of plate efficiency in which we are sometimes interested is the efficiency of energy transfer from the plate supply to the external plate impedance. This tells us the amount of energy dissipated by the plate of the tube and fixes its structure; it differs from the first efficiency only when other current components than the sideband flow in the plate circuit. Inasmuch as we shall deal in the following with low power tubes which have ample load capacity, only the first-mentioned efficiency is important.

We are also interested ordinarily in the quality of the system. This is determined in large part by the width of the transmitted frequency band, by the presence of new interfering frequencies introduced by the process of modulation, and by the linearity of the modulated output in terms of the signal input. The first and last conditions are equivalent to the requirements that the modulating gain be maintained both over the ordinary range of signal input amplitudes, and over the frequency band essential for good signal reproduction. Finally the system as a whole is required to have a high degree of stability, so that ordinary variations of battery potentials, or even the replacement of a tube by another of the same type, will not impair the operation of the system.

The specific forms of grid current modulator with which we shall be concerned as an application of the theory are those adapted to carrier current telephony, in which the carrier current is suppressed and a single sideband transmitted. Comparison with a representative plate current modulator using the same tubes at the same plate potential shows that, by modulating at maximum efficiency in the grid circuit and using the plate circuit solely for amplification, the maximum power output is increased eight times, the power efficiency is increased five times, and the ratio of sideband output to signal input is increased

[1] The reflection is measured by the quotient of the difference by the sum of the two connected impedances; this is known as the reflection coefficient.

approximately three times. From these figures it is evident that the space current or plate power supplied the grid modulator is sixty per cent greater than it is for the plate modulator, and that this greater power is utilized five times more efficiently. Under these conditions more carrier input power is needed for the grid than for the plate modulator. Where the carrier oscillator employs a tube of the same type as that used in the modulators, sufficient carrier power is, however, available. This improved performance has been made possible by a detailed study of the fundamental processes involved, and by a design of the tubes and associated equipment, such as transformers and filters, to permit these fundamental processes to operate to best advantage.

In view then of the close interdependence of the circuit elements, we shall start with a discussion of the theory as developed for the simplest circuits, and accompany it by approximate mathematical analyses wherever it appears profitable. No rigorous mathematical treatment appears to be possible or even desirable because of its complexity in the general case, and the sole purpose of our approximate analyses is to help in building up a physical picture of the operation of modulators. With the theoretical conclusions in mind, the characteristics of tubes, transformers, retard coils, balanced circuits, and filter networks which are important in this connection are examined, and the performance of the complete carrier telephone modulator circuit is covered in some detail. The theoretical conclusions are not limited to the carrier telephone modulator which has been used simply for illustrative purposes; as a matter of fact the same principles have been found operative in different types of circuits over a wide frequency range. It should be noted that we have not attempted to combine the oscillating and modulating functions in a single circuit as is sometimes done, but have maintained these circuits distinct from one another, so that the best performance of each may be realized.

Theory of Vacuum Tube Modulation

In a broad sense the same general phenomena are involved in both grid and plate modulation, since modulation is produced when an impedance is varied in accordance with the amplitudes of the modulating potentials, a condition true of both circuits under appropriate conditions. Thus conductive grid current is suppressed at negative potentials in the high vacuum tubes which we employ, and it flows when the grid potential is positive, the grid impedance depending upon the amplitude of the grid potential.

We can obtain a qualitative idea of the situation when we consider

the grid circuit connected to an a.c. generator in series with a high resistance of the order of a megohm, and with the plate circuit connected to a resistance of the same order of magnitude as the internal plate resistance. With a negative grid the input impedance is mainly capacitive and at low frequencies nearly the entire applied e.m.f. exists across the grid. At positive potentials however, when conductive grid current flows and the grid resistance drops to something like 10,000 ohms, by far the greater part of the drop is taken up in the external resistance so that the positive lobe of the grid potential wave is distorted. This distortion is equivalent to modulation since it implies the presence of new frequencies. As a result of the varying reaction of the tube then, modulation voltages are built up across the grid-filament path. These potentials are amplified in the plate circuit where the applied wave suffers further distortion due to the non-linear relation between plate current and grid potential, which in a similar way gives rise to plate modulation. Evidently plate modulation would be alone effective if the external grid resistance were made small.

General Relations

The production of modulated currents or potentials is characteristic of any device in which the relation between instantaneous values of current and voltage is not a linear one. Theoretically, such a relation can be represented to any required degree of accuracy by an equation of the form

$$i = a_0 + a_1 v + a_2 v^2 + a_3 v^3 + \cdots, \tag{1}$$

in which i represents the current through the device and v the potential drop across it. Now suppose a voltage wave which includes two components of frequency $p/2\pi$ and $q/2\pi$ respectively to be impressed on the non-linear element

$$v = P \cos pt + Q \cos qt. \tag{2}$$

If we substitute this expression in eq. 1 for v, the current wave is found to include components of the two original or fundamental frequencies, together with new frequencies produced by the non-linear element:

$$\begin{aligned} i = i_0 &+ i_p \cos pt + i_q \cos qt \\ &+ i_{2p} \cos 2pt + i_{2q} \cos 2qt \\ &+ i_+ \cos (p+q)t + i_- \cos (p-q)t \\ &+ i_{3p} \cos 3pt + i_{3q} \cos 3qt \\ &+ i_{2p+q} \cos (2p+q)t + i_{2p-q} \cos (2p-q)t \\ &+ i_{p+2q} \cos (p+2q)t + i_{p-2q} \cos (p-2q)t + \cdots, \end{aligned} \tag{3}$$

in which the coefficients i_k involve the characteristic constants of the tube, together with the applied potential amplitudes:

$$\begin{aligned} i_0 &= a_0 + a_2(P^2 + Q^2)/2 + \cdots, \\ i_p &= a_1P + 3a_3P(P^2 + 2Q^2)/4 + \cdots, \\ i_q &= a_1Q + 3a_3Q(Q^2 + 2P^2)/4 + \cdots, \\ i_{2p} &= a_2P^2/2 + \cdots, \\ i_{2q} &= a_2Q^2/2 + \cdots, \\ i_+ &= i_- = a_2PQ + \cdots, \\ &\vdots \end{aligned} \tag{4}$$

The new frequencies produced are made up of sums and differences of integral multiples of the two original frequencies, and an inspection of eq. 3 shows that the frequency of any component may be put in the form

$$|mp \pm nq|/2\pi \qquad m, n = 0, 1, 2.\cdots$$

It is convenient to designate the sum of the two numbers m and n as the order of a wave component, so that the frequencies $2p/2\pi$, $2q/2\pi$, and $(p \pm q)/2\pi$ are products of the second order. The last of these serves as the basis for the operation of all present [2] carrier systems, and the rôle of any modulator is therefore to produce one or both of these components which are known as side frequencies, or as sidebands when the signal wave is made up of a band of frequencies. Now by repeating the modulating process, but this time with the frequencies $(p + q)/2\pi$ or $(p - q)/2\pi$, or both, together with the component of frequency $p/2\pi$, designated as the carrier wave, it is well known that one of the resultant second order products has the frequency of the original signal $q/2\pi$. This second or receiving modulator, sometimes designated as a demodulator or detector, is separated from the first one by a transmitting medium and frequency-selective apparatus so that only the desired components may be transmitted and received. The impedance-frequency characteristics of these elements with which the modulators are associated are of prime importance in determining the modulation, as is best brought out by a discussion of some approximate mathematical analyses which follow.

Modulator Circuit, Small Alternating Potentials

We shall consider the current-voltage characteristic of a vacuum tube to be given, to sufficient accuracy for our purposes, by the first

[2] Higher order products, such as $(2p \pm q)/2\pi$ have been equally well employed but we shall confine our attention here to the usual second order system.

three terms of eq. 1, and shall suppose two generated potentials, of frequency $p/2\pi$ and $q/2\pi$ respectively, to be applied to a tube circuit which includes a series impedance. This impedance Z_k may be a function of frequency as indicated by the subscript k, which refers to the particular frequency at which the impedance is effective. The variable part of eq. 1—the change in current produced by application of the alternating potentials—is clearly

$$J = a_1 v + a_2 v^2. \tag{5}$$

The potential drop across the tube is that impressed minus the Zi drop or

$$v = \Sigma(E_k - Z_k J_k) \tag{6}$$

the summation extending over all current components. As a first approximation to the fundamental currents we may neglect the non-linear term $(a_2 v^2)$ in eq. 5 to obtain

$$\begin{aligned} J_p &= a_1 E_p/(1 + a_1 Z_p), \\ J_q &= a_1 E_q/(1 + a_1 Z_q). \end{aligned} \tag{7}$$

Using these solutions we can obtain a second approximation [3] taking into account the non-linear term which we neglected for the first approximation. Thus

$$\begin{aligned} J_p + J_q + J_2 = a_1 \left(\frac{E_p}{1 + a_1 Z_p} + \frac{E_q}{1 + a_1 Z_q} - Z_2 J_2 \right) \\ + a_2 \left(\frac{E_p}{1 + a_1 Z_p} + \frac{E_q}{1 + a_1 Z_q} \right)^2, \end{aligned} \tag{8}$$

in which the subscript 2 indicates the second approximation. By squaring the second member of eq. 8 it is observed that the second approximation includes direct current, second harmonics of the two impressed frequencies $p/2\pi$ and $q/2\pi$, and the second order sidebands. For these last we obtain the expression

$$J_{\pm} = \frac{a_2}{(1 + a_1 Z_{\pm})} \frac{E_p \dot{E}_q}{(1 + a_1 Z_p)(1 + a_1 Z_q)}. \tag{9}$$

The sideband potential across the variable element is clearly $Z_+ J_+$ or $Z_- J_-$ according to the particular sideband in which we are interested. Thus

$$V_- = -\frac{a_2}{a_1^2} J_p J_q \frac{Z_-}{1 + a_1 Z_-} = -\frac{a_2}{a_1^2} J_p J_q \frac{Z_-}{R_0 + Z_-}, \tag{10}$$

where $1/a_1$ is equivalent to R_0, the plate resistance of the tube.

[3] Carson, "A Theoretical Study of the Three Element Vacuum Tube," *Proc. I. R. E.*, 1919.

These equations are equally applicable to grid and to plate circuits provided the potentials are small and the operating region is expressible by eq. 5. This is not always the case in practice, and modifications in the above comparatively simple analysis are required, which will be treated below. A number of characteristic features of operation are exhibited by the above analysis however and these we proceed to discuss.

If the preceding treatment is applied to the grid circuit we see that in order to make the sideband potential across the grid a maximum with fixed fundamental currents, the external grid impedance at the sideband frequency must be made large compared to the effective internal resistance of the tube. Further, if the generator potentials and impedances are fixed it follows that the generator resistance should be made to match the internal resistance of the tube at the fundamental frequency—with a transformer, if necessary—in order to make the fundamental currents as large as possible. This conclusion regarding the ratio of grid impedances follows immediately without mathematical analysis if we suppose the source of the higher order products to lie in the variable impedance element so that it may be considered equivalent to the presence of generators of the higher order frequencies. The generator voltage is evidently maximum on open circuit, which agrees with the above statement.

In considering the form of external impedance to use for best results, an inspection of eq. 10 shows that in the quantity $Z_-/(R_0 + Z_-)$, which expresses the ratio of effective grid voltage to generated grid voltage, the sideband impedance Z_- should have a large reactive component. This is illustrated by Fig. 1 which gives the ratio for various relative external impedances having phase angles of 0 and 90° and shows that, with the external impedance fixed in magnitude, the ratio has its greatest value for a pure reactance.

Relative Phase of Grid and Plate Sidebands

If the tube acted as a perfect amplifier of the potentials impressed on the grid, there would be no further distortion and the sideband current in the plate circuit would be obtained by multiplying eq. 10 by $\mu/(Z + R_0)$, where Z and R_0 are the external and internal plate circuit resistances respectively, and μ is the amplification factor which is assumed constant here.[4] Unfortunately this ideal situation does not exist of itself, and modulation of the amplified fundamentals takes place in the plate circuit, producing an additional sideband component to

[4] The distortion due to variable μ as treated by Peterson and Evans in the *Bell System Technical Journal* for July, 1927, represents but a small part of the total in efficient modulators, although it is of importance in high quality amplifiers.

combine with that generated in the grid circuit and amplified in the plate circuit.

Inasmuch as the amplification factor decreases, and the plate impedance increases as the grid potential goes negative, the grid

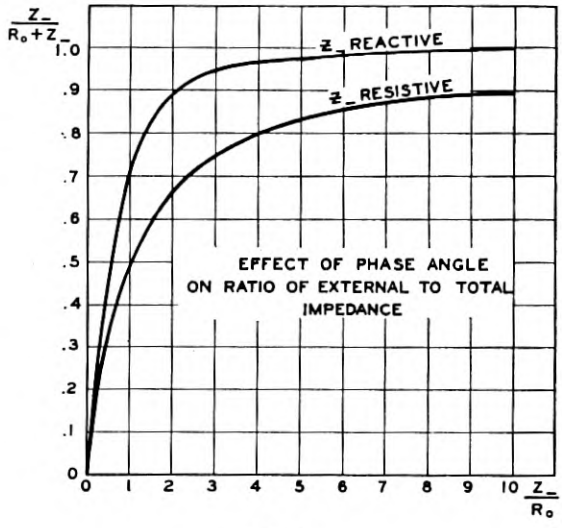


Fig. 1

potential wave is amplified more efficiently on the positive than on the negative lobe, with the result that the plate current wave is limited on one side by the grid current cut-off, and on the other side by plate current cut-off. The second cut-off tends to make the output wave more nearly symmetrical about a horizontal axis; it is therefore equivalent to an increase in the odd order modulation which we do not employ here, and to a reduction of the even order products, one of which—the second order sideband—is used to transmit signal characteristics. It follows that for efficient modulation we must do one of two things—phase the grid and plate products to add, or remove one of the conflicting sources of modulation.

To account for the effect of plate distortion we may apply the same general procedure to the plate circuit as we did to the grid circuit. The plate current-grid potential relation is given as

$$J = b_1\mu v + b_2\mu^2 v^2 \tag{11}$$

and the solution for the current components may be written down directly since the problem presents itself in the same form as the grid circuit situation previously considered. Hence if we change the a

coefficients to b's, and the Z_k of the grid circuit to the Z_k of the plate circuit, we obtain the expressions

$$\left.\begin{aligned} J_1 &= b_1\mu v/(1 + b_1Z_1), \\ J_2 &= \frac{b_2}{(1 + b_1Z_2)}\left(\frac{\mu v}{1 + b_1Z_1}\right)^2. \end{aligned}\right\} \tag{12}$$

Now it is apparent that each of these two terms contributes something to the sideband frequency—the first by amplification of the grid sideband potential, and the second by modulation of the two fundamental components in the plate circuit. The net sideband current in the plate circuit may accordingly be expressed as

$$J_\pm = \left[\left(\frac{\mu b_2}{(1 + b_1Z_p)(1 + b_1Z_q)} - \frac{b_1a_2Z_\pm}{1 + a_1Z_\pm}\right)\right] \left[\frac{\mu E_pE_q}{(1 + b_1Z_\pm)(1 + a_1Z_p)(1 + a_1Z_q)}\right]. \tag{13}$$

Under normal conditions both grid current and plate current characteristic curves are concave upward, so that b_1, b_2, and a_2 are all positive. The two terms of eq. 13 are then in phase opposition, a condition which is responsible for failure to work certain modulators to fullest advantage. For efficient plate modulators the grid modulation term should be suppressed, which may be accomplished by making the external grid impedance to the modulated product of interest equal to zero, or by keeping the grid potential negative at all times. For efficient grid modulators the plate modulation term should be reduced to a minimum by suppressing the fundamental currents in the plate circuit, in which case Z_p and Z_q are made large. Of course the possibility exists of phasing the two sideband components to add rather than to subtract (arithmetically)—and this, it will be readily seen, is obtained by having the phase angle of the entire plate circuit approach 90° at each fundamental frequency when the grid circuit sideband impedance is large. This condition cannot be met without lowering the amount of plate current modulation, so that the first mentioned plate circuit condition is the more practical one.

Other possibilities of more favorable phasing exist by working within appropriate regions of tube operating characteristics where either b_2 or a_2 becomes negative. Generally speaking, operating points of this nature are not stable with variations in tube potentials, nor are they adaptable to large power outputs approaching the maximum load capacity of the tube. Finally, for straight amplification purposes the two terms of eq. 13 should be made equal and opposite in sign.

In order to test directly the conclusions regarding relative phase of grid and plate modulation products, a circuit was set up which permitted two frequencies to be supplied to the grid of a vacuum tube, the resultant currents of sideband frequency being measured in grid and plate circuits by means of a current analyzer.[5] As shown in Fig. 2 the

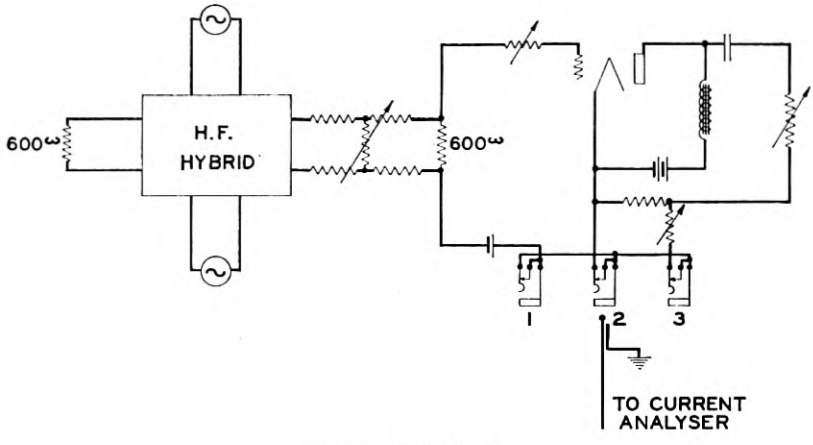


Fig. 2. Test Circuit

grid circuit contains a high external resistance (for producing grid current modulation when conductive grid current flows as previously explained) in series with a "C" battery to vary the relative amounts of grid and of plate modulation. The relative phase of sideband currents produced in grid and plate circuits was calculated from the currents measured separately in jacks No. 1 and No. 3 and their vector sum in jack No. 2. These measurements verified the conclusions drawn from eq. 13,—that with resistances in grid and plate circuits second order modulation products produced in the grid circuit are exactly out of phase with the same frequencies produced in the plate circuit.

The effect of the grid circuit resistance when conductive current flows is of course to limit the positive potentials applied to the grid and so, in effect, to cause the input-voltage—output-current relation of the circuit to be deflected at the upper end more nearly to parallelism with the x-axis than it is for the tube alone. We may therefore consider grid modulation as equivalent to the introduction of a reversed curvature in the operating characteristic. To substantiate this point a tungsten filament tube was used in which the curvature of the lower branch is nearly the same as that of the upper branch, as shown in Fig.

[5] "Analyzer for Complex Electric Waves," by A. G. Landeen, *Bell System Technical Journal*, April, 1927.

3. The plate circuit sideband was measured as a function of the "C" battery with zero grid resistance so that the plate circuit was responsible for the total sideband production; the data are plotted on the

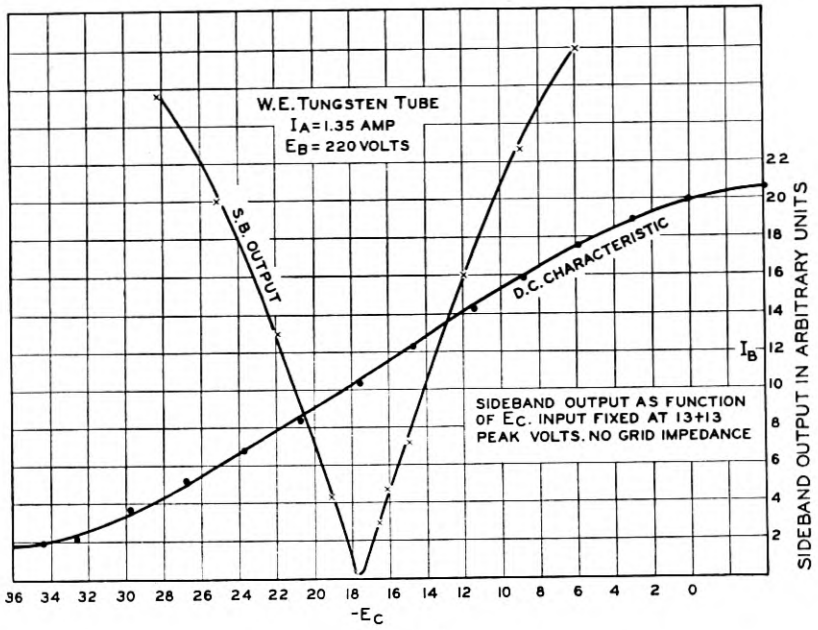


Fig. 3

same Figure. It is seen that the sideband drops nearly to zero when the "C" battery is adjusted so that the input voltage swings symmetrically over the upper and lower branches of the curve. These results could very well be attributed to out-of-phase modulation resulting from reversed curvature. As a matter of fact the algebraic expression for the characteristic involves in general a series of both odd and even powers of applied voltage, but if the axis is taken at the point of symmetry of the characteristic the even powers drop out. Now since the even orders of modulation can be attributed only to the even powers of the static equation, it might be expected that these components would drop to zero.

We may conclude from this discussion that best results will be had in the practical design of grid current modulators, when the external grid impedance at the sideband frequency is made as high as possible and when the impedance to the fundamentals is matched. As to the plate impedances, the situation is the reverse of that existing in the grid circuit since we must have the impedances to the two modulating

frequencies as high as possible, and match the tube impedance at sideband frequencies in order to develop maximum sideband power in the load resistance. It will be observed that with these conditions satisfied, the plate circuit of the tube acts substantially as an amplifier of the sideband produced in the grid circuit, since none is developed in the plate circuit.

Reaction of Sideband Flow on Tube Impedance

In any non-linear system such as the grid circuit or the plate circuit of a tube, the modulation products resulting from the lack of linearity have amplitudes which depend upon one or more of the impressed fundamentals, and react upon the fundamental amplitudes. It follows that the amplitude of any modulation product depends in general upon the amplitudes of all other modulation products, and that the impedance offered to the flow of fundamental depends upon the reaction of the modulation products. Stated otherwise, the amplitude of any one current component depends upon all other components.

This may be demonstrated quantitatively by higher approximations than the two which we have already obtained, in which expressions for the currents are found to contain terms proportional to the sideband voltage across the tube; in fact, if we went to the labor of including a number of distorting components, terms in the fundamental current equation due to their reaction would result. The effects found with a single sideband are simply typical. If, for example, we put (7) and (9) in (8), we get

$$J_p = a_1(E_p - Z_p J_p) - a_2(E_q - Z_q J_q)\,\frac{a_2 E_p E_q Z_+}{(1 + a_1 Z_p)(1 + a_1 Z_q)(1 + a_1 Z_+)},$$

and a similar expression for J_q, as second approximations to the fundamentals. These furnish us with a pair of simultaneous cubics in J_p and J_q. When we assume the reaction of the sideband flow on J_q to be small so that eq. 7 remains valid, the above equation becomes linear in J_p,

$$J_p = E_p \Big/ \left(\frac{1}{a_1} + Z_p + \frac{a_2^2}{a_1^4} J_q^2 \frac{Z_+}{1 + a_1 Z_+} \right). \qquad (13a)$$

This shows that the impedance in the fundamental path has been increased due to non-linearity by the amount of the last term which may be denoted by ΔZ_p where

$$\Delta Z_p = \frac{a_2^2}{a_1^4} \frac{Z_+}{1 + a_1 Z_+} J_q^2.$$

A similar expression exists for ΔZ_q when J_p is substituted for J_q. The reciprocal of a_1 will be recognized as the internal resistance of the variable element for small potential variations.

From these

$$J_p^2 \Delta Z_p = J_q^2 \Delta Z_q,$$

so that the two fundamental circuits share equally in the power dissipation due to sideband flow. This means that when the two modulating currents are not of the same amplitude, the smaller current will have the larger resistance change due to sideband flow, and therefore will suffer a greater percentage amplitude change. This discussion serves to emphasize the point that the tube impedances depend upon the impedance-frequency characteristics of the circuit to which the tube is connected, so that this point must be kept in mind in the design and measurement of modulating circuits.

Grid Current Modulator, Large Alternating Grid Potentials

The comparatively simple analysis we have just employed is not capable of very wide application because of the assumed form of the grid current equation. In the practical forms of grid current modulators, from which comparatively large amounts of modulated power are required, the grid potentials are increased and the grid is maintained negative during an appreciable part of the cycle. The above method then becomes too involved to be extended to this case, since a large number of terms would be required for an accurate representation of the tube characteristic. When we have a large external grid resistance, however, as appeared to be desirable from eq. 10, a fairly exact solution for the modulation products can be obtained by another method which is capable of direct application.

If we determine the relation between impressed potential and output current in this particular case we find that on passing from negative to positive potentials, the plate current curve breaks sharply at about zero grid potential, and becomes nearly parallel to the x-axis, as shown in Fig. 4. We can therefore consider the positive lobe of the input wave to be cut off at zero grid potential under these conditions and the problem can be handled analytically.[6]

We are indebted to Mr. F. Mohr for computations on the sideband amplitude as given by eq. 20, of the Appendix, in which the sideband is expressed in terms of a multiple of P, as function of the ratio Q/P. The relationship between these quantities is given as a single-valued function. For our own purposes, however, we have plotted the

[6] Appendix 1.

sideband potential as a function of one of the modulating potentials with the other as parameter, as shown in the dotted lines of Fig. 5. The experimental data are plotted as the full lined curves and appear

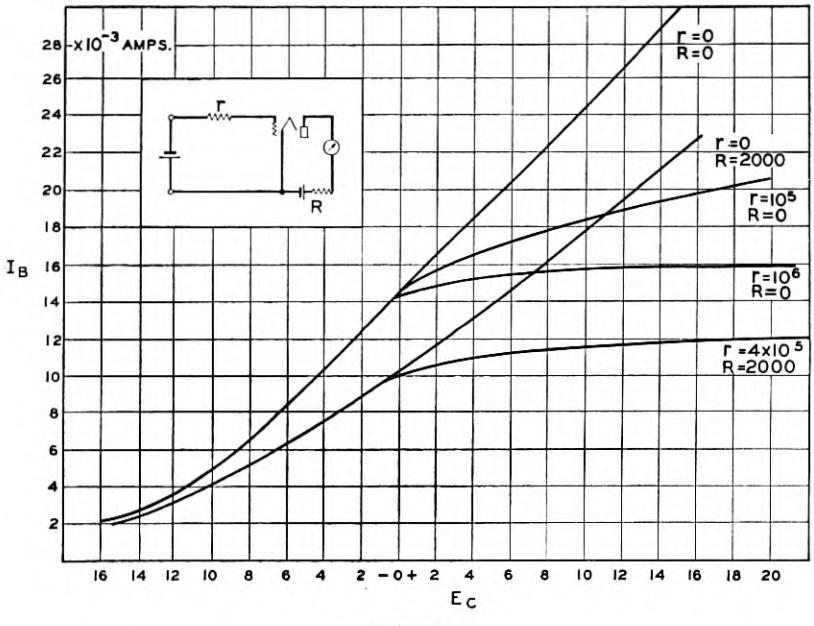


Fig. 4

to be in good agreement with the theory, the divergence being due presumably to the incomplete suppression of the positive lobe in the experimental set-up. The measurements were carried out with the current analyzer as described in connection with Fig. 3, grid current being measured as a function of the two input amplitudes. The sideband grid potential was then determined by multiplying the sideband current by the external grid resistance.

Possibly the most striking thing shown by Fig. 5 is that the sideband amplitude is independent of the larger of the two inputs, when the ratio of one input to the other is made sufficiently great. Hence we must provide sufficient carrier amplitude to insure that the resultant sideband shall be linearly proportional to the impressed signal up to its greatest value so that good quality of speech transmission may be assured.

Our earlier analysis using eq. 5 to represent the grid current characteristic led to a sideband amplitude proportional to the product of the two inputs whereas in this case in which the positive lobe is completely

suppressed, it is proportional to the smaller of the two when the ratio of the two inputs is greater than about 3/2. This proceeds from the fact that eq. 5 must be supplemented by many more terms involving higher powers of the impressed potentials, in order to represent the

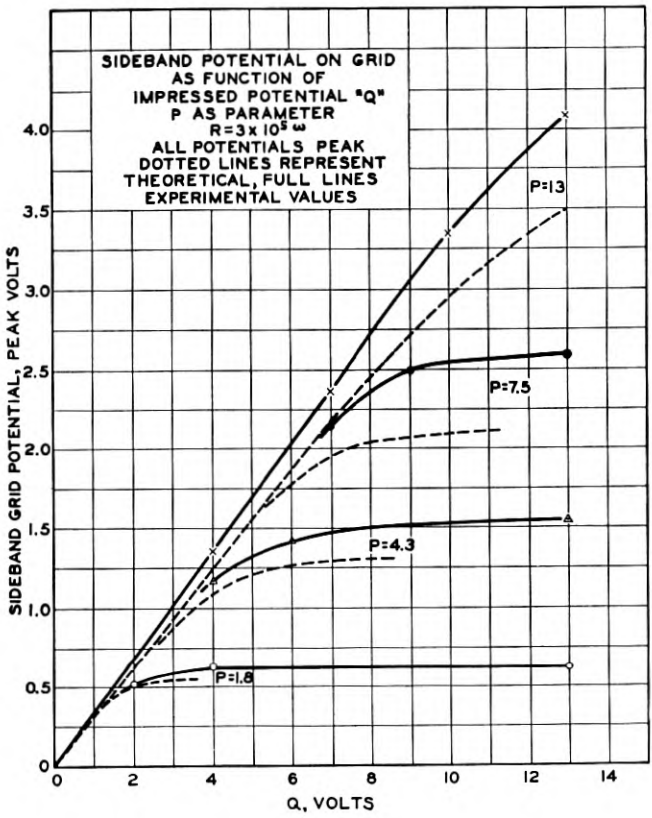


Fig. 5

grid current characteristic to any degree of precision when the grid is driven negative.

The form of the input-output curve is especially valuable for telephony. The relative independence of the larger of the two inputs means that the sideband output will be stable with regard to carrier current variations under the limitations noted. The output approaches a maximum asymptotically, so that the articulation at heavy loads may be expected to hold up better than in those modulating systems in which the output passes through a pronounced maximum. In a system transmitting the carrier, as in radio, and in which a square law

detector is used, the voice output is proportional to the product of the received carrier and sideband. Any change in attenuation, expressed in transmission units (T. U.), between the transmitting and receiving station affects the output current by twice that number of T. U. In the above grid current demodulator, however, the output changes to the extent of the attenuation change, and varies no more than in a carrier suppression system with the carrier locally supplied.

Having determined the grid voltage components, we may now apply the plate circuit coefficients to the grid potential in order to determine the plate current components, just as we did in the previous case. There, it will be recalled, we used a simple representation for the plate current in terms of the grid potential from which amplification and modulation terms were deduced. The same general considerations regarding phase opposition are carried over unchanged.

Limitation of Sideband Output

The above method of treatment is quite satisfactory when the space current is never reduced to zero, but when the grid voltage goes sufficiently negative, precisely the same limitations apply to the plate characteristic equation as applied to the grid equation under similar circumstances, and there exists an additional source of distortion in the plate circuit. In this circumstance the method of expansion in Bessel coefficients cannot readily be used because of the large number of components in the wave subjected to additional distortion, which would lead to prohibitive complication. We may nevertheless obtain a qualitative idea of the result in special cases of interest to us in this connection.

We shall assume that, as we found previously to be advisable, one of the two fundamental currents is substantially suppressed in the plate circuit, so that despite the non-linearity of the plate circuit sideband components are produced only in the grid circuit. We are therefore concerned with the variation of amplification with operating parameters. Now it is clear to start with, that at sufficiently small grid potentials, the entire variation falls within the region of variation of the plate dynamic characteristic so that the result may be written down as in the previous analysis. As the amplitude is increased, the negative end of the grid swing finally has no effect in varying the space current, and the distortion which results tends to limit the magnitude of the amplified components. Hence as the sideband potential on the grid is increased by increasing the applied signal potential and keeping the carrier potential large enough (say one and one half times the signal) so as to get the full efficiency of grid current modulation, the

sideband current in the plate circuit increases linearly with the signal up to a certain point. At this point, which corresponds to the plate current cut-off, the output departs from the linear relation and increases less rapidly. Further increase of the carrier produces no increase in output but a reduction of the output may result because of a greater swing beyond the cut-off point.

Inasmuch as the modulating potentials together with undesired modulated products form a wave having a net amplitude considerably greater than that of the useful sideband, it is clear that the maximum output amplitude can be increased by suppressing the undesired current components thus avoiding the loading and heating effects produced in large part by these other components. A method of attaining this desired result will be treated in connection with balanced circuits. The loading effect may be partially ameliorated very simply since one of the products of modulation is a d.c. component. The presence of series grid resistance means that we have in effect a negative bias applied to the grid which becomes increasingly negative as the input amplitude increases,—just the sort of thing, in other words, to limit sideband production. If, therefore, we use grid reactances instead of grid resistances we can achieve the same degree of modulating efficiency in the two cases at low inputs, and in addition remove effective grid bias, the maximum output power available being increased to a very considerable extent. Of course the insertion of grid reactance changes the details of the conclusions for the grid resistance, but the main features of performance are retained.

When grid resistance is used to provide a high impedance to the sideband, the operation of the grid leak and condenser detector is approached, in respect to the undesirable increase of bias with increase of input. As a consequence the output power is limited at large inputs, although the gain is fairly high at small input amplitudes.

Another point affecting the operation of the grid leak and condenser detector is the plate circuit impedance. According to the conclusions of the above theory for grid current modulation, the output power is increased at large input amplitudes by providing an impedance in the plate circuit which is high to both input frequencies and matches the tube impedance at all desired output frequencies. This conclusion has been verified experimentally at carrier frequencies when operating the tube for maximum output, but is contrary to the usual practice in radio circuits, where the plate circuit impedance to the modulating frequencies is ordinarily made low rather than high compared to the tube impedance. The problem is complicated at radio frequencies by regenerative effects not present to the same degree at the compara-

tively low frequencies used in carrier telephony, and by the comparatively low alternating and battery potentials which raise the relationship of plate and grid voltages to grid current, to importance.

We have now to examine the electrical properties of available circuit elements in the light of our previous analysis, so that their assembly will yield the most favorable results.

Vacuum Tubes

The effect of the shape of the grid-current—grid-voltage curve on the modulating properties of the grid circuit is not as pronounced at large amplitudes as might be expected from experience with plate current modulators at comparatively low amplitudes. As is well known this characteristic of ordinary tubes is much more variable between tubes of the same type than the plate-current—plate-voltage curve. But it has been found that a change of tubes having static grid characteristics varying within wide limits does not vary the modulating gain of a grid current modulator more than one T.U. The reason for this may be seen most easily in the case of an external grid impedance consisting of a pure resistance. If the tube grid resistance were comparatively small for all positive voltages the positive half of the wave would be completely suppressed, and the analysis of Appendix 1 would accurately represent the wave. Even when the tube grid resistance varies considerably it does not alter the wave shape appreciably so long as it remains small compared to the external resistance. This condition may be satisfied with particular ease for large input voltages, and may also be satisfied in a qualitative sense, when reactances are used in place of resistance. The principal effect of a change in grid resistance is then to change the input impedance, which affects the net gain only through the mismatch of impedance at input frequencies.

As a consequence of the tube circuits and range of operating potentials used in the grid current modulator, the details of the grid current characteristics become of relatively small importance and attention is focussed on the functioning of the plate circuit. The plate circuit is used purely for amplification purposes as mentioned above, so that the criteria of usefulness of a tube as a grid current modulator come down ordinarily under the stated operating conditions to the criteria of usefulness of a tube as an amplifier.

Filter and Transformer Networks

Input Filters and Modulating Gain

Since the gain obtainable in a grid current modulator depends primarily on the ratio of external to total grid circuit impedance, it is

necessary to consider how the required high impedance may be obtained in practice. The input transformer must have an impedance looking into the grid side which is high to all sideband frequencies, and must at the same time transmit efficiently all signal input frequencies. A high impedance over the sideband range is best obtained by a filter [7] on the low side of the input coil, care being taken to allow for the effect of the transformer on the filter impedance. In order to determine the actual external grid impedance and to investigate the modification of filter impedance by the input transformer, a high impedance bridge was built in which precautions were taken to prevent errors due to the high impedances involved (up to several megohms). In each case only the end section of the filter adjacent to the modulator was used, since this provided nearly the same impedance as would be given by a complete filter. Low pass filters are used on the input to the modulator and on the output of the demodulator, while band pass filters are used on the output of the modulator and the input of the demodulator. These are to be considered in turn.

Low Pass Filters

The simplest type of low pass filter is the infinity type of section, the impedance characteristics of which are shown in Fig. 6*a*. The filter alone, as shown by the solid lines, has negligible reactance in the transmission band (0–3 K. C.) and practically pure inductance in the attenuated region. The input transformer resonates in the attenuated region when terminated by this filter, as shown by the dotted lines, because of the leakage inductance and distributed capacity of the windings. The resonance peak is quite broad due to the comparatively high a.c. resistance and so covers a considerable frequency range as is shown by the ratio of $Z_-/(R_0 + Z_-)$ in Fig. 6*c*. It may be made to appear at higher frequencies by using a filter with a higher cutoff frequency, and at a lower frequency by replacing the series inductance by a parallel tuned circuit (an *m*-type section).[8] A new type of filter section developed for certain phases of this work and known as the built-out type [9] has a particularly good impedance characteristic in the attenuated region, as shown by the solid lines of Fig. 6*b*. But as shown by the dotted lines the transformer impedance, when terminated in this type of section, is very much modified by the coil constants. The resulting efficiency as show by the ratio $Z_-/(R_0 + Z_-)$ in Fig. 6*c* is not as good as that of the infinity type section.

[7] For a general discussion of filter impedances and attenuations see Campbell, *Bell System Technical Journal*, November, 1922; Zobel, January, 1923; Johnson and Shea, January, 1925.

[8] O. J. Zobel, *Bell System Technical Journal*, October, 1924.

[9] Devised by T. E. Shea of Bell Telephone Laboratories.

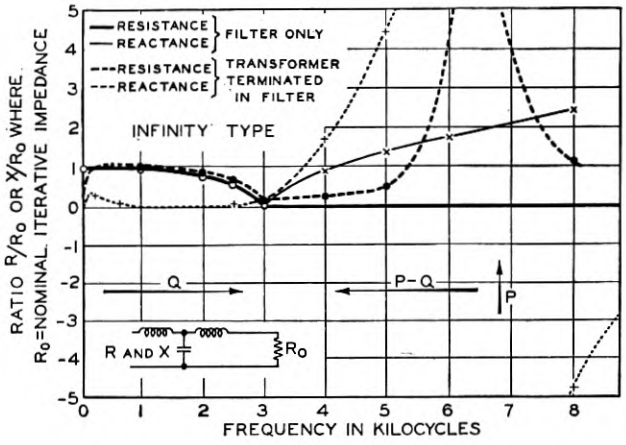


Fig. 6a

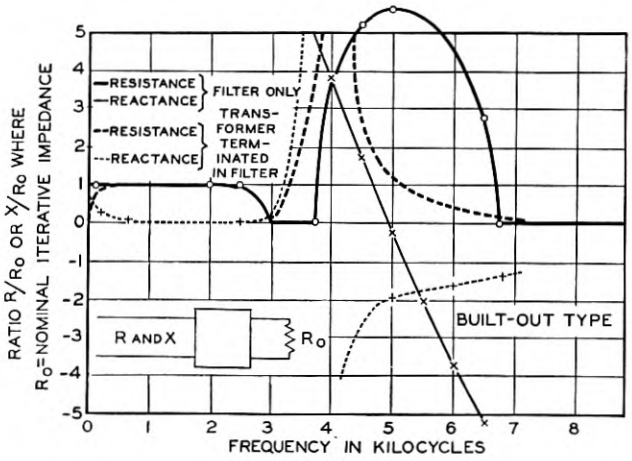


Fig. 6b

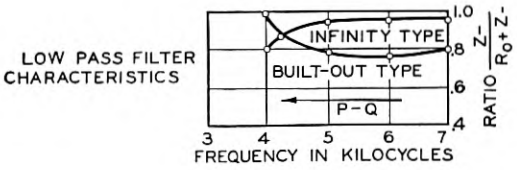


Fig. 6c

Band Pass Filters

The two most important types of band pass filter sections—the confluent and the built-out types—are shown in Fig. 7. The resonance of the confluent section in the voice frequency range does not affect the

ratio of external to total impedance appreciably but the chief defect is in the comparatively low value of this ratio at the higher voice frequencies (2.5 to 2.8 K. C.), although this type of section is satisfactory over the entire voice frequency range at higher carrier frequencies.

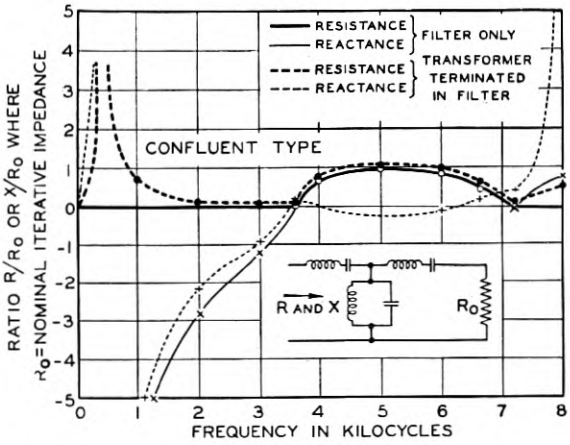


Fig. 7*a*

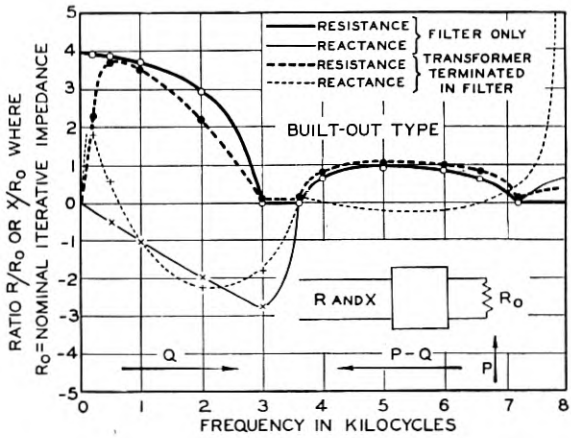
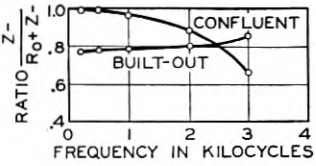


Fig. 7*b*



BAND PASS FILTER CHARACTERISTICS

Fig. 7*c*

The built-out band pass filter shown in Fig. 7*b* has a very satisfactory impedance over the voice range and the modifications introduced by the transformer do not seriously affect its efficiency. From the curves of Fig. 7*c* it is evident that the built-out type of section must be used for channels near the voice frequency range but that the confluent type shown in Fig. 7*a* may be used for the higher frequency channels.

The close relation between input filter impedance and modulating gain is illustrated in Fig. 8. Two band pass filters were built having

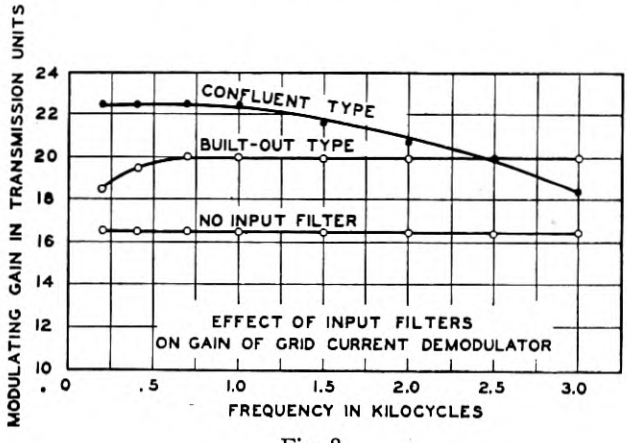


Fig. 8

impedance characteristics approximating the curves shown in Figs. 7*a* and 7*b*. It would then be expected that the modulating gain would be proportional to the ratio $Z_-/(R_0 + Z_-)$. The curves in Fig. 8 show that this is very nearly the case. With no input filter the ratio $Z_-/(R_0 + Z_-)$ would be 0.5 or 6 T.U. less than the maximum possible gain, as is found to be actually the case. This shows that the modulating gain may be calculated for any value of input impedance if it is known for any other value.

Input Impedance

The impedance looking into the low side of the input transformer when the high side is terminated in the grid circuit of a modulator under operating conditions (carrier at normal value) depends on a number of factors, among which the principal variables are the signal and carrier input currents, the input transformer, and the input generator impedance. As might be expected, the input impedance decreases as either carrier or signal amplitude is increased, and the change of impedance with signal amplitude is small when the signal is small compared to the carrier, as is normally the case.

The influence of the input transformer upon input impedance depends not only upon first order, but also upon higher order effects. The first order effect is simply due to the transformer terminated in a network having a linear current-voltage characteristic, which may be calculated from the usual transformer theory. The higher order effect is produced by the effect of the contributions to fundamental frequencies caused by the flow of modulation currents, as discussed in connection with Equations 13*a* and 4. For this reason the impedance of the external grid circuit at other than input frequencies may have a considerable effect on the input impedance. It has been found possible to reduce the reflection from a resistance line to a small value with suitable transformers.

Output Filters and Transformers

The general effect of an output filter or retard coil in the plate circuit with high impedance to all frequencies except the sideband is to increase the output level for large inputs, since the opposing effect of plate modulation is eliminated and the total load capacity of the tube is employed solely in the amplification of the sideband. The output transformer on account of its low ratio has very little effect in altering the impedance-frequency characteristic of the output filter so that we need not enter so thoroughly into the details as we did in the case of input filters.

Output Impedance

The output impedance of a grid current modulator (looking from the line into the output coil) is affected mostly by the transformer ratio and the impedance to carrier in the plate circuit. If the impedance to the carrier frequency is very high, as is usually the case, there will be very little modulation with the carrier in the plate circuit, and neither the carrier input current nor the external output impedance at signal frequencies affects the output impedance appreciably. The reflection may be made quite small over the frequency range without any great difficulty.

Gain-Frequency Characteristic

The problem of obtaining a flat frequency-gain characteristic over the voice range depends upon the attenuation of input and output transformers, the attenuation of filters, and the impedance characteristic of input and output coils when terminated by their respective filters. The transformer attenuation is comparatively small and affects the frequency characteristic mostly at frequencies below 200 cycles. The closer the carrier channels are spaced to each other or to the voice band, the more difficult it becomes to obtain filters with suf-

ficiently sharp cutoff. In most cases a maximum variation of 2 T.U. in the attenuation over the transmitted band is a reasonable figure for a band pass filter. Each transformer and filter tends to increase the attenuation at the edges of the transmitted band more than in the center so that frequencies from 800 to 2,000 cycles are always transmitted with minimum attenuation, which is independent of the frequency-output current characteristic of the modulating elements.

The above consideration of filter attenuation is substantially independent of filter impedance since the latter is determined mostly by the end section. From the previous consideration it is evident that either may have a very pronounced effect, so that in measuring the frequency characteristic of a modulator or demodulator both attenuation and impedance effects must be taken into account. The effects of input and output impedances can be partially separated when the carrier is suppressed because of the fact that the output impedance has but little effect at small inputs and the input impedance has but little effect at large input currents.

Balanced Tube Circuits

The present practice in carrier telephone systems is to suppress the carrier current and one sideband in order to conserve frequency space and to reduce the energy levels and the cross-talk in associated equipment. The elimination of undesired components of a wave may be carried out by two distinct processes,—frequency discrimination by filter networks, and phase discrimination or balance by bridge circuits.[10] Each method is useful and both find places in carrier systems. When the frequency separation between desired and undesired components becomes relatively small, frequency discrimination becomes impractical and expensive. The balance method is used to separate frequencies according to their respective phase relations in two or more similar modulating circuits, the phases of the output components depending on the relative phases of the input currents. Consequently only certain combinations of modulation products can be separated by balance and these only to an extent determined by the balance attainable in transformers and vacuum tubes, both of which are subject to manufacturing variations. Due to the proximity of the carrier and second order sideband frequencies the suppression of carrier current by filter circuits alone is impractical. Balanced circuits must be used for this purpose and in spite of unavoidable variations in tubes and circuits it is usually possible to reduce the carrier on the line to less than five per cent

[10] For an illustration of balanced circuits, reference may be made to U. S. Patent 1,343,306, issued to J. R. Carson.

of its normal unbalanced value. To separate one sideband from the other after the carrier has been suppressed, and to suppress unbalanced components other than the carrier, filter attenuation is customarily employed.

In the usual type of balanced circuit there are two possible input paths with corresponding output circuits, one connected to the two grids in series; and known as the series path; the other to the two grids in parallel, known as the shunt path or midbranch. When carrier is impressed on the midbranch and signal on the series arm—the present arrangement in commercial carrier systems using plate current modulators—we designate the circuit, as a matter of convenience, as the "Conjugate Input Type." The modulation product frequencies are distributed as shown in Fig. 9*a*. When both signal and carrier are impressed on the series branch the modulation product frequencies are

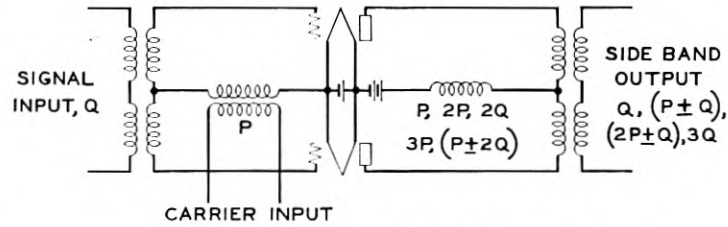


Fig. 9*a*. Conjugate Input Type

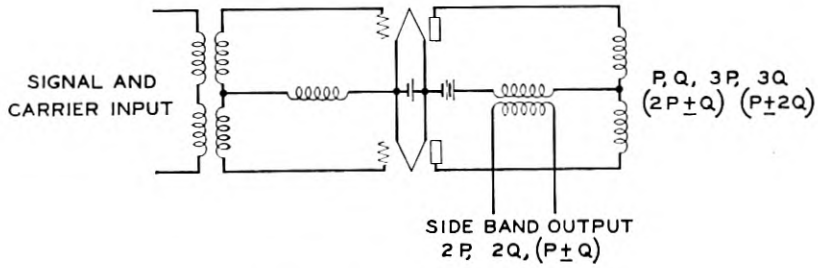


Fig. 9*b*. Common Input Type

as shown in Fig. 9*b* and the circuit is called for convenience the "Common Input Type." The phase of the modulation product of any order may be determined from the consideration that the phase of the frequency

$$\frac{1}{2\pi}|mp \pm nq|$$

depends upon the quantity $(m\theta_p \pm n\theta_q)$ where θ represents the phase

angle between current and voltage of each input frequency. This conclusion is independent of the type of modulation employed. If the phase of the product is then calculated to be identical on the two grids or two plates, it appears in the midbranch; if it turns out to be opposite in phase on the two grids referred to the filament, it appears in the series arm.

Conjugate Input Grid Modulator

With the signal introduced in the series arm of Fig. 9*a*, the sideband potential is built up across the same arm, so that a high sideband impedance must be provided by the input transformer terminated in its filter,—a low pass filter for the modulator, and a band pass filter for the demodulator. The carrier frequency is introduced in the conjugate arm so that the carrier circuit does not directly affect the signal and sideband impedances. There is a second order effect, however, due to the reaction of those modulation products which flow in the common branch. The input impedance may be expected to change also when the coupling between the two high impedance windings of the input transformer is varied, since this effectively changes the impedance to the above mentioned modulation products.

Modulation is largely eliminated in the plate circuit and the load capacity of the tubes is increased by inserting a choke coil in the common plate branch to suppress the carrier current. This, incidentally, tends to reduce carrier leak. The impedance of the choke coil at the carrier frequency is modified by the capacity to ground of the output transformer, and must be designed with this point in mind since the shunt arm impedance may otherwise be materially reduced. Some further increase in load capacity is obtained by having the output transformer and the terminating filter offer a high impedance to frequencies outside the transmitted band. In this way all important components except the sideband are suppressed and the plate circuit of the tube operates as an amplifier so that the plate power dissipation is reduced and the load capacity increased. The same considerations regarding the second order impedance effects of the shunt branch on the series branch exist for the plate circuit as for the grid circuit considered above. The main effect when there is loose coupling between the high impedance windings of the output transformer is to introduce an inductive reactance into the series arm. This tends to increase the reflection coefficient so that it becomes preferable to couple the two windings closely, a comparatively easy thing to do in low impedance circuits. The modulation products accompanying the desired product are indicated in Fig. 9, and it is seen that there will be no introduced distortion up to the third order when the carrier frequency is sufficiently high.

Common Input Grid Modulator

Another useful type of grid current modulator is shown in Fig. 9*b*, in which both signal and carrier are applied across the same input terminals. The modulation currents flow in the plate (and corresponding grid) circuits as shown in the above schematic. Where the ratio of carrier to signal frequency is large so that a single input transformer cannot be used efficiently, separate transformers with associated filter networks may be used for each of the two inputs. Since the second order sidebands ($p \pm q$) appear in the midbranches, it is not necessary to have the impedance high to these frequencies in the input coil, but only from the midpoint of the input coil to ground. This is most conveniently accomplished by a high inductance retard coil in the midbranch of the grid circuit, although transformers and high impedance networks may be used in general. The grid circuit sideband across the midbranch is amplified and appears in the plate circuit midbranch. The fundamental currents together with all odd order modulation products are eliminated by a high impedance, high mutual retard coil in the series arm of the plate circuit.

Since the present practice is to use suppressed carrier, a hybrid [11] coil must be used to introduce the carrier if this circuit is to be used as a demodulator, although the signal and carrier currents may be introduced through filters when used as a modulator. Either frequency discrimination or balance is required in any case to keep carrier current out of the signal circuit.

The chief advantage of the common over the conjugate input type of circuit is that the high impedance required for the modulated product is provided by a distinct element, and no high impedance requirements are placed on other elements in either input or output circuits. Another advantage of this arrangement is that the amplified fundamentals are balanced out, making the singing gain about 20 T.U. less than that of the conjugate input type. The only modulation products (up to the fourth order) not balanced out of the output are the second harmonics of carrier and signal. This type of circuit may be used as a demodulator at any frequency, but as a modulator only when the second harmonic of the highest voice frequency does not come in the sideband range—it is therefore not well adapted to modulate low carrier frequencies where high quality is required.

Although the output of this modulator is affected but little by the filter impedance in either input or output circuits, some care is neces-

[11] By using a hybrid coil having eight times as many turns in the signal circuit as in the carrier circuit, the equivalent current losses to signal and carrier are 0.5 T.U. and 9.5 T.U. respectively instead of 3 T.U. each, as is the case for the usual equality ratio hybrid coil.

sary in selecting the retard coils for the grid and plate circuits. Since the grid retard should have a high impedance to the desired modulation frequencies it must have an inductance of the order of 50 henries or greater at low frequencies in a demodulator. Resonance in the voice band is not harmful so long as the impedance does not drop too much at high voice frequencies.

The plate circuit retard coil is well balanced to reduce the unbalanced carrier transmitted to the line. An important requirement is that of close coupling so that the reactance in the output circuit may not be great enough to cause a transmission loss or large reflection coefficient. The required inductance then depends upon the relative separation of voice and sideband frequencies. If the lowest sideband frequency is very close to the highest voice frequency it may be impossible to prevent positive reactance from coming into the voice circuit of a demodulator, but the effect may be considerably reduced by utilizing this positive reactance in the mid-series section of the adjacent low pass filter.

Double Balanced Circuits

If two balanced circuits of either of the above types are connected with their input and output terminals respectively in series, all the modulation products up to the fourth order except the second order sidebands may be balanced out. There is no hybrid or filter loss and due to more complete suppression of unwanted frequencies the maximum output power obtainable is more than twice that with a single balanced circuit. The complexity of the resultant circuit is such as to rule it out for all ordinary applications.

For purposes of comparison we proceed to consider the experimental results obtained on a conjugate input grid modulator designed in accordance with the ideas set forth above.

Experimental Results

Figs. 10 and 11 represent the results of experiment on a conjugate input grid modulator with a carrier frequency of 6,800 cycles and a signal frequency of 1,000 cycles. The input and output networks previously discussed and represented in Figs. 6 and 7 were used here with 101-D tubes operated at 120 volts plate potential and 1.0 ampere filament current. The grids were connected to the negative terminal of the filaments. Fig. 10 represents the sideband output current in a 675 ohm circuit, plotted as a function of the signal current measured in the 675 ohm input circuit, with the carrier input maintained at 15 mils throughout. The upper four curves represent various experimental conditions designed to bring out the effect of different circuit

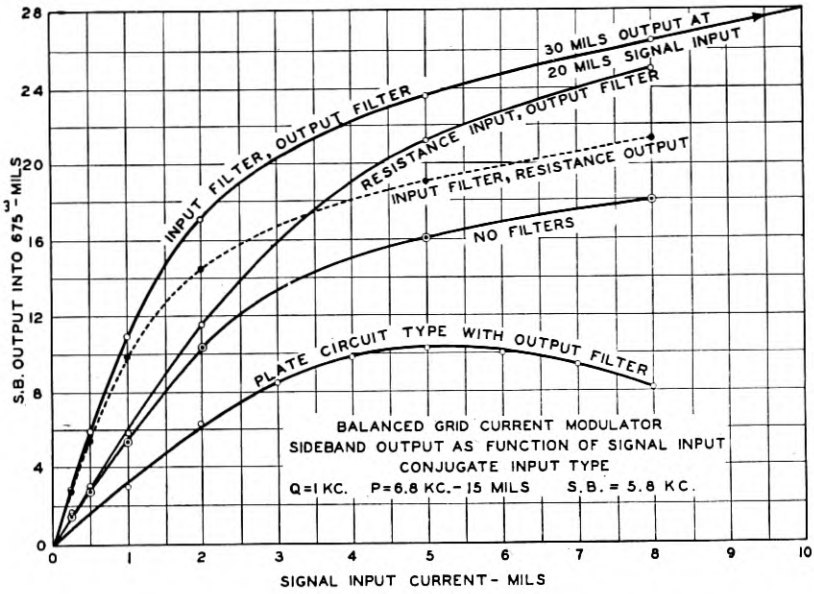

30 MILS OUTPUT AT
20 MILS SIGNAL INPUT
INPUT FILTER, OUTPUT FILTER
RESISTANCE INPUT, OUTPUT FILTER
INPUT FILTER, RESISTANCE OUTPUT
NO FILTERS
PLATE CIRCUIT TYPE WITH OUTPUT FILTER
BALANCED GRID CURRENT MODULATOR
SIDEBAND OUTPUT AS FUNCTION OF SIGNAL INPUT
CONJUGATE INPUT TYPE
Q=1 KC. P=6.8 KC.-15 MILS S.B. = 5.8 KC.
S.B. OUTPUT INTO 675 Ω - MILS
SIGNAL INPUT CURRENT - MILS
0
4
8
12
16
20
24
28
1
2
3
5
6
7
9
10

Fig. 10

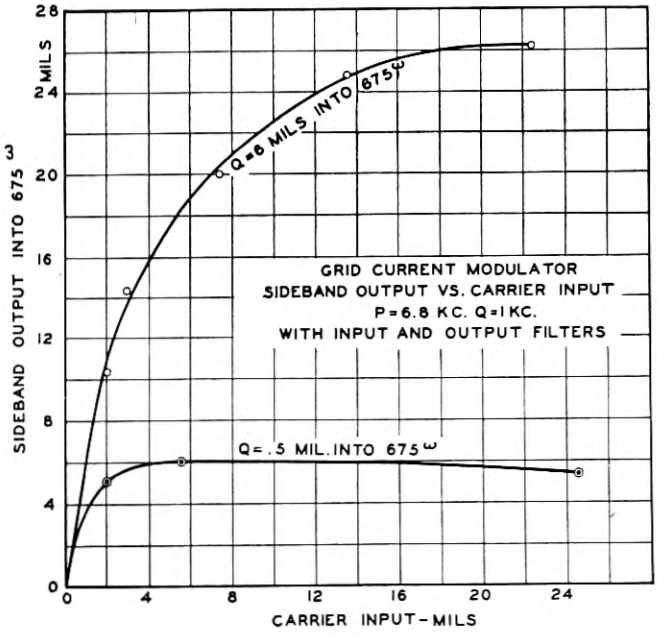

Q=6 MILS INTO 675 Ω
Q=.5 MIL. INTO 675 Ω
GRID CURRENT MODULATOR
SIDEBAND OUTPUT VS. CARRIER INPUT
P=6.8 KC. Q=1 KC.
WITH INPUT AND OUTPUT FILTERS
SIDEBAND OUTPUT INTO 675 Ω - MILS
CARRIER INPUT - MILS
0
4
8
12
16
20
24
28

Fig. 11

elements, while the lowest curve illustrates the performance of a representative conjugate input plate modulator working under conditions prescribed for it into a 600 ohm circuit with the same tubes and plate potential. A direct comparison between the two types of modulator as to sideband current output should include a comparative increase of 0.6 T.U. to the grid current modulator output to take care of the difference in the two load impedances.

The curve labelled "no filters" applies to the circuit of Fig. 9 in which both input and output circuits were connected to 675 ohm resistances. The presence of the retard coil in the plate circuit is accountable for the increase in output at large signal inputs over that of the plate type. When an output filter is added (resistance input, output filter) the gain at low inputs is scarcely affected but the output power for large signals is doubled since the load capacity is increased by the suppression of the signal frequency current in the plate circuit. If now an input filter is inserted and the output connected to a 675 ohm circuit (input filter, resistance output) the gain at low signal currents is increased by about 5 T.U. over that with no filters in circuit, while the increase at high signal amplitudes is of the order of 1.5 T.U. The topmost curve represents the performance of the modulator circuit terminated in the two filters, which shows a modulating gain of 21.5 T.U. at small inputs and a maximum power output of 30 mils into 675 ohms (0.6 watt). Fig. 11 represents the effect of varying the carrier input at two signal inputs—0.5 and 6 mils respectively. This illustrates the lack of dependence of sideband on carrier when the carrier is greater than the signal, which was deduced from eq. 20 as characteristic of this type of modulator. The use of a 15 mil carrier is seen to furnish close to the optimum value for the circuit, at least when the signal amplitude does not greatly exceed 6 mils.

The common input type is capable of yielding much the same results as the conjugate input type with somewhat less care required for the flanking filter impedances, since the proper circuit impedances are obtained by the use of retard coils as shown in Fig. 9*b*. On theoretical grounds, however, as we mentioned in discussing the general properties of balanced circuits, it is not capable of furnishing as high quality as the conjugate type at the low sideband frequency used here. At high sideband frequencies this objection disappears, so that the reduced filter requirements make it perhaps more attractive in application than the conjugate type. It should be noted that the plate modulator may be made to have a greater gain than that shown in Fig. 10 by changing the input transformer (with the same maximum output level) but this restricts the signal input current to correspondingly smaller amplitudes.

A few words on the shape of the signal-sideband curve of plate modulators of the van der Bijl type may not be inappropriate at this point since the curve depends to some extent upon the incidental grid modulation produced. Thus at large signal amplitudes the grid of the modulator tube is driven positive and grid modulation is produced, which tends to oppose plate modulation. By reversing the conditions which we have employed in the grid current modulator to promote grid modulation, the net plate modulation may be increased and the sideband-signal curve may more nearly show an asymptotic maximum which is so desirable from the overloading standpoint. This condition is evidently secured with a flanking input filter having a low impedance to the sideband, or by having an input coil which, while not seriously affecting the transmission of signal frequencies, offers of itself a low impedance to the grid sideband. Thus in plate modulators the input coil would have a high winding capacity, and in plate demodulators it would have comparatively low mutual inductance between primary and secondary windings.

As an indication of the quality obtained with the grid current modulating process, comparative listening tests between carrier telephone systems employing plate and grid demodulators, respectively, conducted by R. W. Chesnut, indicate roughly a 10 T.U. greater load carrying capacity for the grid type over a wide range of input amplitudes at about the same quality in both cases. The carrier leak may be reduced to one half mil by a not very critical tube selection, which is quite satisfactory in general.

The last point remaining is the plate power efficiency, which we have defined as the ratio of the sideband power developed in the load resistance to the d.c. power supplied to the plate circuit under operating conditions—it is really the efficiency of power conversion. At maximum output it is three per cent for the standard plate modulator and fifteen per cent for the above grid modulator. The efficiencies obtained at maximum output for a number of different low power tubes used in the grid current modulator may be tabulated as follows:

Tube	Plate Potential	Sideband Power W_{AC}	Plate Efficiency W_{AC}/W_{DC}
230-D.....	60	0.022	11%
221-A.....	70	0.065	18
221-D.....	90	0.13	14
101-D.....	120	0.50	15
102-D.....	120	0.11	22

For design information and construction of the experimental models,

the authors are indebted to E. B. Payne and H. R. Kimball for filters, and to H. Whittle and A. G. Ganz for transformers and retard coils.

Appendix

Grid Current Modulator, Large Grid Potentials

Making use of the observation that the positive lobes of the input wave are effectively suppressed with a sufficiently large external grid resistance, we first define a function equal to zero when the independent variable is positive, and equal to the variable when the variable is negative. This is evidently a representation of the potential effective on the grid in terms of the applied potential. If we denote the grid potential by $-f(y)$ where y is the impressed potential, it may be expressed as a Fourier series

$$-f(y) = b_0/2 + \Sigma b_m \cos m\pi y/Y + a_m \sin m\pi y/Y, \tag{14}$$

in which the coefficients are determined by the usual relations

$$\left.\begin{aligned} b_m &= \frac{1}{Y}\int_0^Y y \cos\frac{m\pi y}{Y}dy = \frac{Y}{m^2\pi^2}(\cos m\pi - 1), \\ a_m &= \frac{1}{Y}\int_0^Y y \sin\frac{m\pi y}{Y}dy = (-1)^{m-1}\frac{Y}{m\pi}\cos m\pi, \\ b_0 &= \frac{1}{Y}\int_0^Y y dy = Y/2. \end{aligned}\right\} \tag{15}$$

In these equations y represents the generator e.m.f. and Y is its maximum value. If we put (15) in (14), we have

$$\begin{aligned} -f(y) = Y/4 - \frac{2Y}{\pi^2}\sum_{m=1}^{m=\infty}\frac{\cos(2m-1)\pi y/Y}{(2m-1)^2} \\ + \frac{Y}{\pi}\sum_{m=1}^{m=\infty}\frac{(-1)^{m+1}}{m}\sin\frac{m\pi y}{Y}, \end{aligned} \tag{16}$$

the last term of which may be summed to $y/2$ and (16) may be rewritten as

$$-f(y) = Y/4 + y/2 - \frac{2Y}{\pi^2}\sum_{m=1}^{m=\infty}\frac{\cos(2m-1)\pi y/Y}{(2m-1)^2}, \tag{17}$$

which represents the desired solution. It will be observed that the first two terms of the right member contribute only a d.c. term and the fundamentals, so that the other modulation products must come from the summation. This expression is a perfectly general one as far as the form of y is concerned. In the case of a sinusoidal grid e.m.f. it is possible by more customary methods to find the grid potential, but

with a complex grid potential in which we are primarily interested no simpler representation is known to the authors except where the two frequencies involved are harmonically related.

With the two frequency inputs considered, we have the grid potential

$$y = P \cos pt + Q \cos qt \tag{18}$$

and the summation term may be written

$$\frac{2(P+Q)}{\pi^2} \sum_{m=1}^{m=\infty} \frac{\cos\left[(2m-1)\pi(P \cos pt + Q \cos qt)/(P+Q)\right]}{(2m-1)^2}. \tag{19}$$

Upon expansion of (19) as the cosine of the sum of two angles we find the terms cos $(A \cos \theta)$ and sin $(A \cos \theta)$ which require evaluation before the solution can be put in significant form. This might conceivably be done by direct expansion; for example, the first of the expressions would then be

$$\cos (A \cos \theta) = 1 - \frac{(A \cos \theta)^2}{2!} + \frac{(A \cos \theta)^4}{4!} \cdots.$$

Putting the terms of this series in terms of multiple angles gives us an infinite series to be summed as the coefficient of each multiple angle term. This may be done in terms of Bessel coefficients by Jacobi's expansions [12] which are as follows

$$\cos (A \cos \theta) = \sum_{n=0}^{n=\infty} (-1)^n \epsilon_{2n} J_{2n}(A) \cos 2n\theta,$$

$$\sin (A \cos \theta) = \sum_{n=0}^{n=\infty} (-1)^n \epsilon_{2n+1} J_{2n+1}(A) \cos (2n+1)\theta,$$

in which $J_k(A)$ is a Bessel coefficient of the kth order and ϵ_k is Neumann's factor which is two for k not zero, and unity for k zero. Carrying out the expansion, the sideband amplitude comes out as

$$\frac{4(P+Q)}{\pi^2}\left[J_1\left(\frac{P\pi}{P+Q}\right) J_1\left(\frac{Q\pi}{P+Q}\right) + \frac{1}{3^2} J_1\left(\frac{3P\pi}{P+Q}\right) J_1\left(\frac{3Q\pi}{P+Q}\right) + \cdots\right], \tag{20}$$

which may be computed from tables to be found in Watson's book [13] previously cited. Analogous expressions exist for the other components.

[12] Watson, "Theory of Bessel Functions," p. 22.
[13] P. 666 et seq.

A High Efficiency Receiver for a Horn-Type Loud Speaker of Large Power Capacity

By E. C. WENTE and A. L. THURAS

SYNOPSIS: This paper describes a telephone receiver of the moving coil type which is particularly adaptable to the horn type of loud speaker and which represents a notable advance over similar devices at present available. Its design is such as to permit of a continuous electrical input of 30 watts as contrasted with the largest capacity heretofore available of about 5 watts. In addition, measurements show that the receiver has a conversion efficiency from electrical to sound energy varying between 10 and 50 per cent in the frequency range of 60 to 7,500 cycles. Throughout most of this range, its efficiency is 50 per cent or better. This contrasts with an average efficiency of about 1 per cent for other loud speakers either of the horn or cone type. Combining the 50 fold increase in efficiency with a 5 or 6 fold increase in power capacity, a single loud speaker unit of the type here described is capable of 250 to 300 times the sound output of anything heretofore available.

This device is in commercial use in connection with the Vitaphone and Movietone types of talking motion pictures. As commercially produced in quantities numbering several thousand, an average efficiency of the order of 30 per cent has been realized.

BEFORE the advent of radio-broadcasting, practically the only loud-speakers in commercial use were of the horn type. In recent years this type has been largely supplanted by others of more compact design. However, where appearance and size are not of prime importance, a loud speaker with a horn still has a large field of service, as, for instance, in public address equipment or in systems for reproducing speech and music in large auditoriums from wax or film records. For such uses, the greater directivity obtained by the use of a horn has in some respects definite advantages. In the design of the receiver about to be described we have had in view particularly the requirements for such services, where the following qualifications were deemed of the greatest importance: a good response-frequency characteristic up to at least 5,000 p.p.s., large power output without amplitude distortion, high efficiency, and constancy of performance.

As this paper is concerned with the design of a driving unit, or the receiver proper, and not with the horn, we shall confine our discussion to the operation of the receiver when connected to a tube of infinite length and of the same cross-sectional area as the throat of the horn. An ideal horn should have at its throat the same acoustic impedance [1]

[1] The term acoustic impedance as here used may be defined as the ratio of pressure to rate of volume displacement.

as a tube of this character, viz., $\frac{\rho c}{A}$ c.g.s. units, where ρ is the density of air, c, the velocity of sound, and A, the area.[2]

The Coupling Air Chamber and Diaphragm

In order effectively to make use of horns as sound intensifiers it is usually necessary to couple the throat of the horn to the diaphragm through an air chamber. We shall first consider the effect of this air chamber on the sound output of the loud speaker. This coupling air chamber is generally of an indefinite conical shape of the type shown in Fig. 1. If we assume that this air chamber is so proportioned that

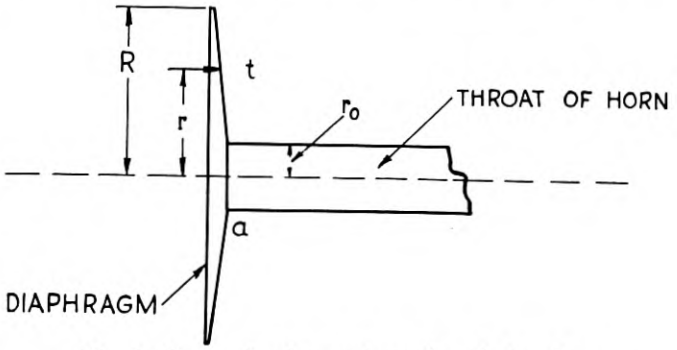


Fig. 1—Conventional type of coupling air chamber.

the annular area, $2\pi rt$, is equal to the throat area, and that the diaphragm is driven so that its displacement is paraboloidal, then, as calculated from formulæ developed in appendix A, the mechanical impedance imposed by the air chamber and horn on the diaphragm is shown in the curves of Fig. 2. Here the ordinates of the curves r_1 and x_1 are proportional to the resistance and reactance respectively, and the abscissæ are equal to the product of the frequency and the radius of the diaphragm. Of particular interest here is the large decrease in the resistance with frequency, for r_1, multiplied by the square of the velocity of the diaphragm, is the acoustic power delivered to the horn. For example, if the radius of the diaphragm were four centimeters, no sound would be emitted at 4,000 p.p.s. We have here one reason why most horn-type loud speakers fail to reproduce high frequency tones at sufficient intensity. Of course, in most cases the high frequency tones are further attenuated by the fact that the mode of motion of the diaphragm changes with frequency. The decrease in

[2] "The Function and Design of Horns," by C. R. Hanna and J. Slepian, *Journal of the A. I. E. E.*, March 1924.

resistance with frequency is largely due to the fact that the disturbances generated at different points of the diaphragm do not reach the throat of the horn in the same phase. To minimize this effect the air chamber should be designed so as to make this phase difference as small as

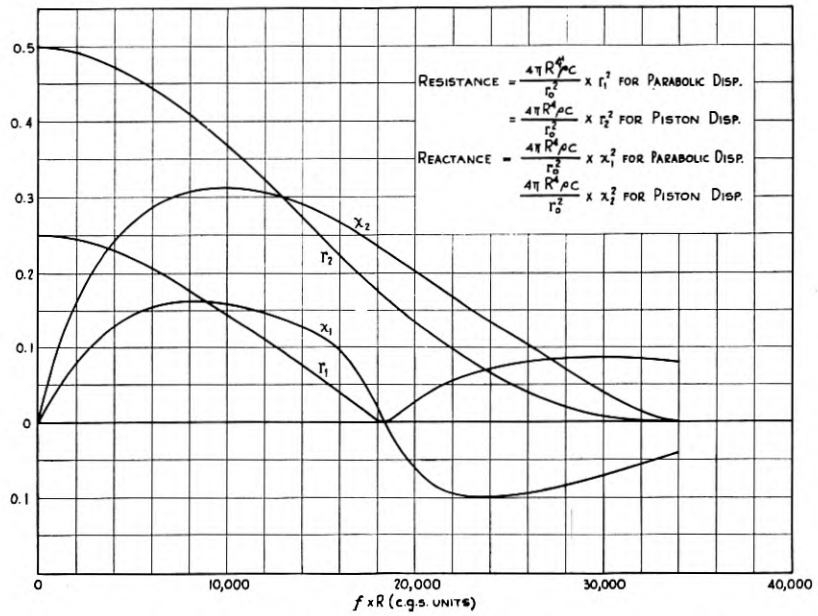


Fig. 2—Mechanical impedance of air chamber and ideal horn.

possible. In the same figure, r_2 and x_2 show the resistance and reactance respectively, if the diaphragm were moved as a plunger, i.e., with the same amplitude and phase over its whole surface. It is seen that the resistance is considerably larger and the cut-off frequency nearly twice as high. These curves show the superiority of the plunger type of diaphragm.

In order to cover the desired frequency range the method of coupling a diaphragm to the horn shown in Fig. 3 was adopted. Here the disturbances reach the horn more nearly in phase without having to pass through any restricted passages. The throat of the horn is flared annularly to the point A. The disturbances reach the throat of the horn from the inner and outer portions of the diaphragm approximately in phase up to comparatively high frequencies. With this type of construction it is possible to use a fairly large diaphragm so that large amounts of power may be delivered without a great sacrifice in efficiency at either the high or the low frequencies. An experimental test

showed that with this type of coupling for a particular size of diaphragm and throat area the cut-off frequency was raised from approximately 3,500 to 6,000 cycles per second.

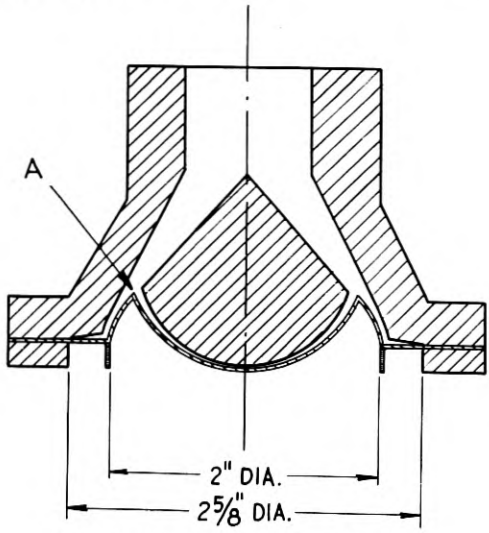


Fig. 3—Diaphragm and air chamber.

Principal Dimensions

Effective mass of coil and diaphragm = 1.0 gm.
Effective area of diaphragm = 28 sq. cm.
Area of throat of horn = 2.45 sq. cm.
Stiffness constant = $\frac{\text{force}}{\text{static displacement}} = 6 \times 10^6$ dynes/cm.
Resistance of coil = 15 ohms.
Length of wire in coil = 760 cm.
Average flux density = 20,000 gauss.

The diaphragm was made of a single piece of aluminum alloy 0.002 inch thick; metal was used in preference to other materials because of its superior mechanical properties. The form and principal dimensions are shown in Fig. 3. A driving coil is attached directly to the diaphragm near its outer edge. With this arrangement the diaphragm can be driven nearly as a plunger and it has little tendency to oscillate about a diametral axis, as there is great rigidity against a radial displacement of any part of the coil. The portion of the diaphragm lying between the coil and the clamping surfaces has tangential corrugations of the same type as described by Maxfield and Harrison [3] in reference to a phonograph sound box. The inner portion of the diaphragm was drawn into the form of two re-entrant segments

[3] *Bell System Technical Journal*, Vol. V, pp. 493–523, July 1926.

of spherical shells; this part was thereby made very rigid so that it should move as a unit up to high frequencies.

Construction of the Driving Coil

For the driving element of loud speakers either a moving coil or a moving armature is commonly used. The latter is in general satisfactory if driven at a small amplitude. However, where large powers are involved, the moving coil drive can be much more simply constructed so that it is free from amplitude distortion; it has the further advantage of having a resistance nearly constant with frequency and a practically negligible reactance. These were the primary reasons for our choosing this type of drive.

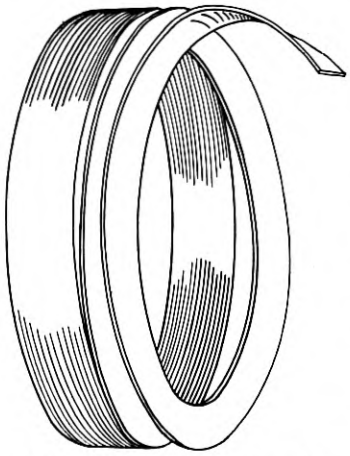


Fig. 4—Receiver driving coil.

The coil that was used in the receiver consisted of a single layer of aluminum ribbon 0.015 inch wide and 0.002 inch thick wound on edge as shown in Fig. 4. The turns were held together with a film of insulating lacquer about 0.0002 inch thick, thoroughly baked after the winding was completed. This type of coil has the following advantages. It is self-supporting, no spool being required; 90 per cent of the volume of the coil is occupied by metal; the distributed capacity between turns is small, giving a coil whose impedance varies only slightly with frequency; the metal is continuous between the cylindrical surfaces, allowing heat to be conducted rapidly outward from the center of the winding and diminishing the possibility of any warping of the coil; it can be accurately made to dimensions, thus permitting small clearances between the coil and the pole pieces. Small clearances not only permit the use of a comparatively small magnet but they facilitate the dissipation of heat. This latter effect is shown in the curves of Fig. 5. These curves give the temperature of the coil as a function of the power input for the coil in open air (*A*), and when it is placed between annular pole pieces with clearances of 0.010 inch between the cylindrical surfaces (*B*).

The Electromagnet

As shown in Fig. 6, the electromagnet is of conventional design except that the central pole piece has an opening through its center to

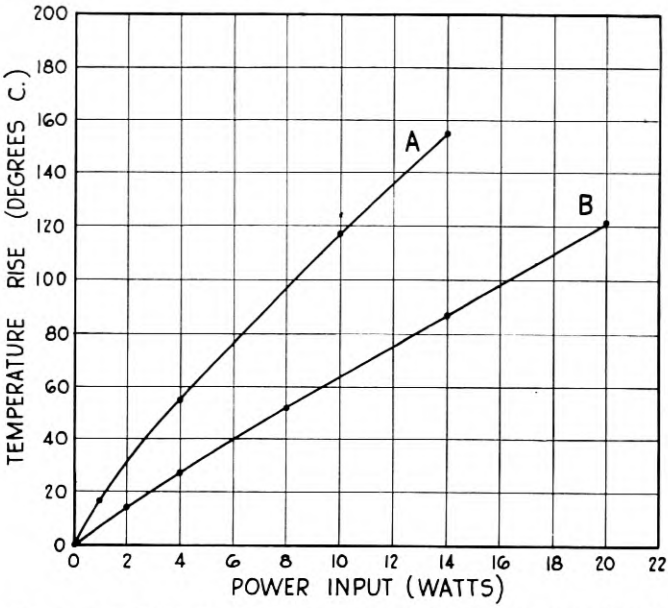


Fig. 5—Variation of temperature of coil with direct current input.

avoid a reaction of any air pockets on the diaphragm. This opening is widely flared to prevent tube resonance.

Experimental Results

It has already been pointed out that an ideal horn should have an acoustic impedance at its throat equal to that of a tube of infinite length and of the same cross-sectional area. In order to measure the

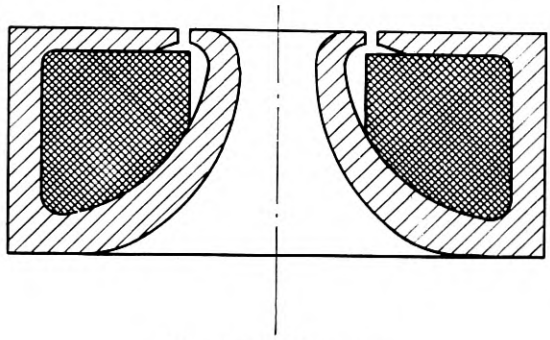


Fig. 6—Field magnet.

efficiency that the receiver would have under this condition, by any method whatever, it is necessary to connect the receiver to a tube having the same acoustic impedance as a tube of this character. This impedance for a tube 2.45 sq. cm. in area is 16.7 c.g.s. units. A tube of finite length but of the same area will have an impedance of this value provided the sound wave reflected at the far end has a relatively small amplitude when it reaches the sending end. To satisfy this condition a tube 50 feet long was terminated in an acoustic resistance unit having an impedance approximately equal to $16.7 + j16\omega \cdot 10^{-4}$ c.g.s. units. The essential elements of this resistance unit comprised a number of short narrow annular slits; its impedance was determined experimentally by a method described in another paper.[4] As this impedance at low frequencies is practically the same as the characteristic impedance of the tube, the amplitude of the reflected wave in this region is small; at the higher frequencies the reflected wave is attenuated sufficiently in the 50-foot tube to produce a negligible effect on the sending end impedance. This tube with the resistance unit was connected to the receiver during the following series of measurements.

Efficiency

One of the simplest methods of determining the power efficiency of a loud speaker is to measure the electrical impedance, first, when the receiver is in operating condition, and, secondly, when the diaphragm is constrained from moving so that no back e.m.f. is generated. The difference between these impedances is known as the motional impedance.[5] The resistance component of this motional impedance when multiplied by the square of the current gives the power that is generated by the motion of the diaphragm. If there is a negligible amount of power lost in viscosity and mechanical hysteresis, the ratio of the motional impedance to the free impedance can be taken as the efficiency of the receiver, i.e., the ratio of the acoustic power output to the total power input. This method of measuring efficiency is well known to the art, but for most commercial receivers the efficiency is so low that the motional impedance cannot be determined with a high degree of accuracy over an extended frequency range. However, for this receiver we have had no difficulty in determining the efficiency in this way up to 8,000 p.p.s. The values so obtained are given by the circles in Fig. 7.

On account of the uncertainty of the magnitude of the mechanical

[4] Wente and Bedell, *Bell System Technical Journal*, January 1928.

[5] Kennelley and Pierce, "The Impedance of Telephone Receivers as Affected by the Motion of their Diaphragms," *Proc. A. A. A. S.*, Vol. 48, No. 6, September 1912.

power losses within the receiver it was deemed desirable to measure the efficiency more directly, viz., to measure the actual sound power generated for a given power input.

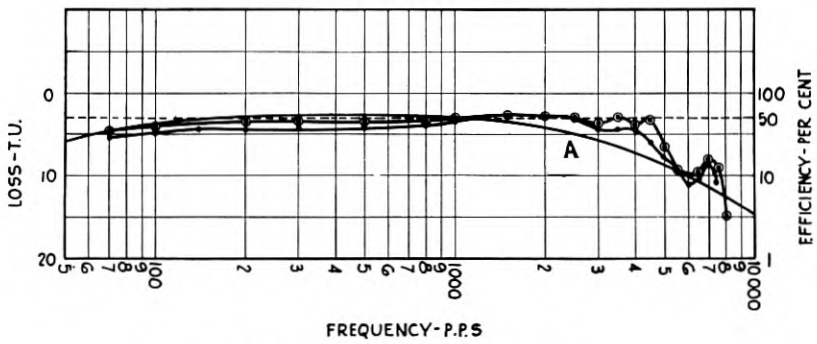


Fig. 7—Efficiency of the receiver.

The power output may be determined directly by measuring the acoustic pressure in the tube at the sending end. In order to measure this pressure an annular slit was provided on the side of the tube a few inches from the receiver as shown in Fig. 8. This annular slit had a

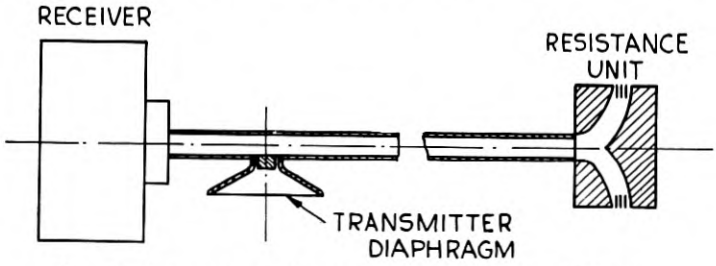


Fig. 8—Arrangement of apparatus for measuring efficiency.

diameter of a quarter of an inch, a width of 0.003 inch and a length of 0.040 inch. A slit of these dimensions has an acoustic impedance about fifteen times as great as the tube, so that it has a negligible effect on the sound wave propagated along the tube. This slit led to a small chamber over the face of the diaphragm of a condenser transmitter. The condenser transmitter was connected to an amplifier and an a.c. ammeter. This combination was previously calibrated, so that from the meter readings the pressure over the slit could be determined.

The input power was determined from the current input and the resistance of the receiver. With this set-up we thus were able to measure both the acoustic power transmitted along the tube, which

is equal to $\frac{(\text{pressure})^2}{16.7}$ ergs per second, and the power input. The operating efficiency is the ratio of these two quantities. The dots plotted in Fig. 7 give the values of the efficiencies so obtained. These values are seen to agree closely with those calculated from the motional impedance. This agreement shows that the mechanical power losses in the receiver are small.

Curve A in Fig. 7 gives the efficiency as calculated from the constants of the receiver by means of the formula given in appendix B, under the assumption that the mechanical impedance imposed on the diaphragm and the air chamber has the same value throughout the whole frequency range, viz., $16.7\,A^2$ c.g.s. units, where A is the effective area of the diaphragm. It is seen that the calculated and measured values are in good agreement except for certain irregularities at the higher frequencies. Whether these irregularities are to be ascribed to the action of the air chamber or to a change in the mode of motion of the diaphragm we are not at present prepared to say.

The curves of Fig. 7 give an efficiency for this receiver of about 50 per cent over a wide frequency range. This efficiency is within 3 T.U. of the possible maximum of 100 per cent. We may remark at this point that it is conceivably possible to build a receiver which will sound louder than one having an efficiency of 100 per cent. If, for instance, a receiver introduces harmonics on account of amplitude distortion, a low frequency driving force may give rise to a tone of higher frequency, where the ear may have a sensitivity many times greater than at the driving frequency. An increase in loudness obtained in this way of course exacts a sacrifice in the faithfulness of reproduction. The difference in loudness between the sound emitted by this receiver and by ordinary commercial types of loud speakers, for the same power input, is considerable, since most of them have an efficiency of less than one per cent for speech frequencies. Not only does this receiver have a high efficiency over a wide frequency range but it is free from any sharp variations in efficiency with frequency, a condition of great importance in the quality of reproduction.

Amplitude Distortion

Thus far we have discussed only the frequency characteristic of the receiver. There still remains to be considered the proportion of harmonics that are generated by the receiver when supplied with a current of sine wave form. These harmonics are generated when the displacement of the diaphragm is not proportional to the input current. At low frequencies the amplitude of motion for a given power output

is comparatively large and the diaphragm for large powers will be driven beyond the point where Hookes' law holds. At the higher frequencies no trouble is to be expected from this source. With the aid of an electrical filter we have therefore made measurements on the harmonic content in the sound output when the receiver was supplied with a sixty-cycle sine wave current. The values so obtained as a function of the power input are plotted in Fig. 9, where curve *A* is the

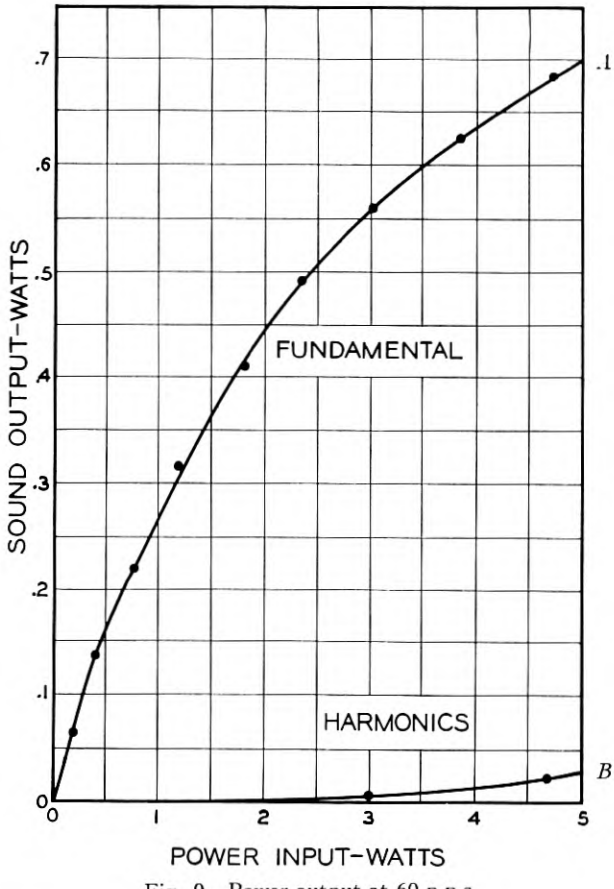


Fig. 9—Power output at 60 p.p.s.

output power of the fundamental tone and *B* that of the higher harmonics. These curves show that even at 60 cycles an output power of 0.5 watt may be obtained without the introduction of higher harmonics to an amount greater than 1.0 per cent. The total power in the harmonics would in this case be 20 T.U. below that in the fundamental

tone. If a horn were connected to the receiver in place of the tube, in addition to the resistance, a mass reactance would generally be imposed on the diaphragm at the lower frequencies. Under these conditions the proportion of harmonics introduced would be still lower than that indicated in Fig. 9.

At the higher frequencies the power output is limited solely by the current-carrying capacity of the coil. At these frequencies the steady power input for a temperature rise of 100 degrees C. is about 30 watts. With an efficiency of 50 per cent the corresponding output would be 15 watts.

After the work described in this paper was for the most part done and as a result of the extremely promising performance of the first models, a design of the receiver built along essentially these lines was worked into a form suitable for commercial production by Mr. W. C. Jones and Mr. L. W. Giles. These receivers are now in commercial use in Vitaphone and Movietone installations. As commercially produced in quantities numbering several thousand, efficiencies of the order of 30 per cent have been realized.

In conclusion, we wish to express our appreciation for the valuable assistance given by Mr. T. F. Osmer in carrying out most of the experimental work described in this paper.

Appendix A

Consider a diaphragm and connecting air chamber of the form shown in Fig. 1. Assume that the air chamber is of a form such that the cross-sectional area at any distance r from the center is equal to the throat area of the horn, i.e., $2\pi rt = \pi r_0^2$. This form of connecting air chamber then differs but little from that used in most commercial types of horn speakers. The sound output is in general dependent on the mode of motion of the diaphragm. In most loud speakers this mode of motion varies with the frequency. However, let us assume that we have a paraboloidal displacement at all frequencies. The velocity at any radial distance may then be represented by

$$\xi = \xi_0 \left[1 - \left(\frac{r}{R}\right)^2\right] e^{i\omega t}$$

if $\xi_0 e^{i\omega t}$ is the velocity at the center.

Under the assumed conditions, the sound transmitted through the throat is very nearly the same as that which would be transmitted along the positive direction through the tube sketched in Fig. 10, which extends to infinity in both directions, provided the portion of the wall

of the tube from a' to 0 and from a to 0 had a radial velocity equal to

$$\frac{2\pi r}{2\pi r_0} \cdot \dot{\xi}_0 \left[1 - \frac{r^2}{R^2} \right] e^{i\omega t}.$$

The velocity potential at a point, P, at a distance y from a, if r_0 is small compared with the wave-length of sound, is then

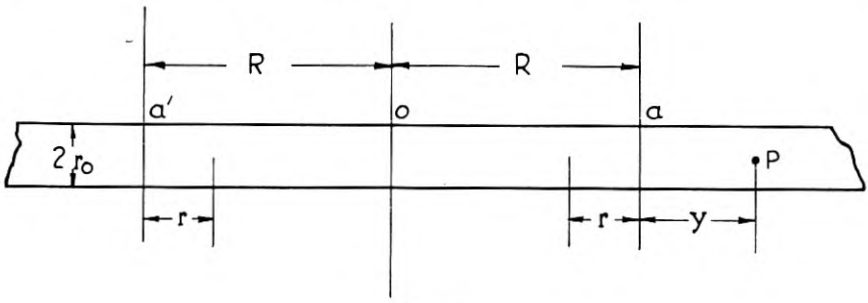


Fig. 10.

$$\varphi_y = -i \left[\int_0^R \frac{\left[1 - \frac{r^2}{R^2} \right] r}{r_0^2 k} e^{ik(ct-y-r)} dr + \int_0^R \frac{\left[1 - \frac{r^2}{R^2} \right] r}{r_0^2 k} e^{ik(ct-y-2R+r)} dr \right] \dot{\xi}_0,$$

where c is the velocity of sound and

$$k = \frac{\omega}{c}$$

or

$$\varphi_y = A \frac{2\dot{\xi}_0 e^{ik(ct-y-R)}}{-ir_0^2 k^3},$$

where

$$A \equiv \left(1 + \frac{6}{k^2 R^2} \right) \cos kR + \left(2 - \frac{6}{k^2 R^2} \right).$$

If we take the real part of

$$\rho \frac{d\varphi_y}{dt},$$

we get the instantaneous pressure at the point, P, where ρ is the density of the air. This gives

$$p_y = \frac{-2\dot{\xi}_0 \rho c}{r_0^2 k^2} \cdot A \cos k(ct - y - R).$$

If we substitute r for $-y$, this expression gives the instantaneous pressure on the diaphragm at the radial distance r from the center. The instantaneous power delivered by the diaphragm is then

$$W = -\frac{2\rho cA}{r_0^2 k^2}\dot{\xi}_0 \cos kct \cdot 2\pi \int_0^R \left(1 - \frac{r^2}{R^2}\right) \cos k(ct + r - R) \cdot r dr$$

$$= \frac{4\pi A \rho c}{r_0^2 k^4}\dot{\xi}_0^2 [A \cos kct + B \sin kct] \cos kct,$$

where

$$B \equiv \left(1 + \frac{6}{k^2 R^2}\right) \sin kR - \frac{6}{kR}.$$

The effective force on the diaphragm is then

$$F = \frac{W}{\dot{\xi}_0 \cos kct} = \frac{4\pi A \rho c \dot{\xi}_0}{r_0^2 k^4} [A \cos kct + B \sin kct].$$

The effective resistance is therefore

$$\frac{4\pi A^2 \rho c}{r_0^2 k^4}$$

and the effective reactance

$$-\frac{4\pi \rho c A B}{r_0^2 k^4}.$$

Expressions for the resistance and reactance for other modes of motion of the diaphragm may be obtained in a similar manner.

Appendix B

The motional resistance, R_m, of a moving coil receiver is equal to

$$\frac{B^2 l^2 (r + r')}{(r + r')^2 + \left(m\omega - \frac{S}{\omega} + x\right)^2}$$ ohms,[6]

where B is the average flux density,

l, the length of wire in the receiving coil,

$r + jx$, the mechanical impedance imposed on the diaphragm by the horn through the coupling air chamber,

[6] Kennelley and Pierce, loc. cit.

$r' + j\left(m\omega - \frac{S}{\omega}\right)$, the mechanical impedance of the diaphragm aside from that imposed by the horn.

If r' is negligible, the efficiency of the receiver, i.e., the ratio of power output to power input, is

$$\eta = \frac{R_m}{R_m + R_d},$$

where R_d is the resistance of the coil when its motion in the field is completely damped.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

A Photo-electric Process of Halftone Negative Making Applicable over Telephone Lines.[1] H. E. IVES. The commercial process of making halftone engravings breaks the picture up into a large number of dots by means of a screen. Unless special intermediate processes are adopted this does not reproduce the tones correctly because the size of the dots is not directly proportional to the light intensity. This paper describes an adaptation of the photo-electric system of picture transmission, as an alternative for use of screen, which produces individual dots more accurately proportional in size to the light intensity.

The method proposed thus affords a means of improving quality in halftone engraving by using an outfit similar to the picture transmitting and receiving outfit. Used in connection with the commercial picture transmission service, it can provide pictures with an accurate tone structure during the transmission process so that the resulting copy is ready for the engraver without the use of any screen process. Full details of several arrangements of the apparatus are given together with engravings produced by the photo-electric method.

The picture transmission system as used transmits the picture in the form of a continuous strip of varying intensity. By the introduction of a synchronized sectored disc this strip is made discontinuous, forming the dots for the halftone process.

Advance Planning of the Telephone Toll Plant.[2] J. N. CHAMBERLAIN. A general review of the commercial studies which precede the design of telephone toll plant is given together with some specific data concerning the toll line conditions in the northern California area of the Pacific Telephone and Telegraph Company. Attention is called to the special conditions requiring a large amount of submarine cable plant resulting from the peninsular location of San Francisco. The Pacific Company has plans for about 1,000 miles of toll cable network in its present program which it expects to install at the rate of 100 miles a year. The article places special emphasis on the commercial factors governing the choice between cable and open wire construction for future extensions of the toll plant.

[1] *Opt. Soc. Amer. Jl.*, Vol. 15, p. 96, August, 1927.

[2] *Jl. Am. Inst. El. Engrs.*, Vol. 46, p. 994, October, 1927.

The Adsorption of Gases by Solids with Special Reference to the Adsorption of Carbon Dioxide.[3] H. H. LOWRY and P. S. OLMSTEAD. This paper presents a theory of adsorption and its mathematical development, which is similar to but differs somewhat from that of Polanyi. A detailed description is included of a relatively convenient method of application of the theory to experimental data. Using this procedure, a test of the theory has been made on the data obtained by Homfray, Titoff, Richardson, Chappuis and S. O. Morgan for the adsorption of carbon dioxide by charcoal. The very satisfactory agreement obtained between experiment and theory gives support to the fundamental assumptions underlying the theory.

The Densities of Coexisting Liquid and Gaseous Carbon Dioxide and the Solubility of Water in Liquid Carbon Dioxide.[4] H. H. LOWRY and W. R. ERICKSON. The densities of coexistent liquid and gaseous carbon dioxide are measured over the temperature range − 5.8 to 22.9° and it is shown that they can be satisfactorily represented by equations involving the first and one third powers of the temperature on the critical scale. The data are shown to be in substantial agreement with those of other observers. It is also shown that the density of saturated carbon dioxide vapor is the same within the experimental error in the presence or absence of water, from which it is concluded that the solubility of water in liquid carbon dioxide is less than about 0.005 per cent by weight over the temperature range of the investigation. Attention is called to qualitative evidence of the formation of a solid hydrate of carbon dioxide at about 4°.

Atomic Grouping in Permalloy.[5] L. W. MCKEEHAN. This is a theoretical paper which follows a long series of experimental papers. In a solid solution of two metals, e.g., in the solid solution of nickel and iron known as permalloy, the atoms of both kinds occupy in each crystal the points of a single space-lattice. It is important to know whether the points so occupied by atoms of a single kind are located at random or have some regularity of arrangement. If the latter is the case, it may be asked further whether the regularity is due to the frequent occurrence of definite groupings of unlike atoms or merely to a tendency for atoms of one kind to separate from each other as widely as possible. The magnetic properties of permalloy are here taken to show that the last-mentioned possibility is probably nearest

[3] *Journal of Physical Chemistry*, Vol. 31, pp. 1601–1626, November, 1927.
[4] *Journal of the American Chemical Society*, Vol. 49, pp. 2729–2734, November, 1927.
[5] *Jl. Franklin Inst.*, 204, 501–524 (1927).

to the truth. The method is a combination of analytical and graphical analysis applied to several hypothetical solid solution crystals each containing more than a thousand atoms. It may also, as is pointed out in the paper, be used in problems concerning other than magnetic properties of solid solutions.

The Short Wave Limit of Vacuum Tube Oscillators.[6] C. R. ENGLUND. A study of the shortest attainable undamped waves which can be produced with vacuum tube oscillators resulted in the production of one-meter waves as the fundamental mode of oscillation of the circuit. The shortest waves "attainable with reasonable ease" with several types of standard tube are found to be:

Tube	Meters
W. E. 230-D (60 mil. fil.)	2.0
" 205-D ("E" 5-watt)	3.2
" 221-D (1/4 amp. fil.)	3.3
" 211-D ("G" 5-watt)	3.5

In order to reduce the capacity as far as possible some experiments were made with unbased tubes. A special 5-watt tube produced four-meter signals which were received up to one-mile distances. An interesting feature of the work was the interference caused by the presence of the observer while conducting the experiments, because of the action of the human body as a tuned antenna at these wavelengths.

A General Theory of the Correlation of Time Series of Statistics.[7] M. K. ZINN. In mathematical physics elaborate methods have been devised for dealing with oscillatory systems, damped or undamped, like pendulums, vibrating strings, vibrating telephone diaphragms, and with systems that are essentially damped but have no oscillatory characteristics, like the flow of heat and diffusion. The writer of this article approaches the problem of the economic structure with a point of view which sees the business world as behaving like such an oscillatory system.

The structure of economic society is a system exhibiting certain structural factors which express the tensions that are set up by a shift in prices here on some other factor like employment there. The problem is: How are these tensions to be inferred from statistics which describe the observed behavior of the economic world? The economist, unfortunately, cannot disconnect, say, the Federal Reserve system from the rest of the economic structure, connect it to a portable test

[6] *Proc. I. R. E.*, 15, 914, November, 1927.
[7] *The Review of Economic Statistics*, Harvard Economic Service, *9*, 184 (1927).

set and read off its economic impedance in the same way that an engineer can test a transformer.

He has to take instead the observed fluctuations in time of two series, say "wholesale commodity prices" and "commercial paper rates," and try to infer from such data the way in which changes in one of these quantities react on the other. The paper is concerned with the general way of doing this with a full analysis of the relation between these two economic variables from this point of view.

Just as most of the theory of oscillatory systems in mathematical physics is confined to linear systems, so the writer finds it convenient to assume linear relations between the economic variables. This greatly simplifies the mathematics. The formulas come out to be analogous to some formulas of electric circuit theory when approached from the Heaviside operational standpoint. It thus appears that much alternating current theory may come to be of value in studying economic variations. It is pleasant to think that some years hence we may be using the language of "a.c." theory (reluctance, susceptance, impedance, etc.) to describe the functional relation of one economic unit to the rest of society. Perhaps to study the relation of conditions in one part of the country to those in another we shall be using the long line transmission theory. Perhaps the theory will show us how to build economic band-pass filters which will protect us from too great fluctuations in business conditions, etc.

The application of the ideas of oscillatory systems to economics, here well started, is a subject which, it is believed, will strike a responsive chord in the heart of every electrical engineer and every mathematical physicist.

Contributions of Chemical Science to the Communications Industry.[8] CLARENCE G. STOLL. The author considers the improvements that have been made under a four-fold grouping of materials into electrically conducting, magnetic, insulating, and materials for apparatus structures. Particular emphasis is laid on the influence the chemist has exerted on the control of materials for manufacturing purposes. So many undesirable variations in manufactured products are caused by lack of uniformity in the raw materials that the aid of the chemist in standardizing testing and sampling methods cannot well be overestimated.

In conclusion, he pays homage to the readiness with which chemists have responded to demands for improvement in many of the raw materials and suggests that because of it similar demands may be of

[8] *Journ. Ind. and Engr. Chem.*, Vol. 19, p. 1132, 1927.

more frequent occurrence in the future. Among such possible innovations he mentions a better conductor of electricity, a cable covering superior to the present lead alloys, higher grade insulation, and a contact material less subject to corroding and freezing. The recent advances made in metallic alloys give every promise of a more brilliant future.

The Thickness of Spontaneously Deposited Photoelectrically Active Rubidium Films, Measured Optically.[9] H. E. IVES and A. L. JOHNSRUD. Measurements of the phase shift on reflection of polarized light from a reflecting surface on which there is deposited a very thin film of metal are described. The materials were the thin films of rubidium which are slowly deposited on glass or platinum when these have been thoroughly out-gassed. The apparatus was so arranged that measurements of the photo-electric activity could be made simultaneously with determinations of the optical effect due to the thin film. The data were then interpreted in terms of the electromagnetic theory of light using special developments due to T. C. Fry, to be published in the *Journal of the Optical Society* for January, 1928.

It is concluded from the measurements that the film of rubidium deposited on glass after fourteen days is of the order of magnitude of one atom thick. The theoretical effect of a layer of rubidium on platinum of the order of even several atoms thick is, however, so small as to lie within the errors of observation. "It thus appears, if the validity of the optical measurements of thickness is conceded, that the photo-electric emission is obtained when a layer of rubidium of approximately one atom in thickness is present," they conclude.

[9] *Opt. Soc. Amer. Jl.*, Vol. 15, 374, December, 1927.

Contributors to this Issue

E. C. WENTE, A.B., University of Michigan, 1911; S.B. in Electrical Engineering, Mass. Inst. of Technology, 1914; Ph.D., Yale University, 1918; instructor in physics and mathematics, Lake Forest College, 1911–12; Engineering Department, Western Electric Company, 1914–16, 1918–24; Bell Telephone Laboratories, 1924–. Mr. Wente has worked principally on general acoustic problems and on the development of special types of acoustic devices.

E. H. BEDELL, S.B., Drury College, 1924; University of Missouri, 1924–25; Bell Telephone Laboratories, Inc., 1925–. Since coming to the Laboratories Mr. Bedell has studied various acoustic problems, notably those of an architectural character.

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 is well known through his theoretical transmission studies and has published extensively on electric circuit theory and electric wave propagation.

PAUL P. COGGINS, Harvard University, A.B. in Mathematics, 1920, A.M. in Physics, 1921; Department of Development and Research, American Telephone and Telegraph Company, 1921–27. Statistician, New Jersey Bell Telephone Company, October 1927–. Up to October 1, 1927, Mr. Coggins dealt with the application of the mathematical theory of probabilities, including sampling theory, to various telephone problems.

W. J. SHACKELTON, B.S. in E.E., University of Michigan, 1909; Western Electric Company, Manufacturing and Installation Department, 1909–10; Bell Telephone Laboratories, 1910–. Mr. Shackelton's principal activities have been in connection with the design of loading coils and the development of methods of high frequency measurement.

J. G. FERGUSON, B.S., University of California, 1915; M.S., 1916; research assistant in physics, 1915–16; Bell Telephone Laboratories,

1917–. Mr. Ferguson's work has been in connection with the development of methods of electrical measurement.

C. J. DAVISSON, B.Sc., University of Chicago, 1908; Ph.D., Princeton University, 1911; instructor in physics, Carnegie Institute of Technology, 1911–17; research engineer, Western Electric Company and Bell Telephone Laboratories, 1917 to date. Dr. Davisson's work since coming with the Bell System has related largely to thermionics and electronic physics.

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., Bell Telephone Laboratories, 1919–.

C. R. KEITH, B.S., 1922, California Institute of Technology; M.A., 1925, Columbia University; Carrier Research Department, Bell Telephone Laboratories, 1922–. Mr. Keith's work has related to the study of vacuum tube and magnetic modulators and other carrier apparatus.

A. L. THURAS, B.S., University of Minnesota, 1912; E.E., 1913; laboratory assistant with U. S. Bureau of Standards, 1913–16; graduate student in physics, Harvard, 1916–17; scientific observer with U. S. Coast Guard, 1917–19; oceanographer, 1919–20; Bell Telephone Laboratories, 1920–. Since joining the Laboratories staff, Mr. Thuras' work has had to do largely with mechanical impedance studies and bridges.