The Bell System Technical Journal
October, 1923

Mutual Impedances of Grounded Circuits

By GEORGE A. CAMPBELL

SYNOPSIS: Formulas are derived for the direct-current mutual resistance and inductance between circuits grounded at the surface of the earth. For circuits composed of straight filaments, the mutual inductance is reduced to known Neumann integrals which involve only comparatively simple expressions for the case of horizontal, coplanar conductors above, below or on the surface of the earth. Numerical values for these integrals may be readily obtained from new and accurate graphs for straight filaments which meet at a point or start from a common perpendicular. It is shown that these new results supply a useful first approximation to the actual alternating-current mutual impedance of grounded circuits, when the frequency and extent of the circuits are not larger than occur in many practical applications.

1. INTRODUCTION

THE important discovery of the possibility of using the earth as the return conductor for electric telegraphic communication was announced by Steinheil in the *Comptes Rendus* of September 10, 1838, and throughout the entire development of telegraphy grounded circuits have been extensively employed. Considering the extensive application of such a capital discovery extending over a period of 85 years, it is surprising that so little is known quantitatively about grounded circuits. We have, however, long known that conditions are not of the extreme simplicity pictured under the early view that the earth acts as a reservoir presenting no resistance to the return current and introducing no interference between parallel returns. This view was expressed in 1857 by Bakewell as follows: "There is no mingling of currents, the electric current of each battery being kept as distinct as if separate wires were used both for the transmitted and the return currents. It would indeed be as impossible for the separate currents transmitted from the two batteries to be mingled together as it would be for the written contents of two letters enclosed in the same mail bag to intermix."

Measurements made a few years ago of the mutual impedances between grounded circuits which are restricted to a territory six miles square, at frequencies of 25 to 60 cycles per second, showed that within 10 per cent. the mutual reactance increased in the same ratio as the frequency. It was inferred that the effective inductance under the conditions of these tests was approximately the same as for direct current, or in other words, the incomplete penetration of the alternating currents into the earth was not of controlling importance in tests upon this scale.

This led to my making a theoretical investigation of the mutual inductance between direct-current grounded circuits which did, in fact, show that the calculated numerical results are in reasonable agreement with these actual experimental data. It is the purpose of this paper to describe this work; the mathematical discussion of the theoretical corrections for the incomplete penetration of alternating currents into the earth will form the subject of another paper.

2. Distribution of Current, Potential and Magnetic Force with Direct-Current Earth Return Flow

On the assumption of an infinite earth of uniform resistivity the lines of flow and the equipotential surfaces for a direct current I entering the earth at a point source at A and leaving the earth at a point sink at B, both A and B being on the surface of the earth, assumed flat, and the distance $AB=2b$, are given by the equations,[1]

$$\left.\begin{aligned}\frac{C}{I} &= \tfrac{1}{2}(\cos\theta_1-\cos\theta_2)\\ &= \tfrac{1}{2}\left(\frac{x_1}{r_1}-\frac{x_2}{r_2}\right)\\ &= \sin\tfrac{1}{2}(\theta_1+\theta_2)\sin\tfrac{1}{2}(\theta_2-\theta_1)\\ &= b\frac{\sin^2\theta}{r}, \text{ if } \frac{b}{r} \text{ is small,}\end{aligned}\right\} \quad (1)$$

$$\left.\begin{aligned}\frac{V}{I} &= \frac{\rho}{2\pi}\left(\frac{1}{r_1}-\frac{1}{r_2}\right)\\ &= -\frac{b\rho}{\pi}\frac{\cos\theta}{r^2}, \text{ if } \frac{b}{r} \text{ is small,}\end{aligned}\right\} \quad (2)$$

where C is the total current flowing in the earth, from the source at A to the sink at B, outside the current sheet of revolution defined by (1); and V is the potential, with respect to the midplane, upon the equipotential surface of revolution defined by (2).

These equipotential lines and stream lines are identical with the equipotential lines and lines of force for a uniformly magnetized filament. The formulas may be checked by regarding the return flow from A to B as being due to the superposition of two flows, a

[1] The coordinates used in this paper (x_1, y, z), (x_2, y, z), (x, y, z) and (r_1, θ_1, ϕ), (r_2, θ_2, ϕ), (r, θ, ϕ) are rectangular and spherical coordinates with origins at A, B and the midpoint of AB, respectively, the direction AB being the polar axis or positive x-axis in all cases, z being vertical, and ϕ being measured from the earth's surface in the plane perpendicular to AB.

return flow of direct current I from the point A to some infinitely distant point and a second return flow of direct current I from this infinitely distant point to the point B. For these component flows the current diverges from A or converges towards B radially and with equal intensity in all directions in the earth; the total current for one of the component flows flowing through any surface in the earth will thus be equal to $I/2\pi$ times the solid angle subtended at A or B, respectively, by the boundary of the surface, since the entire solid angle filled by the earth at a point on the surface is 2π. The total radial flow from A through the lower half of the circular cone having its axis in AB, the elements of the cone making the angle θ_1 with AB, is $\frac{1}{2}I(1-\cos\theta_1)$; similarly, the total radial flow toward B through the lower half of a cone with the angle $\pi-\theta_2$ will be $\frac{1}{2}I(1+\cos\theta_2)$. For the combined superposed flows the total current flowing through the semicircle in which the cones intersect is the sum of these two values or $\frac{1}{2}I(2-\cos\theta_1+\cos\theta_2)$, from which (1) is immediately obtained, since the total current flowing in the earth from A to B is I. This assumes that the semicircle lies between A and B, but the same formula holds for the entire current sheet of revolution. The lines of flow in the earth are symmetrical about AB and lie in planes through AB, since, in the earth, both component flows are symmetrical about AB.

For the component flows the equipotential surfaces are hemispherical and, since the resistance of a hemispherical shell of radius r, thickness dr, is $\rho dr/2\pi r^2$, the potentials at distances r_1 or r_2 from A or B, referred to the potential at infinity, are $I\rho/2\pi r_1$ or $-I\rho/2\pi r_2$, respectively, from which equation (2) follows by addition.

Fig. 1 accurately reproduces the flow and equipotential lines as given by formulas (1) and (2). At the midpoint of a line of flow its distance from each electrode is $r_1=r_2=bI/C$ and it may be shown that every other point of a line of flow is at a still shorter distance from the nearer electrode. It follows, for example, that less than 1/10 of the total current reaches, in its flow through the earth, any point lying at a distance greater than $5AB$ from the line AB connecting the electrodes.

If a uniform radial flow of current I in the horizon plane converging on the point A is combined with the uniform radial flow in the earth outward from A, we have a closed flow which is symmetrical about the vertical axis through A. Below the horizon plane the magnetic lines of force will be horizontal circles and the magnetic force at any point distant r_1 from A, α being the angle included between r_1 and the nadir, is $H=2I(1-\cos\alpha)/(r_1\sin\alpha)$

$=2I\ r_1^{-1}$ tan $(\alpha/2)$. Above the horizon plane there is no magnetic field, since any magnetic lines of force are, by symmetry, horizontal circles and the intensity is zero, since there is no current threading any horizontal circle above the surface of the earth.

Superposing this closed flow and a similar closed flow through B from the earth to the horizon plane, we obtain a closed flow from A

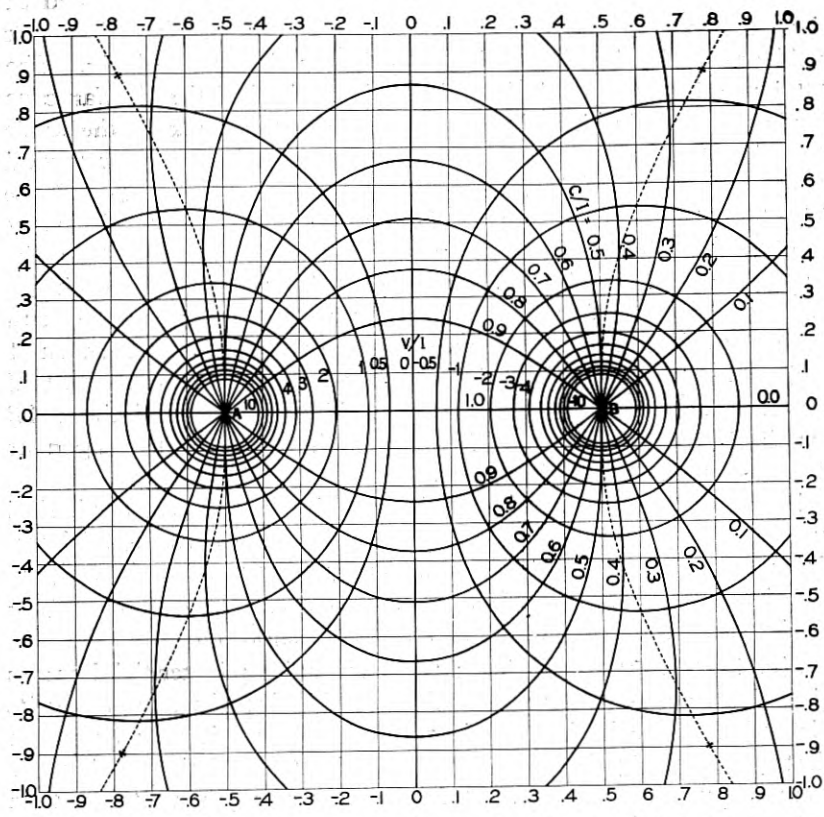


Fig. 1—Flow and equipotential lines on the earth's surface for an earth return flow from A to B. C/I is the fraction of the current flowing in the earth outside the flow surface of revolution; V/I is the resistance to the flow of the portion of the earth lying between the equipotential surface of revolution and the mid or zero potential plane, if the earth's resistivity is $\rho = 2\pi$. The flow and equipotential lines through each point on the dotted curve are perpendicular and parallel, respectively, to AB

to B in the earth and back from B to A in the horizon plane. The magnetic field for this closed flow, being the sum of the magnetic fields for the component flows, will be zero above the horizon plane, while below it will consist of lines of force in horizontal planes. This

result also applies to any closed flow which does not extend above the horizon plane and may be resolved into any number of component flows, each of which is radially symmetrical about a vertical axis.

3. Mutual Resistance of Grounded Circuits

By definition, if e.m.f.'s E and e in grounded conductors AB and ab produce the currents I and $i=0$ in the conductors, the mutual impedance between the two conductors is e/I. In the present case we are dealing with direct current and thus the mutual impedance is a mutual resistance Q, and by (2) its value is [2]

$$\left.\begin{aligned} Q &= \frac{\rho}{2\pi}\left(\frac{1}{Aa}-\frac{1}{Ab}-\frac{1}{Ba}+\frac{1}{Bb}\right) \\ &= \frac{\rho}{2\pi}\int\int \frac{d^2}{dSds}\left(\frac{1}{R}\right) dSds \\ &= \frac{\rho}{2\pi}\int\int \frac{-2dUdu+dVdv+dWdw}{R^3} \\ &= \frac{\rho}{2\pi}\int \left\{ \frac{\cos(\theta_1-\epsilon)}{r_1^2} - \frac{\cos(\theta_2-\epsilon)}{r_2^2} \right\} ds. \end{aligned}\right\} \quad (3)$$

The third form of (3) shows that the mutual resistance falls off as the inverse third power of the distance between grounded circuits when this distance has become large compared with the length of these circuits between grounding points.

The first form of (3) shows that the mutual resistance between grounded circuits does not depend upon the location of the conductors but only upon the location of the terminal grounding points A, B, a, b.

The mutual resistance for the case $\rho=2\pi$ is obtained from Fig. 1 by taking the value of V/I at the point corresponding to a reduced by its value at the point corresponding to b; if b is anywhere on the center line, for which $V/I=0$, the diagram gives directly the value of the mutual resistance. Employing ordinary units the diagram gives the mutual resistance directly in ohms if $AB=1$ mile and the earth has a resistivity of about one million ohms per centimeter cube (more exactly 1.011×10^6) which is its actual order of magnitude.

[2] In addition to the earlier notation there are employed in the different expressions for formula (3), and also in formula (5) below, the following: R is the distance between two elements dS and ds of any two paths extending from A to B and a to b; the rectangular projections of these elements along and perpendicular to R are dU, dV, dW and du, dv, dw, the two sets being parallel and with the same positive directions; θ_1, θ_2, ϵ are the angles which r_1, r_2 and ds make with AB, when the path ab lies in a plane with A and B.

4. Neumann Integrals for Return Flows

The required mutual inductances of grounded circuits will be found by means of the Neumann integral

$$N=\int\int\int\int(\cos\epsilon/r)dI\,dS\,di\,ds$$

extended over every current filament in both flows. Since the earth return portions of the two flows are independent of the flows in the arbitrarily located conductors on the earth's surface, it is convenient to divide the Neumann integral into four partial integrals which involve either no return flow, one return flow or both return flows according to the following formula [3]

$$\begin{aligned} N(\mathcal{X}-\mathcal{C})(x-o) &= N\mathcal{X}x - N\mathcal{X}o - N\mathcal{C}x + N\mathcal{C}o \\ &= N\mathcal{X}x - (\tfrac{1}{2}+\tfrac{1}{2}-1)\Delta, \text{ by Table I,} \\ &= N\mathcal{X}x. \end{aligned} \tag{4}$$

Checking the entries of Table I may be accomplished without performing more than two integrations. It will be convenient to make the integrals somewhat more general than is required in checking the table and find $N\mathcal{F}s$ and $N\mathcal{X}'h$ where $\mathcal{F}$ is any flow in space from A to B, which need not be coplanar points with the terminals a and b of s, and $\mathcal{X}'$ is any flow in a plane parallel to the horizon plane between terminal points A' and B'.

Consider first the part of a space return flow s which is radial from a in connection with an element dS on any filament of current dI of a flow $\mathcal{F}$ from A to B. The component dx of dS along the line x from a to ds is the only component which need be considered, since by symmetry the normal component contributes nothing to the Neumann integral. As the total radial flow is to be taken equal to unity, the amount flowing out through a ring, taken as the volume element, lying between the spheres of radii s and $s+ds$ and between the cones making angles θ and $\theta+d\theta$ with x will be $\frac{1}{2}\sin\theta\,d\theta$. If this ring lies at a distance r from dS the Neumann integral will be

$$N=\int_{Aa}^{Ba}dx\int_0^\infty ds\int_0^\pi\frac{\cos\theta\sin\theta\,d\theta}{2r},\quad r^2=x^2+s^2-2xs\cos\theta,$$

$$=\frac{1}{4}\int_{Aa}^{Ba}\frac{dx}{x^2}\int_0^\infty\frac{ds}{s^2}\int_{|x-s|}^{(x+s)}(x^2+s^2-r^2)dr$$

[3] Each term indicates the Neumann integral for the pair of flows designated by the script letter subscripts, as explained in the note accompanying Table I. Both $(\mathcal{X}-\mathcal{C})$ and $(x-o)$ are arbitrary flows on the earth's surface closed by earth return flows from A to B and from a to b, respectively.

$$=\frac{1}{3}\int_{Aa}^{Ba}\frac{dx}{x^2}\left\{\int_0^x sds+x^3\int_x^\infty\frac{ds}{s^2}\right\}$$

$$=\frac{1}{2}\int_{Aa}^{Ba}dx=\tfrac{1}{2}(Ba-Aa).$$

TABLE I

Entries are the value of k in the formula $N=k\Delta$ for the Neumann integral between the specified flows, $\Delta=-Aa+Ab+Ba-Bb$, or $2AB$ if $A=a$, $B=b$, and points A B, a, b are all on the earth's surface.

Flows†	$\mathcal{h}$	$\mathcal{s}$	$\mathcal{o}$	$\mathcal{a}$	$\mathcal{n}$	$\mathcal{x}$	$(\mathcal{h}-\mathcal{s})$	$(\mathcal{h}-\mathcal{o})$	$(\mathcal{h}-\mathcal{a})$	$(\mathcal{h}-\mathcal{n})$	$(\mathcal{h}-\mathcal{x})$	$(\mathcal{o}-\mathcal{a})$
Any surface $=\mathcal{X}$	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	1	0
Space return $=\mathcal{S}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0	0
Earth return $=\mathcal{E}$	$\frac{1}{2}$	$\frac{1}{2}$	1	0	$\frac{3}{2}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$	$\frac{1}{2}$	-1	1	1
Air return $=\mathcal{A}$	$\frac{1}{2}$	$\frac{1}{2}$	0	1	$-\frac{1}{2}$	$\frac{3}{2}$	0	$\frac{1}{2}$	$-\frac{1}{2}$	1	-1	-1
Nadir return $=\mathcal{N}$	0	$\frac{1}{2}$	$\frac{3}{2}$	$-\frac{1}{2}$	*	-1	$-\frac{1}{2}$	$-\frac{3}{2}$	$\frac{1}{2}$	*	1	2
Zenith return $=\mathcal{Z}$	0	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{3}{2}$	-1	*	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{3}{2}$	1	*	-2
Closed $(\mathcal{X}-\mathcal{S})$	$\frac{1}{2}$	0	0	0	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	1	0
" $(\mathcal{X}-\mathcal{E})$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{3}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	0	2	0	-1
" $(\mathcal{X}-\mathcal{A})$	$\frac{1}{2}$	0	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{3}{2}$	$\frac{1}{2}$	0	1	0	2	1
" $(\mathcal{X}-\mathcal{N})$	1	0	-1	1	*	1	1	2	0	*	0	-2
" $(\mathcal{X}-\mathcal{Z})$	1	0	1	-1	1	*	1	0	2	0	*	2
" $(\mathcal{E}-\mathcal{A})$	0	0	1	-1	2	-2	0	-1	1	-2	2	2

* Not of the assumed form and in general infinite.

† Each type of flow is designated by a script letter. Any flow in the surface of the earth, assumed to be a plane, is designated by $\mathcal{X}$, there being no restriction on the number of filaments of current but each filament must start from a common point, A, and extends to another common point, B. A special case of $\mathcal{X}$ is $\mathcal{H}$ the horizon return flow made up of two superposed uniform radial flows, from A to every point of the horizon and back to B. The space return flow, $\mathcal{S}$, is made up of two superposed uniform radial flows, one outward in all directions from the point A and the other inward from all directions toward the point B. The earth return and air return flows, $\mathcal{E}$ and $\mathcal{A}$, are similar except that the flows are uniformly distributed over all directions in the earth and air, respectively. The nadir return and zenith return flows, $\mathcal{N}$ and $\mathcal{Z}$, consist of two infinite vertical filaments from A and back to B by way of the nadir and zenith, respectively. Small script letters indicate similar types of flow with any independent terminal points, a and b. Differences such as $(\mathcal{X}-\mathcal{S})$ designate closed flows; thus $(\mathcal{X}-\mathcal{E})$ designates any flow on the earth's surface from A to B where it enters the earth and after spreading out uniformly through the earth returns to the terminal A, thus closing the flow.

To this is to be added the corresponding expression $\frac{1}{2}(Ab-Bb)$ for the radial flow converging on b, giving the result $\frac{1}{2}\Delta$. As the path of the line of flow between A and B does not enter into the result, it is immaterial whether the flow is confined to a single filament or is spread out in any way whatsoever in space, provided only all stream lines extend from A to B, as assumed for $\mathcal{F}$. Thus

$$N_{\mathcal{F}_s} = \tfrac{1}{2}\Delta = N_{\mathcal{S}_f}.$$

To find $N_{\mathcal{X}'_A}$ let r and s be the distances of any element dS of a line of flow forming a part of $\mathcal{X}'$ from a and from any element of a plane radial flow from a, respectively, the projections of r and s on either of the planes being x and y, which include the angle ϕ; $s^2=y^2+z^2$, $r^2=x^2+z^2$. The component of dS parallel to x will be dx and this is the only component which need be considered, since, on account of the symmetry of the radial flow, the normal component in the plane of flow contributes nothing to the integral.

$$\begin{aligned}
N &= \int_{Aa}^{Ba} dx \int_0^\infty \int_0^{2\pi} \frac{x-y\cos\phi}{x^2+y^2-2xy\cos\phi}\,\frac{y\,dy\,d\phi}{2\pi s} \\
&= \frac{1}{2\pi}\int_{Aa}^{Ba}\frac{dx}{x}\int_z^\infty ds\int_0^\pi\left(1+\frac{x^2-y^2}{x^2+y^2-2xy\cos\phi}\right)d\phi \\
&= \frac{1}{2\pi}\int_{Aa}^{Ba}\frac{dx}{x}\int_z^\infty ds\left[\phi-\sin^{-1}\frac{(x^2+y^2)\cos\phi-2xy}{x^2+y^2-2xy\cos\phi}\right]_0^\pi \\
&= \int_{Aa}^{Ba}\frac{dx}{x}\int_z^r ds,
\end{aligned}$$

since inspection shows that the two values of the definite integral 2π and 0 are to be used for $s \lesseqgtr r$ respectively, and therefore

$$\begin{aligned}
N &= \int_{Aa}^{Ba}\frac{r-z}{x}\,dx \\
&= \int_{A'a}^{B'a}\frac{r\,dr}{r+z} \\
&= \Big[r-z\log(r+z)\Big]_{A'a}^{B'a} \\
&= (B'a-A'a)-z\log\frac{B'a+z}{A'a+z}.
\end{aligned}$$

This is for the radial flow from a. Adding the corresponding expression

for the radial flow towards b, we have finally for the complete integral

$$N\mathcal{X}'_{\mathcal{A}} = (-A'a + A'b + B'a - B'b) - z \log \frac{(A'b+z)\,(B'a+z)}{(A'a+z)\,(B'b+z)}, \quad (4a)$$

which becomes, if $z=0$,

$$N\mathcal{X}_{\mathcal{A}} = \Delta = N\mathcal{H}_{\mathcal{K}}.$$

The first line of Table I can now be filled in at once since the integrations have shown that the first two values of k are 1 and $\frac{1}{2}$; the next two entries are also $\frac{1}{2}$ since by symmetry $N\mathcal{X}_{\mathcal{S}} = N\mathcal{X}_{\mathcal{O}} = N\mathcal{X}_{\mathcal{A}}$; $N\mathcal{X}_{\mathcal{N}} = N\mathcal{X}_{\mathcal{Z}} = 0$ since the nadir and zenith flows are perpendicular to the $\mathcal{X}$ flow in the horizon plane. The remaining six entries in the first row are for closed flows which are expressed as differences between the flows already considered, and the corresponding k's are the differences of the k's for the component flows. The first column of k's may also be filled in, the table being symmetrical, since interchanging capital and small script letters leaves N unchanged and $\mathcal{A}$ is a special case of $\mathcal{K}$.

The second row of the table involves only special cases of $N\mathcal{S}_{\mathcal{I}}$ and only the values $\frac{1}{2}$ and 0 occur.

From the flows included in the table nine pairs of closed flows may be formed having zero mutual inductances, because one of the closed flows of each pair has no magnetic field below the surface of the earth and the other closed flow includes no current above the surface of the earth, and thus there is no interlinkage of induction between the two closed flows. The portion of Table I referred to is repeated in Table II, where the flows at the top are those for which any difference such as $(\mathcal{A}-\mathcal{A})$ has no magnetic field below the earth's surface, just as $(\mathcal{H}-\mathcal{E})$ has no magnetic field above the earth's surface, as was proved above, while no flow at the side penetrates above the earth's surface.

TABLE II

	$\mathcal{A}$	$\mathcal{A}$	$\mathcal{Z}$
$\mathcal{X}$	1	$\frac{1}{2}$	0
$\mathcal{E}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
$\mathcal{N}$	0	$-\frac{1}{2}$	-1

The top row of Table II includes only values of k already found, and the remainder of the first column follows from symmetry and the

fact that λ is a special case of the general flow $\varkappa$. Any other entry in Table II is now found in terms of three of the entries in this border, thus

$$0 = N(\varkappa - \varepsilon)(\lambda - a) = N\varkappa\lambda - N\varkappa a - N\varepsilon\lambda + N\varepsilon a = (1 - \tfrac{1}{2} - \tfrac{1}{2})\Delta + N\varepsilon a,$$

and we have the interesting result $N\varepsilon a = 0$; the remainder of the table follows.

The result $N\eta\varepsilon = -\Delta$ may be readily checked directly since it involves only the mutual Neumann integral between straight parallel filaments, and by using the expanded form of the integral for equal filaments beginning at a common perpendicular with opposite positive directions[4] the result can be written down at once.

The important difference Δ which is utilized in Table I may be expressed in the following useful forms:

$$\left.\begin{aligned} \Delta &= (-Aa + Ab + Ba - Bb) \\ &= -\int\int \frac{d^2R}{dSds}\, dSds \\ &= \int\int \frac{dVdv + dWdw}{R} \\ &= 2\int \sin \tfrac{1}{2}(\theta_2 - \theta_1) \sin [\tfrac{1}{2}(\theta_1 + \theta_2) - \epsilon]\, ds, \end{aligned}\right\} \qquad (5)$$

where the notation is the same as for formula (3) above. The third form of (5) shows that when the separation R is great the mutual inductance varies inversely as the first power of the separation.

5. Mutual Inductance Between Grounded Circuits Lying on the Surface of the Earth

It has now been shown that *for direct currents the mutual inductance between grounded circuits consisting of conductors lying on the surface of the earth and grounded at their terminals is equal to the mutual Neumann integral between the conductors alone*, since in the complete Neumann integral for the closed flows the total contribution of those parts which involve the ground returns is zero. For low frequencies the effective inductance can differ but little from the direct-current inductance, and it is therefore of practical importance to investigate

[4] Mutual Inductances of Circuits Composed of Straight Wires. *Physical Review*, 5, pp. 452–458, June, 1915, formula (6).

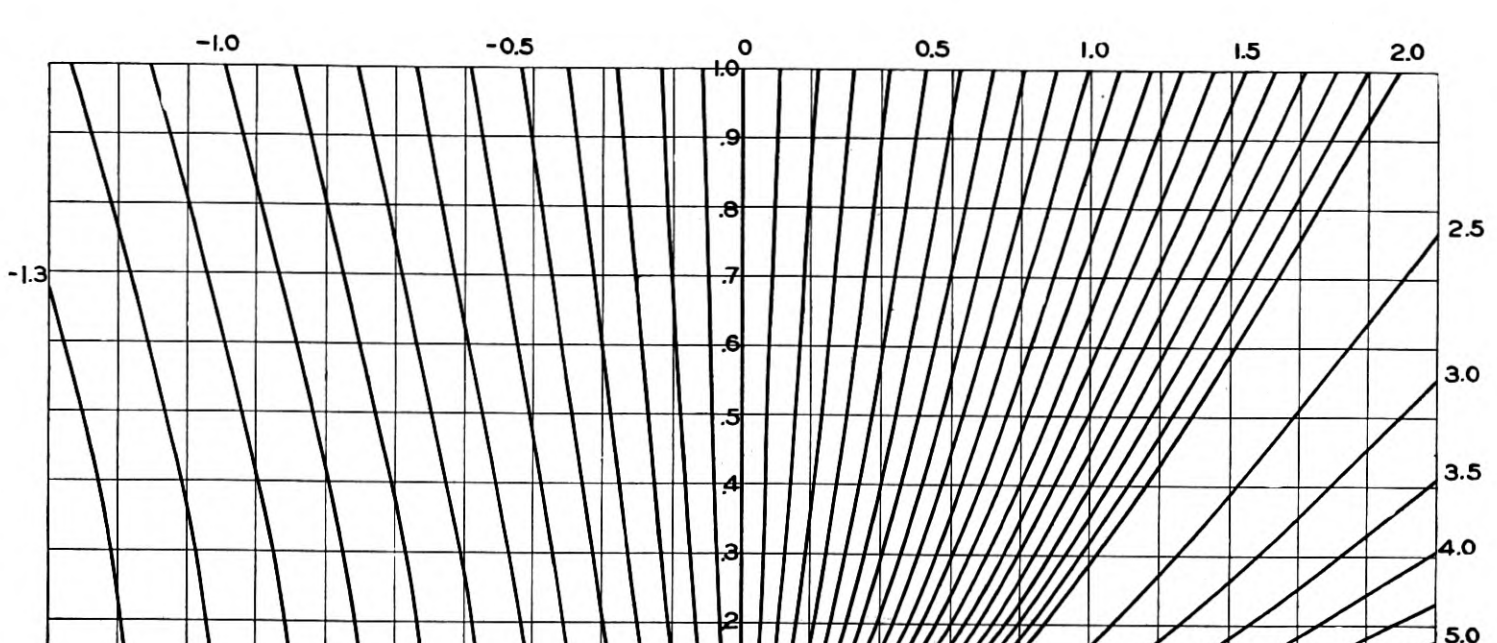


Fig. 2—Contour lines of the mutual Neumann integral between two straight filaments meeting at a point, one filament being of unit length

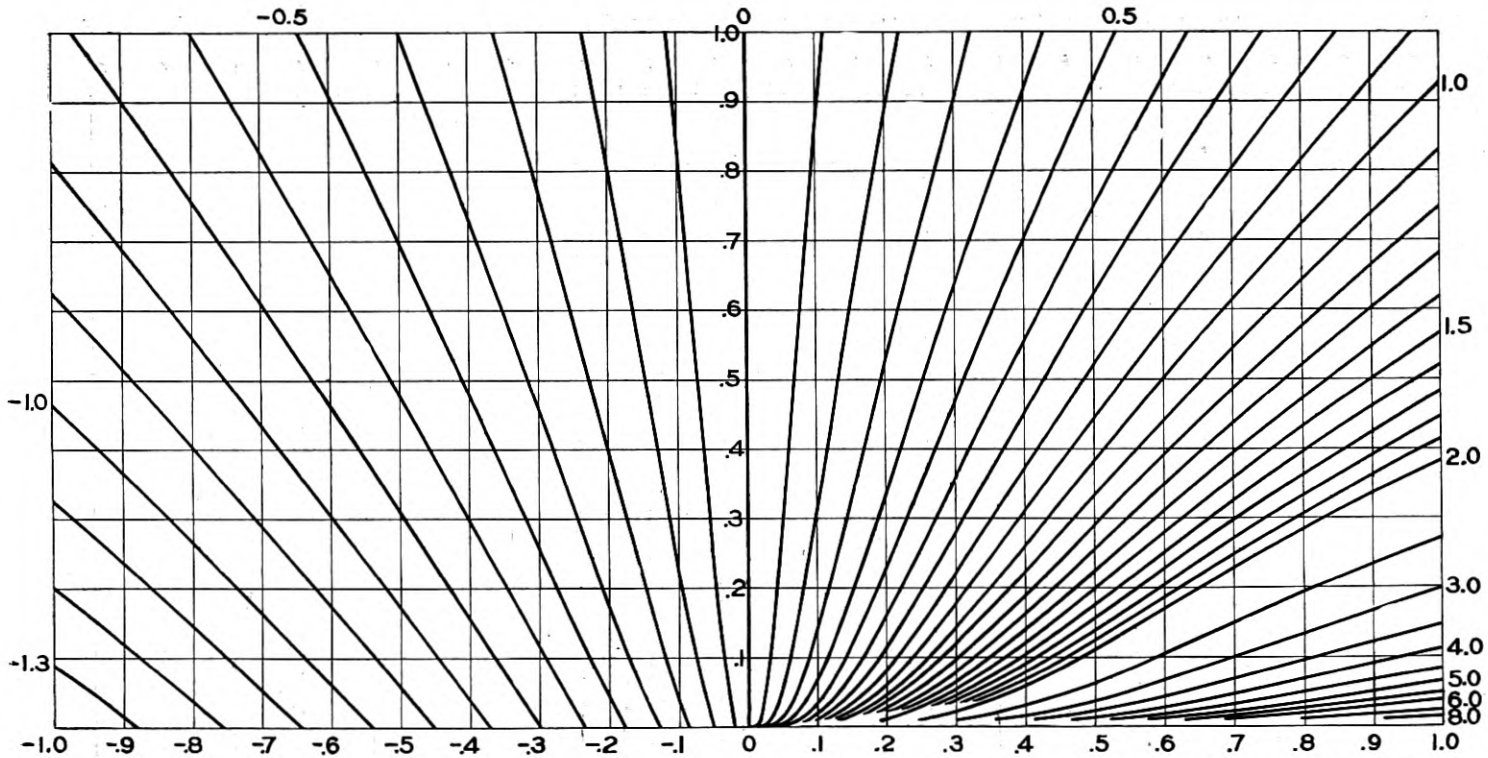


Fig. 3—Contour lines of the mutual Neumann integral between two parallel filaments starting at a common perpendicular, one filament being of unit length

the numerical magnitude of these Neumann integrals between grounded circuits. In order to visualize the magnitudes involved and supply means by which they may be readily calculated, a number of diagrams have been prepared for the important case of straight conductors.[5]

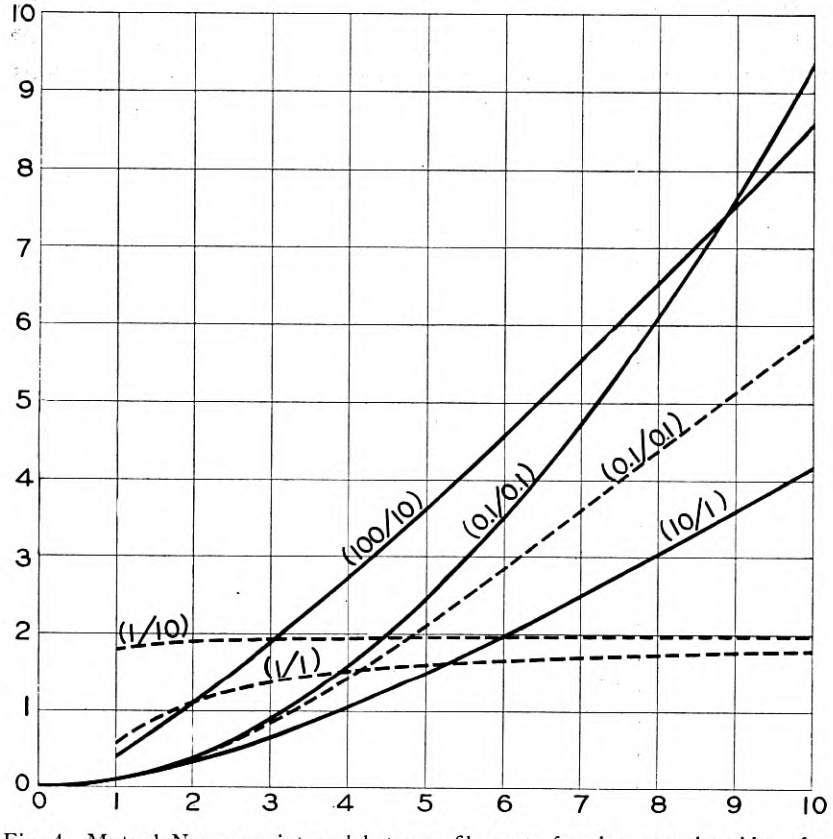


Fig. 4—Mutual Neumann integral between filaments forming opposite sides of a rectangle with unit separation. The dashed curves show the mutual resistance between the filaments if the circuits are grounded through the earth and its resistivity is $\rho = 2\pi$. The numerator and denominator of the bracketed fraction on each section of the curve show the factors by which the vertical and horizontal scales must be multiplied for use on this section

If the two straight conductors OA, Oa start from a common point O, the mutual Neumann integral is shown by Fig. 2; the curves give the locus of terminal a for constant values of the integral when the other conductor OA is the unit base. The Neumann integral between

[5] The necessary formulas are given in the paper loc. cit. Additional transformations of these formulas are given in the appendix to the present paper.

any two straight filaments $\mathscr{R}$ and r having terminals A, B and a, b may be expressed as

$$N_{\mathscr{R}r} = N_{(Aa)} - N_{(Ab)} - N_{(Ba)} + N_{(Bb)}, \tag{6}$$

where $N_{(Aa)}$ stands for the Neumann integral between the two straight filaments, OA and Oa, beginning at O, the point of intersection of $\mathscr{R}$ and r, extended if necessary, and ending at the terminals A and a; thus

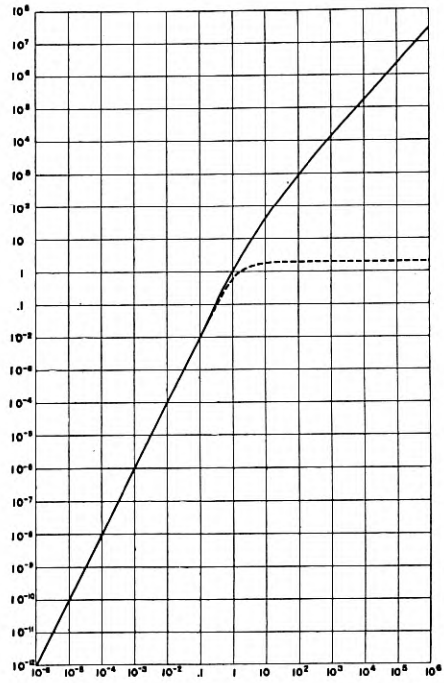


Fig. 5—Mutual Neumann integral between filaments forming opposite sides of a rectangle with unit separation. The dashed curves show the mutual resistance between the filaments if the circuits are grounded through the earth and its resistivity is $\rho = 2\pi$. This is Fig. 4, but with logarithmic scales

the integral between any two filaments which would intersect within a finite distance may be readily found after reading four values from Fig. 2.

In the special case of parallel filaments Fig. 2 fails; the corresponding curves for this case are presented by Fig. 3, which assumes a unit base filament and a parallel filament starting at a point on the left hand perpendicular to the base. Differences will give the general case

of parallel filaments which do not start at a common perpendicular, which may thus be derived from Fig. 3.

The mutual Neumann integral between any two parallel filaments may also be obtained by means of the formula

$$N_{\mathcal{R}r} = \tfrac{1}{2}[-N_{(Aa)} + N_{(Ab)} + N_{(Ba)} - N_{(Bb)}], \qquad (7)$$

where $N_{(Aa)}$ now stands for the Neumann integral between the projections of Aa on $\mathcal{R}$ and r, extended if necessary, the projections

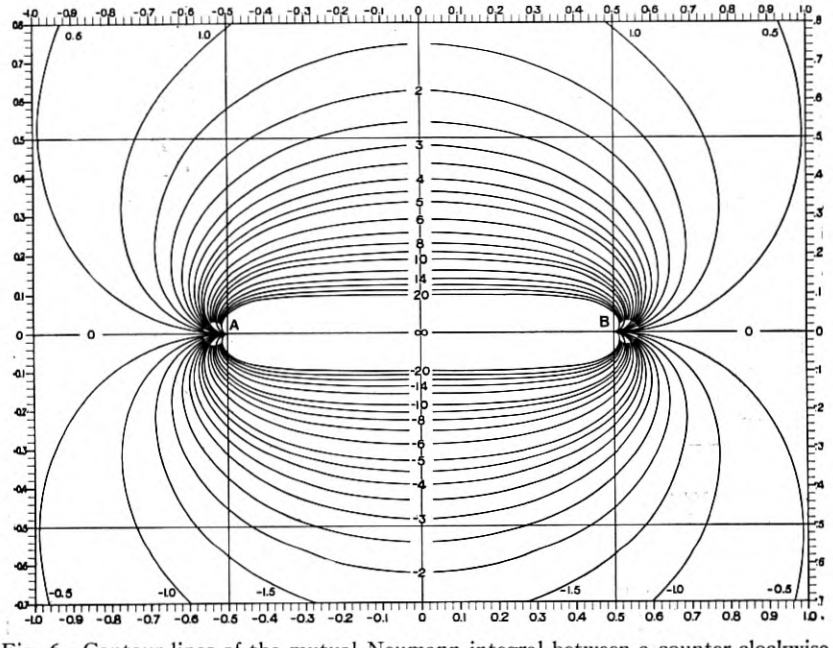


Fig. 6—Contour lines of the mutual Neumann integral between a counter-clockwise small loop on the surface of the earth, per unit area, and a straight grounded filament AB of unit length

having the same or opposite positive directions in agreement with $\mathcal{R}$ and r. This formula for the mutual Neumann integral presents the advantage of requiring only a single entry diagram, which is supplied by Fig. 4 and on a logarithmic scale by Fig. 5.

The mutual inductance may be required between a small, closed loop lying upon, but insulated from, the surface of the earth and a straight grounded conductor. The value depends upon the location, area and assumed positive direction around the loop, but is independent of the shape of the loop. Contour curves for the mutual inductance per unit area of the loop are given by Figs. 6 and 7; the

positive direction around the loop is counter-clockwise; the straight grounded conductor AB of Fig. 6 is of unit length while in Fig. 7 grounded terminal B alone appears, terminal A being at an infinite

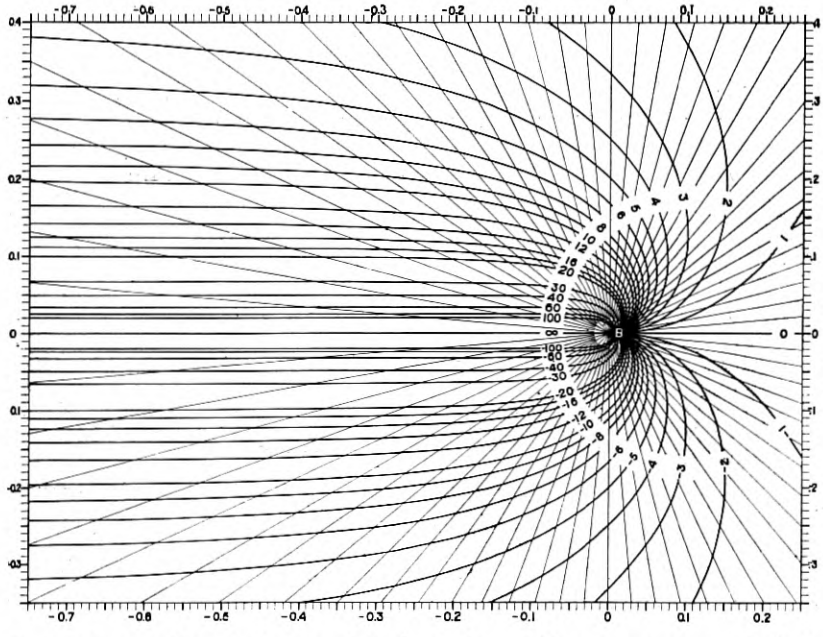


Fig. 7—Contour lines of the mutual Neumann integral between a counter-clockwise small loop on the surface of the earth, per unit area, and a straight grounded filament of infinite length

distance to the left. These curves show the vertical component of the magnetic field due to unit current in AB. The formulas employed for $AB=1$ and infinity, respectively, are [6]

$$-\frac{d^2N}{dxdy}=\frac{2y(r_1+r_2)}{r_1r_2[(r_1+r_2)^2-1]}=\frac{r_1+r_2}{r_1r_2}\sqrt{\frac{1-(r_1-r_2)^2}{(r_1+r_2)^2-1}}, \tag{8}$$

$$-\frac{d^2N}{dxdy}=\frac{y}{r_2(x_2+r_2)}=\frac{1}{r_2}\tan\tfrac{1}{2}\theta_2. \tag{9}$$

Large loop mutual inductances may be calculated either by integrating the value of $-d^2N/dxdy$ over the loop or by integrating the value of dN/ds around the boundary. If the boundary may be approximated to by a broken straight line, the curves of Figs. 2 and 3 may be employed.

[6] These formulas may be derived by differentiating (8) of the paper loc. cit., $dN/dx=2\tanh^{-1}[AB/(As+sB)]$, with respect to y.

6. Mutual Impedances for Conductors Lying on the Surface of the Earth

In order to arrive as directly as possible at a concrete numerical idea of the magnitudes and angles occurring in the mutual impedances encountered in engineering work, we may advantageously start with the following specific constants:

Base Length (AB, OA or Aa) ..... 1 Mile,
Frequency of the Alternating Current. .... 1 Kilocycle,
Resistivity of the Earth per Centimeter Cube. 1 Megohm,

which are of the right order of magnitude and make the factors

$$\rho/(2\pi\, AB) = 10^6/(2\pi\ 0.1609 \times 10^6) = 0.989,$$
$$2\pi f \times AB \times 10^{-9} = 2\pi\ 0.1609 = 1.011,$$

which are both equal to unity within about 1 per cent., so that the approximate resistance and reactance components of mutual impedances may be read directly from Figs. 1–8 without applying multipliers. On the other hand, when mutual impedances are required for other lengths, frequencies and specific resistances, the correcting factors are readily applied. The tangent of the angle of the mutual impedance is proportional to the frequency, to the square of the linear dimensions of the circuits and to the reciprocal of the earth's resistivity.

Grounded circuits separated by a distance large compared with the dimensions of the circuits have a mutual impedance with negligible resistance component since ultimately this component varies inversely as the cube of the distance by (3), whereas ultimately the mutual reactance varies inversely with the distance.

For two parallel grounded conductors separated by one mile and forming opposite sides of a rectangle, the two components of the mutual impedance are shown by Fig. 5 for the assumed constants (approximately a kilocycle and a megohm). The resistance component of the mutual impedance is then always less than the reactance component; when the rectangle becomes a square, the mutual impedance angle is $\tan^{-1}1.595 = 57.9°$. Reducing the frequency to 0.627 kilocycles or reducing the side of the square to 0.792 miles or increasing the resistivity to 1.595 megohms would reduce this angle to 45°.

Consider two straight grounded conductors AB and ab, the latter distance being small compared with the other dimensions of the

system and assume that while ground a is fixed, ground b is rotated about a in a circle of fixed radius ab. This will vary the mutual impedance between the two grounded conductors. The reactance component will vary as the cosine of the angle between ab and AB. The resistance component will also vary as the cosine of an angle, measured, in general, from a direction other than AB. The maximum resistance component will also differ, in general, from the maximum reactance component. The locus of the mutual impedance obtained for all positions of ab will be an ellipse. The ellipse becomes a straight line when ground a lies on the bisecting normal of AB, for then the direction giving the maximum resistance component is parallel to AB and thus the same as for the maximum reactance component. The straight line limit is also obtained when ground a lies on the prolongation of AB in either direction, for the resistance component then has its maximum value in the direction opposite to AB. The maximum resistance component will be perpendicular to AB at the points on Fig. 1 where the C/I contour, if drawn, would be vertical. The locus of these points is

$$y^2 = (x^2 - b^2)^{2/3}[(x+b)^{2/3} + (x-b)^{2/3}]. \tag{10}$$

At remote points on this locus $y = \pm\sqrt{2}\,x$ and the maximum resistance is negligible compared with the maximum reactance since the circuits are widely separated, while in the neighborhood of A and B it is the maximum reactance which is negligible compared with the maximum resistance. At some intermediate point the two maxima are equal, and, since they occur for directions differing by 90°, the elliptical impedance locus becomes a circle; and the mutual impedance between the two grounded circuits does not change in magnitude as ground b is rotated about ground a. For the assumed constants (1 mile, 1 kilocycle and 1 megohm) this point lies at distances of 1.562 and 0.939 miles from the two terminals A and B, and its four possible locations are shown by the four small crosses on Fig. 1.

If a is rotated counter-clockwise about AB, the direction giving the maximum resistance component also rotates counter-clockwise, making two complete revolutions, while the ground a makes one revolution about AB.

7. Equivalent Ground Plane

To a first approximation the direct-current mutual inductance between two straight conductors AB and ab, forming opposite sides of a rectangle on the surface of the earth, at a separation Aa which

is small compared with the length AB is, by (25) of the appendix, neglecting the first and higher powers of $1/s$,[7]

$$N = 2AB \log \frac{\frac{2}{e}AB}{Aa} = 2AB \log \frac{0.736\,AB}{Aa}. \tag{11}$$

This expression has the form ($2\,l \log s/r$) of the commonly employed mutual inductance formula for two long parallel conductors, each of length l, separated by distance r, the common return being a perfectly conducting earth in which the image of each conductor is at the distance s from the other physical conductor. For our direct-current case, therefore, the effective distance to the images is about $\frac{3}{4}$ of the length of either grounded conductor. Since this distance is by assumption large compared with the distance between the conductors, the images are approximately at this same depth below the actual surface of the earth, and the hypothetical perfectly conducting earth would be at one-half this depth, or $\frac{3}{8}AB$. The effective image distance is necessarily directly proportional to the dimensions of the grounded circuits and independent of the earth's resistivity because the shape and relative distribution of the lines of flow are independent of the resistivity and of the length of the grounded circuits. Inspection of Fig. 1 shows that somewhat over $\frac{1}{2}$ of the return flow attains a distance $\frac{3}{4}AB$ from AB, while the remainder of the current remains closer to the grounded conductor.

It may be inquired what would be the effect of confining both return currents to a thin uniform conducting layer on the earth's surface, so that they become horizon return flows. For the closed flows ($\mathcal{X}-\mathcal{H}$) and ($\mathcal{x}-\mathcal{h}$) in general and the particular flows ($\mathcal{R}-\mathcal{H}$) and ($\mathcal{r}-\mathcal{h}$), where $\mathcal{R}$ and $\mathcal{r}$ are close parallel straight conductors,

$$\begin{aligned} N_{(\mathcal{X}-\mathcal{H})(\mathcal{x}-\mathcal{h})} &= N_{\mathcal{X}\mathcal{x}} - N_{\mathcal{X}\mathcal{h}} - N_{\mathcal{H}\mathcal{x}} + N_{\mathcal{H}\mathcal{h}} \\ &= N_{\mathcal{X}\mathcal{x}} - \Delta; \\ N_{(\mathcal{R}-\mathcal{H})(\mathcal{r}-\mathcal{h})} &= N_{\mathcal{R}\mathcal{r}} - 2Ab + 2Aa \\ &= 2AB \log \frac{\frac{2}{e^2}AB}{Aa} = 2AB \log \frac{0.271AB}{Aa}. \end{aligned} \tag{12}$$

[7] If the term $1/s = Aa/AB$ of the expansion is retained the equivalent ground plane has the depth $(AB+Aa)/e$ and thus becomes deeper as ab is moved away from AB. But the equivalent ground plane may be kept fixed at the distance AB/e from AB provided it is tipped at the angle $\sin^{-1}2/e = 47°$ so that ab moves away from the ground plane as it moves away from AB. If it were worth while, still closer approximations might be secured by using a perfectly conducting cylindrical earth of suitable cross-section.

Thus, the assumption that the return current is confined to the earth's surface does not change the order of magnitude of the effective image distance, but reduces it from about $\frac{3}{4}$ to about $\frac{1}{4}$ of the length of the exposure. For space returns the effective image distance is

$$2AB/e^{3/2}=0.446\ AB.$$

Now take another practical case by assuming that the conductor ab is of negligible length compared with its separation r from the parallel conductor AB and the separation is, in turn, negligible compared with the length AB. The formula for dN/dx given in footnote 6 shows that the required inductance depends only on the ratio $AB/(r_1+r_2)$ and is thus constant upon an ellipse. Equivalent expressions in logarithmic form are

$$N=2\ ab\log\left(2\cos\tfrac{1}{2}\theta_1\sin\tfrac{1}{2}\theta_2\frac{\sqrt{r_1r_2}}{y}\right) \tag{13}$$

$$=2\ ab\log\left(\frac{\cos\frac{1}{2}\theta_1}{\cos\frac{1}{2}\theta_2}\sqrt{\frac{r_1}{r_2}}\right), \tag{14}$$

or approximately

$$N=2\ ab\log\frac{AB}{r}, \text{ if } ab \text{ is opposite the midpoint of } AB, \tag{15}$$

$$N=\frac{1}{2}\left[2\ ab\log\frac{AB}{r_2}\right], \text{ if } ab \text{ is at distance } r_2 \text{ beyond } B.^{8} \tag{16}$$

Thus, from (13) the effective image distance is never greater than twice the geometrical mean distance from ab to A and B. Its maximum value is approximately AB and occurs when ab is opposite the midpoint of AB. Its minimum value is approximately $\sqrt{r_2AB}$ and occurs when ab is at the distance r_2 from A or B in the prolongation of AB. This makes the inductance one-half of what it is at the symmetrical position. Thus, wherever ab is placed, its mutual inductance with AB lies somewhere between 50 per cent. and 100 per cent. of the mutual inductance, due to an effective image distance AB, ab remaining always parallel to AB and the locus of ab being a rectangle with semicircular ends of radius r_2 and centers A and B.

8. Mutual Impedances of Grounded Circuits Which Depart from the Surface of the Earth

Consider a system of conductors following any paths in space and insulated from the earth except at two grounding points on the sur-

[8] The shortest distance from ab to AB might have been designated by a single letter in place of using y, r and r_2 in formulas (13), (15) and (16).

face of the earth. Any flow of current through this system of conductors may be divided into elementary filaments each of which is made up of segments beginning and ending at the earth's surface and not crossing the earth's surface between these terminals. Ground each segment at both ends. Let $\mathcal{W}$ and $\mathcal{U}$ designate segments having terminals A and B, $\mathcal{W}$ (for example, an open wire circuit) never going below the surface of the earth and $\mathcal{U}$ (for example, an underground cable circuit) never going above the surface of the earth. Add the underground flow $\mathcal{U}$ to Table II at the foot of the left-hand column and, proceeding as before,

$$0 = N_{(\mathcal{U}-\mathcal{X})(z-a)} = N_{\mathcal{U}z} - N_{\mathcal{U}a} + \tfrac{1}{2}\Delta = N_{\mathcal{U}z} + N_{\mathcal{U}o} - \tfrac{1}{2}\Delta,$$

$$0 = N_{(\mathcal{U}-\mathcal{X})(h-a)} = N_{\mathcal{U}h} - N_{\mathcal{U}a} - \tfrac{1}{2}\Delta = N_{\mathcal{U}h} + N_{\mathcal{U}o} - \tfrac{3}{2}\Delta,$$

$$\text{or } N_{\mathcal{U}o} = \tfrac{1}{2}\Delta - N_{\mathcal{U}z} = \tfrac{3}{2}\Delta - N_{\mathcal{U}h},$$

and similarly for the flow $\mathcal{W}$ above ground,

$$N_{\mathcal{W}o} = \tfrac{1}{2}\Delta + N_{\mathcal{W}n} = N_{\mathcal{W}h} - \tfrac{1}{2}\Delta.$$

Hence the three cases which may occur give

$$\left.\begin{aligned} N_{(\mathcal{W}-\mathcal{E})(w-o)} &= N_{\mathcal{W}w} - N_{\mathcal{W}n} - N_{\mathcal{H}w} \\ &= N_{\mathcal{W}w} - N_{\mathcal{W}h} - N_{\mathcal{H}w} + 2\Delta, \end{aligned}\right\} \tag{17}$$

$$\left.\begin{aligned} N_{(\mathcal{U}-\mathcal{E})(u-o)} &= N_{\mathcal{U}u} + N_{\mathcal{U}z} + N_{\mathcal{Z}u} \\ &= N_{\mathcal{U}u} + N_{\mathcal{U}h} + N_{\mathcal{H}u} - 2\Delta, \end{aligned}\right\} \tag{18}$$

$$\left.\begin{aligned} N_{(\mathcal{W}-\mathcal{E})(u-o)} &= N_{\mathcal{W}u} - N_{\mathcal{W}n} + N_{\mathcal{Z}u} \\ &= N_{\mathcal{W}u} - N_{\mathcal{W}h} + N_{\mathcal{H}u}. \end{aligned}\right\} \tag{19}$$

The importance of these equations lies in the fact that the earth return flows are replaced by the simpler nadir, zenith and horizon return flows. *If the conductors comprise only broken straight filaments, making any angles with each other and the earth, the required Neumann integrals, if we use the expressions involving the nadir and zenith returns, are the known expressions between straight filaments.* If the conductors lie in horizontal planes with vertical ground connections, it is convenient to employ the expressions involving the horizon return flows, since the required integral is (4a) derived above. Formulas (33)–(41) of the appendix are the resulting formulas for the three general cases and for a number of important special cases.

In general we may say that the effect of changing the height of one or both conductors by an amount which is small compared with the length of the conductors will be relatively small, since the effective

image distance has been shown above to be of the order of the length of the conductors. To illustrate this fact, in Fig. 8 four dotted curves have been added to the curve of Fig. 5 showing the mutual inductances of the two parallel grounded conductors when they are in the

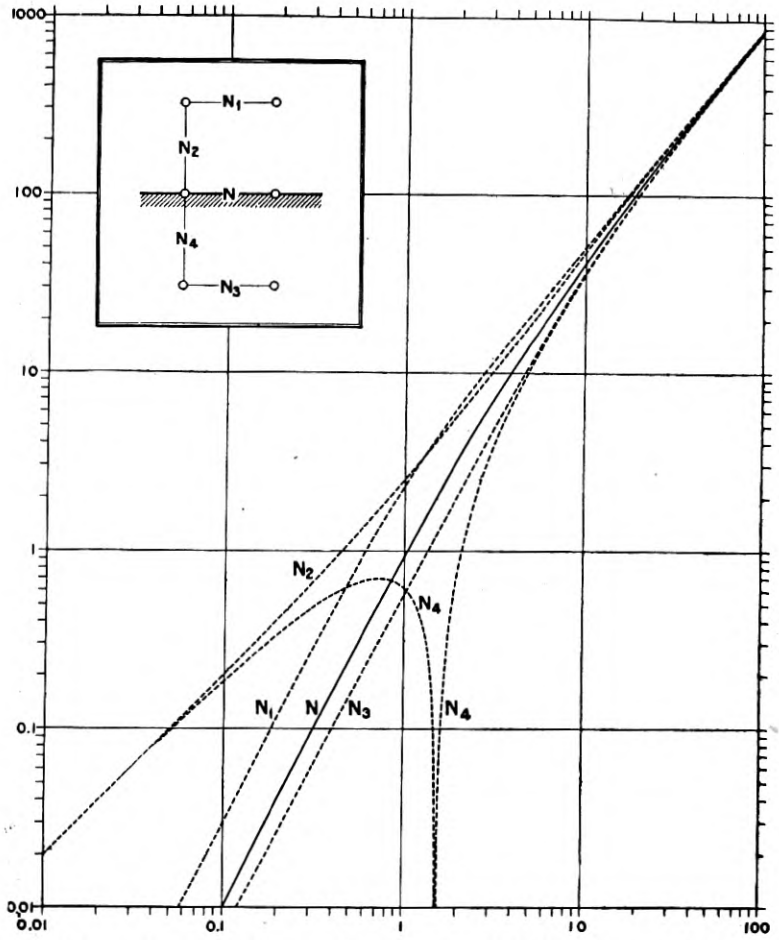


Fig. 8—Mutual Neumann integral between grounded filaments forming opposite sides of a rectangle with unit separation, on the surface of the earth (from Fig. 5) and between the same filaments when one or both of the filaments is raised above or depressed below the surface of the earth as shown by the insert

four positions indicated by N_1, N_2, N_3 and N_4 on the insert of Fig. 8, calculated by formulas (28)–(30) of the appendix, which are special cases of (35), (36), (39) and (40). When the conductors are long, the relative change in the mutual inductance is small. Depressing the

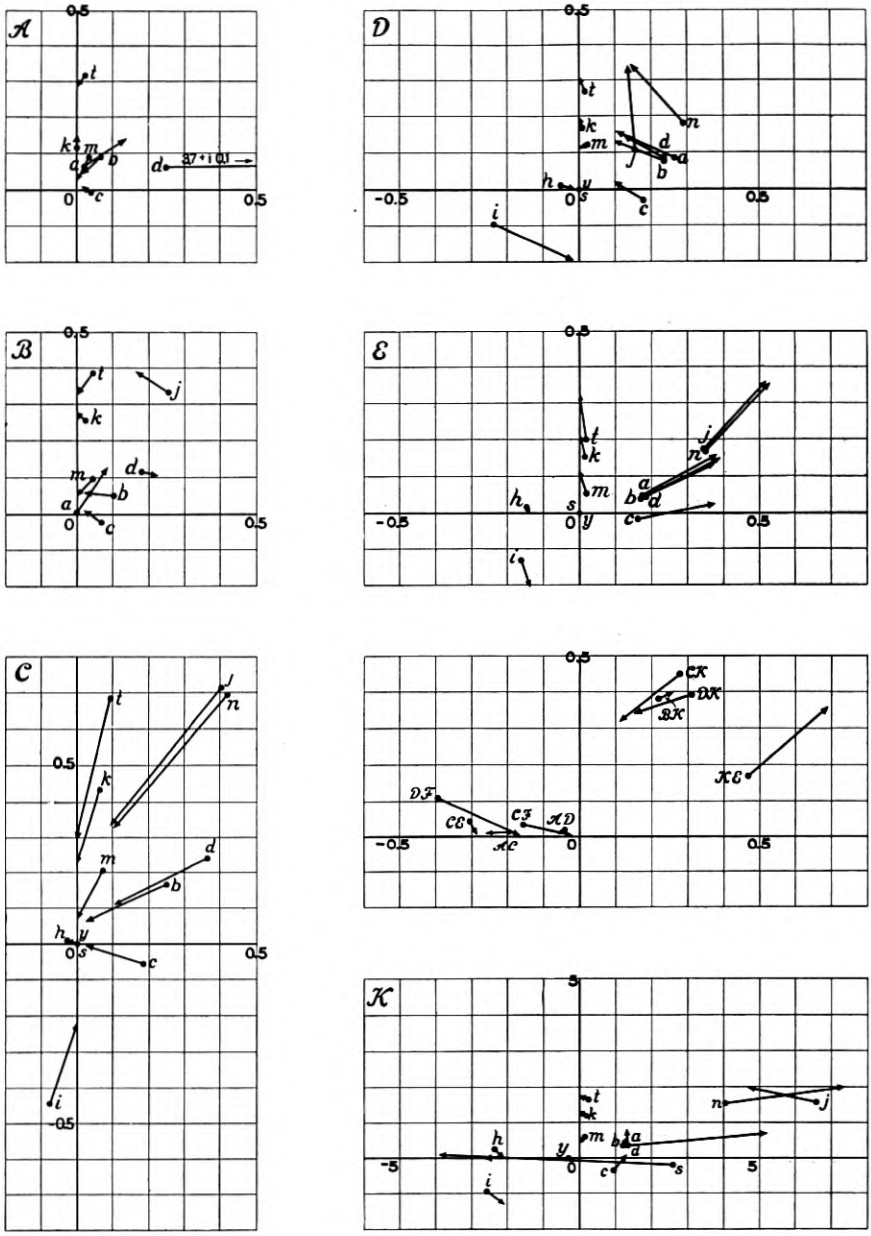


Fig. 9—Comparison of measured and calculated mutual impedances between the indicated circuit pairs. Each arrow extends from the measured impedance located at its tail to the calculated impedance at its head. The location of and positive directions for these circuits are shown by Fig. 10

wires reduces the inductance; the curves show that in the case of N_4 the inductance passes through zero and is reversed in sign when $s = 1.560$.

When the departure of the circuits from the earth's surface may be neglected, all terms, but the first, on the right-hand side of formulas (17)–(19) drop out, and each reduces to the simple, fundamental grounded circuit formula (4).

9. Comparison of Theoretical Results with Measurements at 25 and 60 Cycles

Fig. 9 shows, by means of arrows, the impedances which must be added to each of a large group of measured 25-cycle mutual impedances to obtain the results calculated by means of the preceding formulas, on the assumption that the earth has a uniform resistivity

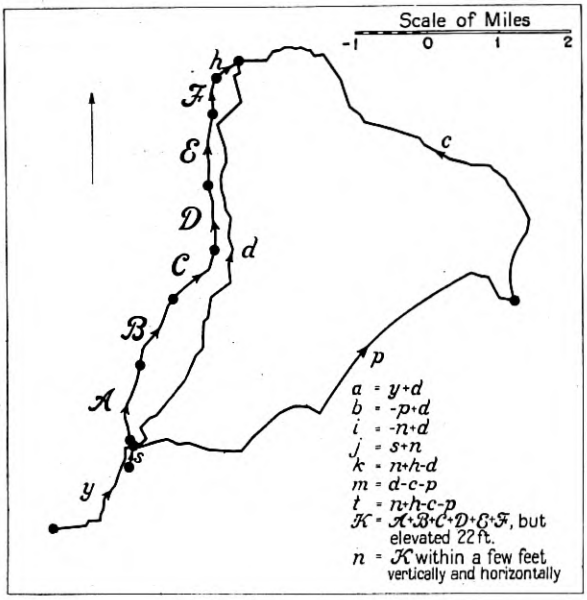


Fig. 10—Location of test circuits, the arrow-heads showing the positive directions in the conductors. When circuits are combined in series, with the removal of intermediate grounds, the new circuit designation is shown by the equations. The test circuits k, m and t are large metallic loops from which all grounds have been removed. The horizontal and vertical displacements of the conductor by a few feet, which render the indicated equalities for $\mathcal{K}$ and n only approximate, were allowed for in determining the calculated results for Fig. 9. The grounds of the capital letter circuits were isolated sections of single track about one mile in length; the midpoint of the section is shown as the effective ground but it may have actually been displaced and have varied with the moisture of the road-bed

of 0.5 megohm per centimeter cube. The measurements were not made for the purpose of this comparison, for which they are not well adapted, but they do give both components of the mutual impedance, which is the absolutely essential requirement. The geometrical irregularity of these circuits is shown by Fig. 10. This was completely allowed for by making detailed computations after substituting an approximately equivalent broken line for each circuit. The variability in the earth's resistivity with location, depth and changing moisture content on different days could not be allowed for. The effect of buried gas and water pipes and of other grounded conductors was also necessarily neglected.

The direct-current theory leaves but one arbitrary constant at our disposal after the frequency and the geometry of the circuits have been fixed. This constant is the earth's resistivity. By trial it was found that 0.5 megohm gave a good average agreement between the calculated resistance components and the entire set of measurements, only a part of which is included in Fig. 9. The individual discrepancies are large but are not so large as to be disconcerting, considering the variations in effective earth resistivity from place to place and from day to day during the progress of the tests.

The calculated reactance component of the mutual impedance based on the direct-current mutual inductance is independent of the earth's resistivity and is uniquely determined by the frequency and geometrical relations. Even a general agreement between the calculated and the measured reactances is significant and Fig. 9 shows not only this, but also a great many good individual agreements. The outstanding discrepancies for circuits $\mathcal{C}$ and $\mathcal{E}$ are systematic, and are apparently to be explained by the effective grounding of these circuits at some other points than the midpoints of the track sections. On the basis of this comparison, it appears that the direct-current theory proves itself adequate to give an approximation to the actual mutual reactances, provided the linear scale and the frequency involved do not greatly exceed those of these tests.

Measurements were also made at 60 cycles. The resistance component remained roughly the same as for 25 cycles; the reactance component doubled as shown by Fig. 11; each component therefore agreeing approximately with the results which would obtain if the direct-current distribution is maintained.

Other comparisons have been made with the same conclusion, but tests should be made, throughout a range of frequencies, at a locality where it is known that the conductivity of the upper layer of the earth's surface is reasonably uniform so that the effect of the lower

layers may be determined. Small scale models, having the proper propagation constant, could advantageously be used in determining the alternating-current impedances for uniform earth resistivity or any assigned distribution of resistivity.

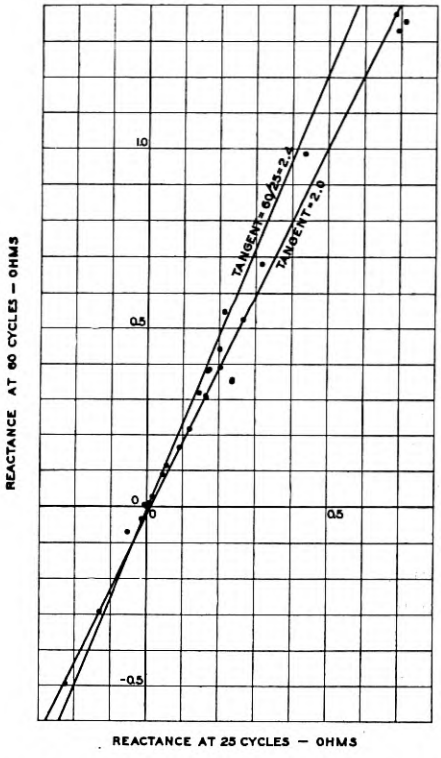


Fig. 11—Comparison of measured mutual reactances at 25 and 60 cycles. A right line with the slope 2.0 fits the point somewhat better than a line with the slope 2.4 which corresponds to the ratio of the frequencies

10. Summary

Formulas for the direct-current mutual resistances and mutual inductances for grounded circuits on the assumption of uniform earth resistivity have been derived and useful diagrams prepared. The applicability of these results as a first approximation to many practical alternating-current cases has been shown.

If, as I hope, this paper is free from ambiguities and errors, it is due to a thorough revision by Mr. R. M. Foster; and I am indebted

to Miss Frances Thorndike for the accuracy attained in the numerous curves of Figs. 1–5 and 8, which should make them of practical value in numerical calculation.

Mathematical Appendix

Additional mathematical results which have been employed in connection with the figures and discussion of this paper are brought together below for convenient reference.

Formulas for Fig. 2

$2s=\rho+d+1$, if $OA=1$, $Oa=\rho$,

$d^2=1+\rho^2-2\rho\cos\theta$,

$\sin\theta=d\sin(\theta+\phi)$,

$\sin\phi=\rho\sin(\theta+\phi)$,

$\cos\theta=(2s-1)-2s(s-1)/\rho$,

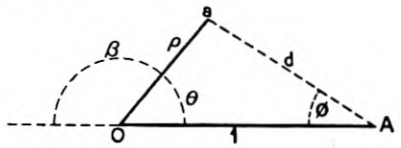


Fig. 12

$$\frac{N\mathcal{R}r}{AB}=f(\rho,\theta)=\rho f(\rho^{-1},\theta)$$

$$=\cos\theta\left(\log\frac{s}{s-\rho}+\rho\log\frac{s}{s-1}\right) \tag{20}$$

$$=\cos\theta\left\{\log\frac{\tan\frac{1}{2}(\theta+\phi)}{\tan\frac{1}{2}\theta}+\rho\log\frac{1}{\tan\frac{1}{2}\theta\tan\frac{1}{2}\phi}\right\} \tag{21}$$

$$=2\cos\theta\log\frac{2+d}{d}=2\cos\theta\log(1+1/\sin\tfrac{1}{2}\theta),\text{ if }\rho=1,$$

$$=\tfrac{1}{2}\rho\log\frac{2+\rho}{2-\rho}+\tfrac{1}{2}\rho^2\log\frac{2+\rho}{\rho},\text{ if }d=1,$$

$$=\cos\theta\left[\rho\log\frac{4}{\rho\theta^2}-(1-\rho)\log(1-\rho)\right]$$
$$+\frac{\rho(1+2\rho)}{12(1-\rho)}\theta^2+\ldots,\text{ for }\rho<1,$$

$$=\cos\beta[\rho\log\rho-(1+\rho)\log(1+\rho)]-\frac{\rho}{4(1+\rho)}\beta^2$$
$$+\frac{\rho(11+19\rho+11\rho^2)}{96(1+\rho)^3}\beta^4+\ldots,$$

$$\theta=\frac{2}{\sqrt{\rho}}e^{-[N+(1-\rho)\log(1-\rho)]/2\rho}+\ldots, \tag{22}$$

$$R=\frac{2\rho^2 \log\left[(1+\rho)/\rho\right]}{2\log(1+\rho)-\rho/(1+\rho)} = (\text{radius of curvature when } \theta=\pi) \tag{23}$$

$=0,\ 1.564,\ \infty$ at $\rho=0,\ 1,\ \infty$, which checks the sharp curvature of the curves which cross the axis just to the left of the origin.

Formulas for Figs. 4 and 5

$$s=AB/Aa,$$

$$\frac{N\mathscr{R}_r}{Aa}=2[s\log(s+\sqrt{s^2+1})+1-\sqrt{s^2+1}] \tag{24}$$

$$\left.\begin{aligned}&=2s\left[\log 2s-1+\frac{1}{s}-\frac{1}{4s^2}+\frac{1}{32s^4}-\frac{1}{96s^6}+\cdots\right]\\&=2s\log\left[\frac{2s}{e}\left(1+\frac{1}{s}+\frac{1}{4s^2}-\frac{1}{12s^3}-\frac{1}{48s^4}+\cdots\right)\right]\end{aligned}\right\} \tag{25}$$

$$=s^2\left[1-\frac{s^2}{12}+\frac{s^4}{40}-\frac{5s^6}{448}+\cdots\right], \tag{26}$$

$$Q_{\mathscr{R}r}(Aa)\left(\frac{2\pi}{\rho}\right)=2-\frac{2}{\sqrt{s^2+1}}, \tag{27}$$

$$\frac{N_1}{Aa}=\frac{N\mathscr{R}_r}{Aa}+2\log(s^2+1), \tag{28}$$

$$\frac{N_2\ (\text{or } N_4)}{Aa}=\frac{N\mathscr{R}_r}{Aa}\pm 2[\log\tfrac{1}{2}(1+\sqrt{s^2+1})-\sqrt{s^2+1}+s+1], \tag{29}$$

$$\frac{N_3}{Aa}=\frac{N\mathscr{R}_r}{Aa}+2\left[\log\frac{(s^2+1)\ (\sqrt{2}+1)^4}{(1+\sqrt{s^2+2}\,)^4}+4(\sqrt{s^2+2}-\sqrt{s^2+1}-\sqrt{2}+1)\right]. \tag{30}$$

Formulas for the Mutual Inductance Between Any Flows in Two Horizontal Planes Grounded by Vertical Filaments

Let the arbitrary flows be $\mathscr{K}'$ and $\mathscr{k}'$ between points A', B' and a', b' in the two horizontal planes, grounded by vertical filaments connecting these four terminals with the points A, B, a, b on the surface of the earth. In order to indicate briefly which of these eight points are involved in each term of the result, we imagine a vertical line which cuts the horizontal planes in the points P', p', P, p, where P and p are the same point on the surface of the earth, since the non-

primed points are all in this plane, and we agree that A' or A occurs in a term, according as P' or P is found in the subscript of the symbol Δ or Γ used to designate the term, where

$$\Delta_{P'p'} = -A'a' + A'b' + B'a' - B'b', \tag{31}$$

$$\Gamma_{P'p'} = \log \frac{(A'b' + P'p')(B'a' + P'p')}{(A'a' + P'p')(B'b' + P'p')}. \tag{32}$$

In these expressions every distance between points, such as $A'b'$, $P'p'$, is a positive quantity. The formulas below are perfectly general, but require the assignment of the capital letters $\mathscr{X}'$, A', B', P' to the upper plane when both flows are above the earth and to the lower plane when both flows are below the earth. They are most readily checked by employing formulas (4a) and (24) in formulas (17), (18) and (19). The results show that the mutual inductance is equal to the Neumann integral between $\mathscr{X}'$ and $\mathscr{x}'$ augmented by terms which depend only upon the arithmetical distances between the eight points A', B', a', b', A, B, a, b.

$$N(\mathscr{W}-\mathscr{E})(\mathscr{w}-\mathscr{e}) = N\mathscr{X}'\mathscr{x}' + P'p'\Gamma_{P'p'} + 2Pp'\Gamma_{Pp} - \Delta_{P'p'} + \Delta_{Pp}, \quad \text{where } P'p \geq Pp', \tag{33}$$

$$= N\mathscr{X}'\mathscr{x}' + 2Z \log \frac{(Ab)(Ba)}{(Aa)(Bb)}, \text{ if } P' \text{ and } p' \text{ are both at height } Z, \tag{34}$$

$$= N\mathscr{X}'\mathscr{x}' + 2Z \log (1+s^2), \text{ if } A'B'b'a' \text{ is a horizontal rectangle and } AB = s(Aa), \tag{35}$$

$$= N\mathscr{X}'\mathscr{x}' + 2Z[\log \tfrac{1}{2}(1+\sqrt{1+s^2}) + 1 - \sqrt{1+s^2} + s], \text{ if } A'B'b'a' \text{ is a vertical rectangle with one side on the earth and } AB = s(A'a) = sZ, \tag{36}$$

$$N(\mathscr{U}-\mathscr{E})(\mathscr{u}-\mathscr{e}) = N\mathscr{X}'\mathscr{x}' + P'p'\Gamma_{P'p'} - 2P'p\Gamma_{P'p} - 2Pp'(\Gamma_{Pp'} - \Gamma_{Pp}) - \Delta_{P'p'} + 2\Delta_{P'p} + 2\Delta_{Pp'} - 3\Delta_{Pp}, \quad \text{where } P'p \geq Pp', \tag{37}$$

$$= N\mathscr{X}'\mathscr{x}' - 2Z \log \frac{(Aa)(Bb)(Ab'+Z)^2(Ba'+Z)^2}{(Ab)(Ba)(Aa'+Z)^2(Bb'+Z)^2} + 4(\Delta_{P'p} - \Delta_{Pp}), \text{ if } P' \text{ and } p' \text{ are both at distance } Z \text{ below the earth}, \tag{38}$$

$$= N\mathcal{X}'_{x'} + 2(Aa)t \log \frac{(1+s^2)(t+\sqrt{1+t^2})^4}{(t+\sqrt{1+s^2+t^2})^4}$$
$$+ 8(Aa)(\sqrt{1+s^2+t^2} + 1 - \sqrt{1+s^2} - \sqrt{1+t^2}),$$ if $A'B'b'a'$ is a horizontal rectangle at the distance $Z = t(Aa)$ below the earth and $AB = s(Aa)$, (39)

$$= N\mathcal{X}'_{x'} - 2Z[\log \tfrac{1}{2}(1+\sqrt{1+s^2}) - \sqrt{1+s^2} + 1 + s],$$ if $A'B'b'a'$ is a vertical rectangle with one edge at the surface of the earth and $AB = ab = s(A'a) = sZ$, (40)

$$N(\mathcal{W}-\mathcal{E})_{(u-o)} = N\mathcal{X}'_{x'} + P'p'\Gamma_{P'p'} - 2Pp'\Gamma_{Pp'} - \Delta_{P'p'} + 2\Delta_{Pp'} - \Delta_{Pp}. \quad (41)$$

Thermionic Vacuum Tubes and Their Applications

By ROBERT W. KING

NOTE: The present material was originally prepared for the National Research Council for use in a proposed *Manual* on "Physical Research Methods and Technique." As the appearance of the *Manual* has been postponed, the Committee in charge of its preparation has kindly consented to the separate publication of some of the sections in various technical magazines. In order to meet the requirements of the *Manual*, the form of expression has been made as compact as possible with practically no discussion of theory and no derivation of formulas. Since this style of presentation leaves much to be desired from some points of view, references have been given to the original literature wherever possible. However, many of the vacuum tube circuits presented have not as yet been treated in the literature. In the preparation of the new material the author has been greatly helped by persons whose contact with these subjects is at first hand.

I. INTRODUCTION

1. *Thermionic Emission.* By thermionic vacuum tubes we shall understand those whose operation depends in an essential manner upon thermionic emission.

The design of the various types of thermionic tubes at present in use requires no knowledge of the exact mechanism of thermionic emission. It may be said, however, that the work of O. W. Richardson and others leaves little question but that this emission is a physical as distinguished from a chemical process, and occurs from certain substances as the result of the large velocities of thermal agitation acquired by electrons when these substances are raised to a high temperature.

On the basis of certain plausible assumptions, O. W. Richardson derived [2] the expression,

$$I_s = Ne = A T^{\lambda} \epsilon^{-\frac{w_o}{kT}}, \qquad (1)$$

for the thermionic emission per cm.^{2} in which I_s is the saturation current formed by drawing all the emitted electrons to a positively charged electrode placed near the emitting surface, e is the electronic charge and ϵ the Naperian base, A is a constant dependent on the emitting substance but independent of the absolute temperature T,

[2] Richardson, The Electron Theory of Matter, 1916 Edition, page 441.

w_o represents the energy lost by each electron as a result of becoming free, λ is a number which does not differ much from unity, and k is the gas constant per molecule. Experiments show that the value of λ for a wide range of substances is about unity, but its exact value is of little practical importance, since the variation of I_s with T is almost entirely controlled by the term in which T enters as an exponent.

The quantity w_o which expresses the *electron affinity* of the emitting substance is usually called the internal work of evaporation. In Equation 1, it is in terms of ergs per electron. Calling v_o the value of w_o when expressed in equivalent volts, $w_o = 1.59 \times 10^{-12}\, v_o$.

The term v_o is of great importance when considering the economy with which a substance acts as a thermionic emitter. Assuming that the emission of an electron occurs when its velocity acquires a value sufficiently high to overcome the potential drop v_o, it is apparent that the smaller v_o, the more copious will be the thermionic emission at any given temperature. For the substances thus far examined, v_o ranges between about two volts and five volts.[3] The two substances whose thermionic properties we shall consider particularly are platinum, coated with a mixture of barium and strontium oxides, and tungsten. For tungsten the value of v_o is approximately 5 volts, and for coated platinum it varies between 1.67 and 2.05 volts. The value of $(v_o)_A - (v_o)_B$ for two substances A and B is equal (except for a small term expressing the Peltier coefficient) to their contact difference of potential.[4]

2. *Thermionic Properties of Filaments.* In designing electron tubes with predetermined characteristics knowledge of the thermionic emissivity of the proposed filament is necessary. This property may be conveniently represented by curves of the type shown in Figs. 1, 2 and 3. The coordinates have been so proportioned [5] that, provided the electronic emission varies with the temperature as indicated by Equation 1 and the thermal radiation from the filament varies as the fourth power of the temperature, then the relation between the thermionic emission and the heating power supplied to the filament is a straight line.

Fig. 1 gives data for tungsten and for coated platinum filaments, Fig. 2 compares thoriated tungsten filament with pure tungsten and Fig. 3 gives data relating to a special coated filament, the core of which consists mainly of platinum alloyed with about 5% of nickel. Since

[3] For a Table of Values of v_o for the materials commonly used see Van der Bijl—Thermionic Vacuum Tube, page 29; also Langmuir, Trans. Am. Electro-chem. Soc. 29,166, 1916.

[4] Richardson, Electron Theory of Matter, 1916, p. 455.

[5] See Van der Bijl, The Thermionic Vacuum Tube, p. 82.

the thermionic emission of tungsten, and thoriated tungsten when freshly activated, do not vary much between different samples, they are given with sufficient accuracy for tube design purposes by single

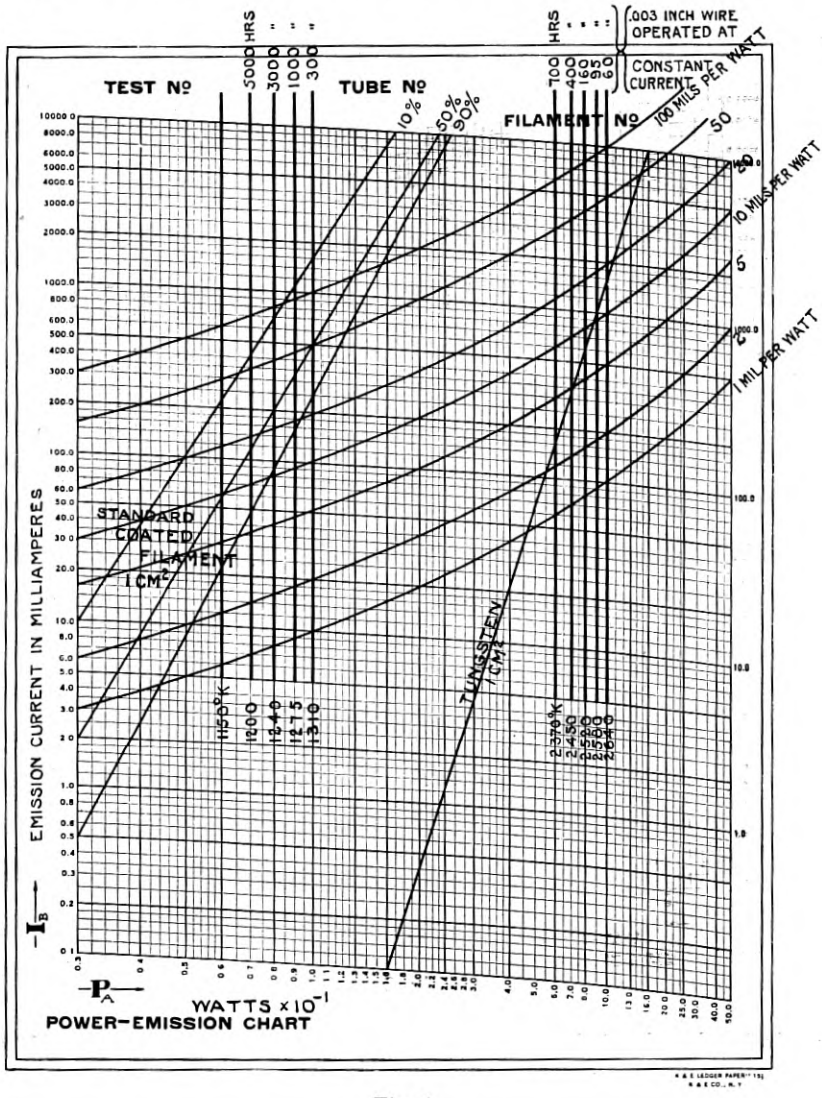


Fig. 1

lines. Coated filaments, however, show a rather wide variation as is indicated. The range of variation shown in Fig. 1 was obtained in the study of several thousand tubes; 10 per cent. showed activity

greater than that represented by the 10% line, 50 per cent. showed activity greater than the 50% line and 90 per cent. showed activity greater than the 90% line. The three lines illustrating the range of

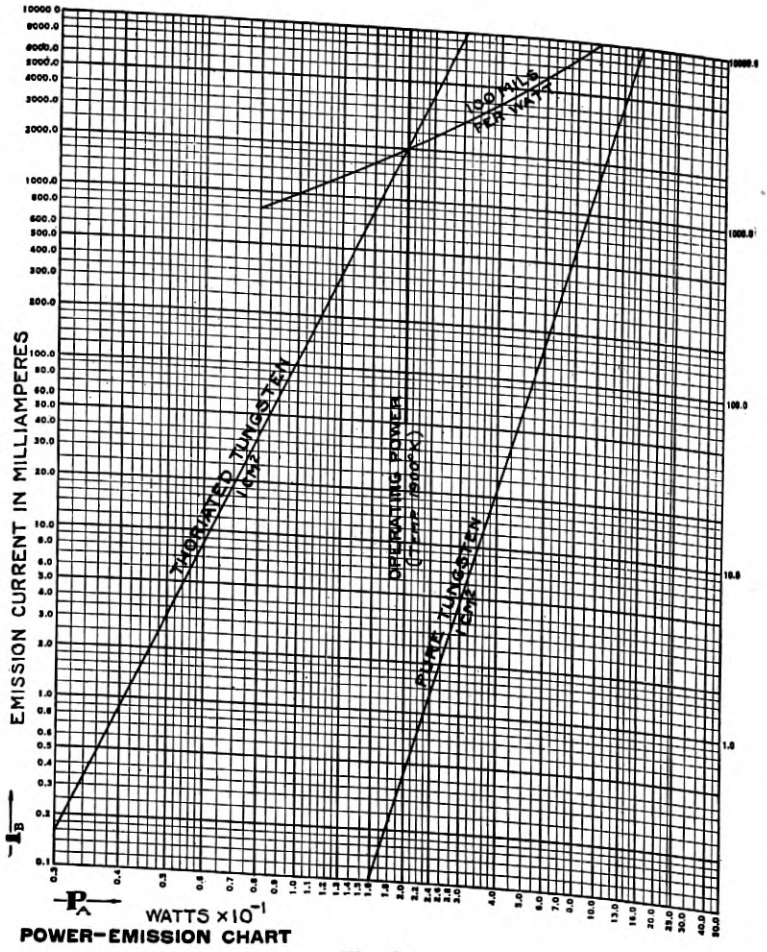


Fig. 2

variation in Fig. 3 do not correspond to these same percentages but have been labeled so as to be readily interpretable.

The efficiencies in milliamperes of emission current per watt used to heat the filament are shown by the curved lines that cross each chart. Operating temperatures and corresponding filament life are given for certain of the ordinates. The tungsten life data in Fig. 1 are for 3-mil wire and a constant heating current. If operating at a con-

stant temperature, the life would be somewhat longer; furthermore, the larger the wire the longer the life for a given temperature. The activity of thoriated filaments tends to diminish with use but may

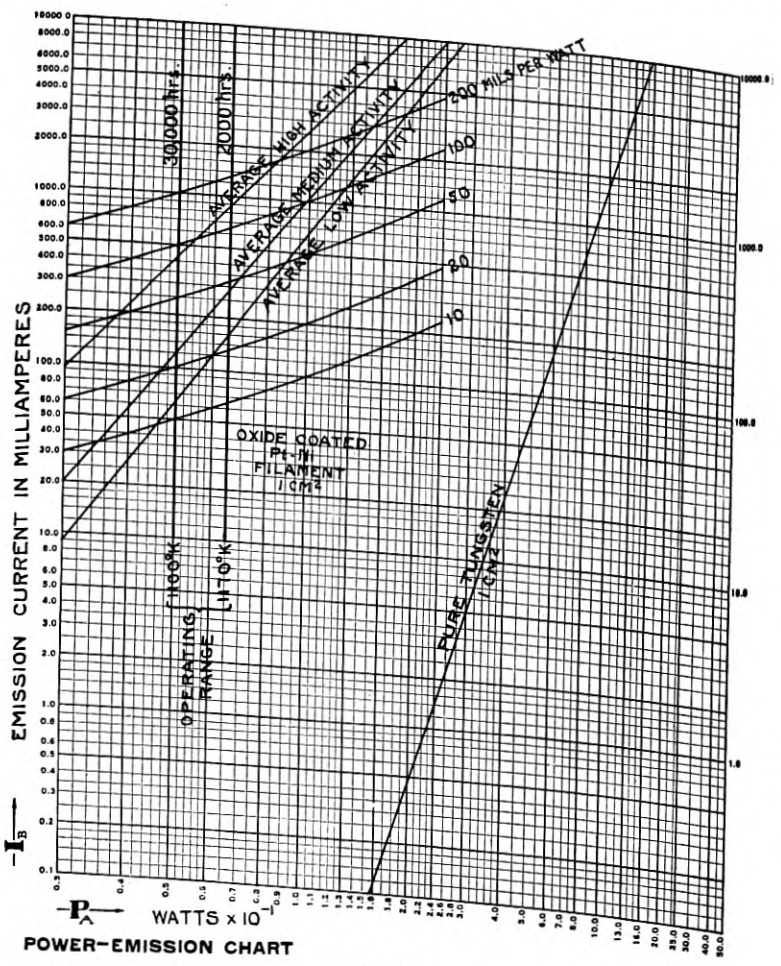


Fig. 3

be re-activated by heating to a temperature higher than the normal operating temperature. The useful life of these filaments is, therefore, somewhat indefinite but taking the possibility of re-activation into account, it is probably well in excess of 2,000 hours.

In using Figs. 1, 2, 3, it should be borne in mind that the emission values given represent saturation currents and that in general the normal operating space current in a tube must, for reasons which will

appear later, be appreciably less than the saturation current. The difference between the saturation current and the maximum operating space current varies with the duty to which the tube is assigned. In the case of high voltage rectifiers, the space current may at certain points in the cycle reach the saturation value, while in a tube which is used as an amplifier it is often desirable, in order to avoid distortion, to have the total emission two to three times as great as the maximum working space current.

3. *Space Current-Voltage Characteristic.* Experiment shows that in a vacuum tube containing an emitting electrode and a conveniently placed anode, the space current I_p, varies with the temperature of the emitter and the anode potential E_p, as in Fig. 4. The three curves shown are for three temperatures such that $T_1 < T_2 < T_3$. It will be observed that between points O and A the three curves coincide;

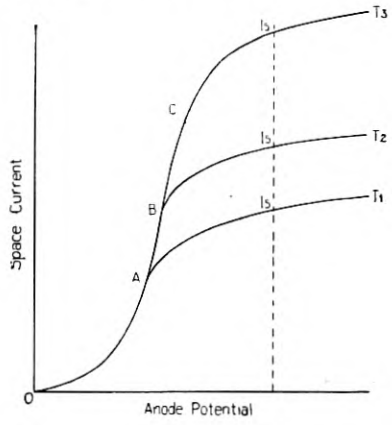


Fig. 4

between O and B the curves for the two higher temperatures coincide. The saturation values of the filament emission at the various temperatures are shown by I_s.

For values of I_p ranging from zero to somewhat less than I_s, the relation between I_p and E_p may be expressed with a fair degree of accuracy by $I_p = \kappa E_p^{\eta}$ in which the exponent η does not differ greatly from 3/2. This relation, therefore, is frequently known as the 3/2 power law. It has been deduced theoretically by Child [7] and Langmuir [8] and has been studied lately in greater detail by Fry.[9] Fry's analysis takes account of the initial velocities of emission of electrons,

[7] *Phys. Rev.*, Vol. 32, p. 498, 1911.
[8] *Phys. Rev.* (2), Vol. 2, p. 450, 1913.
[9] T. C. Fry, *Phys. Rev.*, Vol. 17, p. 441, 1921.

and, as he shows, the effect of the space charge is to create a region of negative potential immediately around the emitter. Let Fig. 5 represent the value of the potential as one proceeds from the cathode in the direction x, and V' represent the minimum value of the potential. Assuming indefinitely large emission from the cathode, V' (which is a function of E_p), determines the space current corresponding

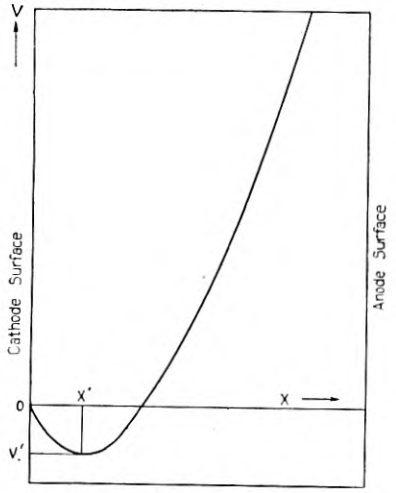


Fig. 5

to any particular value of E_p. The lower the value of V' the fewer the electrons with initial velocities sufficient to carry them past the equipotential surface V' into the region where they are attracted by the anode. Assuming the average initial velocity to be 0.3 volts (roughly a temperature of 2400° K), Fry finds an appreciable deviation from the 3/2 power law for $E_p < 30$ volts, but initial velocities need be considered only in tubes which operate at low E_p.

Another factor which, for low E_p causes an appreciable deviation from the 3/2 power law, is the potential gradient in a filament cathode due to the heating current. Whereas velocity of emission tends to make $\eta < 3/2$, the potential gradient in the filament has the reverse effect. In general, the latter more than overbalances the former and for small anode voltages $\eta > 3/2$. The value of η when the cathode potential gradient is considered, but initial velocities are neglected, has been given by W. Wilson,[10] who finds that when E_p is less than the potential drop across the filament $\eta = 5/2$, while for higher E_p it gradually approaches the limiting value 3/2.

[10] For discussion of the 5/2 power relation, see Van der Bijl, The Thermionic Vacuum Tube, p. 64.

4. *Temperature Saturation and Voltage Saturation.* When a vacuum tube operates at such filament temperature and E_p that an increase in temperature produces no increase in I_p the tube is said to show *temperature saturation.* On the other hand when temperature and E_p are such that an increase in E_p does not increase I_p, the tube shows *voltage saturation.* Referring to Fig. 2 the curve T_1 shows temperature saturation between O and A and approaches voltage saturation beyond the point A. Similarly the curve T_2 shows temperature saturation up to the point B and approaches voltage saturation beyond.

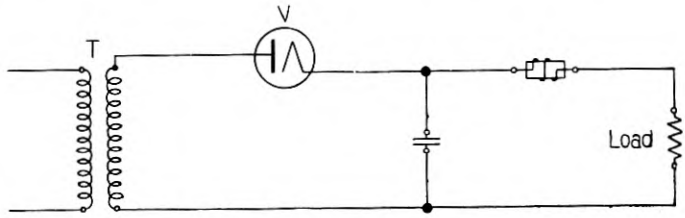


Fig. 6

5. *Effect of Gas.* In thermionic tubes as usually pumped the gas pressure is between 10^{-5} and 10^{-6} mm. At this pressure the gas generally does not manifest its presence in the operation of the tube. However, at higher pressures, and particularly above 10^{-4}mm, it produces certain objectionable disturbances. Thus many gases seriously reduce the filament activity; also for E_p greater than about 20 volts, ionization occurs and the resulting discharge differs in important respects [11] from the pure electron discharge of Fig. 4.

II. Two-Electrode Tubes

The two-electrode tube, which was first due to Edison, found an early practical application when Fleming used it to detect wireless telegraph signals.

However, since the advent of the three-electrode tube of De Forest, the earlier device has been almost entirely superseded as a detector and finds its principal application as a rectifier of a.c. voltages. Its range of applicability in this field is extremely large. With properly designed tubes, Hull [12] has succeeded in rectifying 5 k.w. at a potential of 100,000 volts, and in its transatlantic radio telephone experiments,

[11] See Van der Bijl: The Thermionic Vacuum Tube, p. 86.

[12] A. W. Hull, *General Electric Review*, Vol. 19, p. 177, 1916. Another good source of information is Van der Bijl's "The Thermionic Vacuum Tube."

the American Telephone and Telegraph Company is using a six phase rectifier giving about 200 kw. at about 10,000 volts.[13]

6. *Two-Electrode Tube as Rectifier.* Three typical forms of circuit are shown in Figs. 6, 7, 8, each of which has certain characteristics not possessed by the others. The circuit shown in Fig. 6 rectifies the full transformer voltage, but utilizes only one-half of the current wave; the circuit of Fig. 7 gives a d.c. voltage of only about half the transformer peak voltage but utilizes both halves of the wave, and Fig. 8 illustrates a circuit making use of the full transformer voltage and both halves of the a.c. wave.

A rectifier circuit employing a tuning condenser for the secondary of the high voltage transformer and giving a rectified d.c. voltage as large again as the a.c. voltage of the transformer and providing automatic control of the maximum d.c. voltage supplied by the rectifier, has been described by Webster.[14]

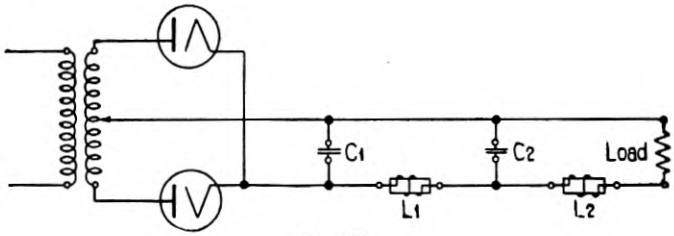


Fig. 7

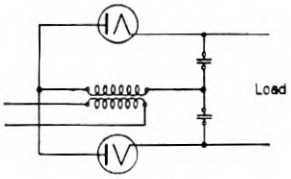


Fig. 8

The circuits shown in Figs. 6 and 7 provide means in the form of condensers C_1, C_2 and inductances L_1, L_2 for smoothing out the rectified voltage. Such an arrangement of conductors is essentially a network whose attenuation for electric currents of a certain range of frequencies is very low and for all other frequencies is high. This type of network is a special form of the *electric wave-filter* which is finding many ap-

[13] Arnold and Espenschied, *Journal of A. I. E. E.*, August, 1923; also *Bell System Technical Journal*, October, 1923.

[14] D. L. Webster, Procd. Nat. Acad., Vol. 6, p. 28, p. 269, 1920. These articles also suggest certain modifications of Hull's method.

plications at the present time and a complete account of its properties is to be found in recent literature.[15]

No definite statement can be made as to the exact range of frequencies over which the rectification of alternating currents can be satisfactorily accomplished by means of thermionic tubes, but it is apparent that this range is large. The degree of smoothness required in the d.c. output is of primary importance in setting the lower limit of the frequency range; on the other hand, the smaller the load resistance, the higher the frequency which may be satisfactorily rectified before the internal capacity of the tubes permits the flow of an objectionable amount of alternating current. Whenever an output with a minimum of ripple is required it is in general desirable to use as high a frequency as is readily available.

III. Three-Electrode Tubes

In 1906 De Forest brought out the three-electrode tube [16] in which a grid is interposed between filament and plate. Since the introduction of this tube, much study has been devoted to its properties and many investigations have been made concerning the best substances to use as thermionic emitters, the best metals for plates and grids, the best varieties of glass for the containing bulbs,[17] and the best methods of exhaustion,[18] so that today problems of design are well understood. At present the three electrode tube finds use as a rectifier, amplifier of small currents and voltages, detector of small a.c. voltages, modulator of alternating currents, and generator of electric oscillations. Tubes have been built which range in size from those about one inch long with a space current of about a milliampere to others which are water-cooled and have an individual output capacity of 100 k.w.[19] Amplifiers with a capacity of 150 k.w. are in operation (see footnote 13).

7. *Action of the Grid.* The general theory of the grid action is simple. As pointed out by Fry [19] the space charge creates a region of

[15] G. A. Campbell, *Bell System Technical Journal*, Nov., 1922; U. S. Patents 1,237,113 and 1,237,114, May 22, 1917; O. J. Zobel, *Bell System Technical Journal*, Jan., 1923; Carson and Zobel, *Bell System Technical Journal*, July, 1923; G. W. Pierce, Electric Oscillations and Electric Waves, p. 186, 1920; Karl W. Wagner, *Arch. f. Electr.*, Vol. 3, p. 315, 1915; Vol. 8, p. 61, 1919.

[16] Variously called the audion, vacuum tube, triode, pliotron, etc.

[17] Measurements of Gases Evolved from Glasses of Known Chemical Composition—Harris & Schumacher, *Jour. Ind. & Eng. Chem.*, Feb., 1923; also *Bell System Technical Journal*, Jan., 1923.

[18] For methods of exhaustion, see Dushman, *Gen. Elect. Review*, Vol. 23, p. 493, 1920, et seq.

[19] See W. Wilson, *Bell System Technical Journal*, July, 1922.

[19] T. C. Fry, l. c.

negative potential immediately around the cathode which persists for all values of E_p less than that required to produce voltage saturation. It is the minimum potential V' (see Fig. 5) that limits I_p, and any increase in V' will result in an increase in I_p. Because the grid is close to the filament, small changes in the grid potential, E_g, are as effective in changing V' and therefore I_p, as large changes in E_p. This leads to the so-called amplification constant μ of the tube which may be taken as

$$\mu = \frac{\pm e_p}{\mp e_g},$$

in which $\pm e_p$ and $\mp e_g$ are changes in E_p and E_g, the changes being opposite in sign as indicated, and such that they leave I_p unchanged.

It has also been shown[20] that if ΔE_p and ΔE_g are increments of E_p and E_g which cause equal increments in the electric field at the surface of the cathode (considered simply as an equipotential surface and not as a source of electrons) the amplification constant, μ, of the tube will be the ratio $\Delta E_p/\Delta E_g$.

8. *Characteristic Equation.* The electrical characteristics of the three-electrode vacuum tube may be represented [21] by the equation

$$I_p = \kappa\left(\frac{E_p}{\mu} + E_g + \epsilon\right)^{\eta}. \tag{2}$$

The constant κ is related in a simple way to the internal resistance of the tube. A consideration of ϵ which expresses the contact difference of potential between grid and filament is usually essential only in tubes which operate at a low E_p and particularly in detectors and amplifiers. In tubes with coated filament, ϵ may not only vary within a range of two or three volts between different tubes, but may also change during the life of any one tube. The exponent η varies in a given tube with applied voltage, being usually equal to about 2 when the effective voltage $\left(\frac{E_p}{\mu} + E_g + \epsilon\right)$ is comparable with the potential drop in the filament (see Fig. 9), and tending to approach the theoretical value 3/2 with increasing effective voltage. It has been found possible to deduce relations between the constants μ and κ of Equation 2 and the structure and dimensions of any tube which are in very fair agreement with experimental values.[22]

Typical curves corresponding to the characteristic Equation 2 are shown in Figs. 10 and 11. These curves are referred to as *static* char-

[20] R. W. King, *Phys. Rev.*, Vol. 15, p. 256, 1920.
[21] Van der Bijl, *Phys. Rev.*, 12, 180, 1918.
[22] King, l. c.

acteristics one parameter being fixed in each case. For the *dynamic* characteristic see section 12. Equation 2, of course, fits only the portions of the curves characterized by temperature saturation.

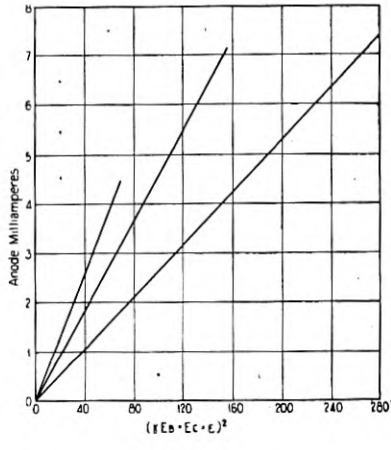


Fig. 9

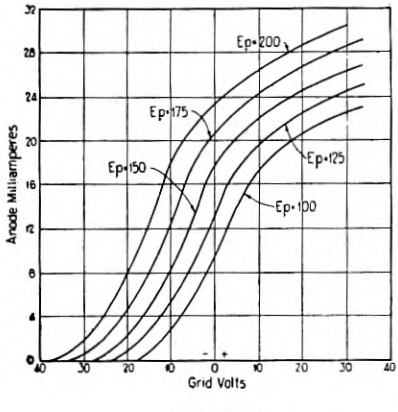


Fig. 10

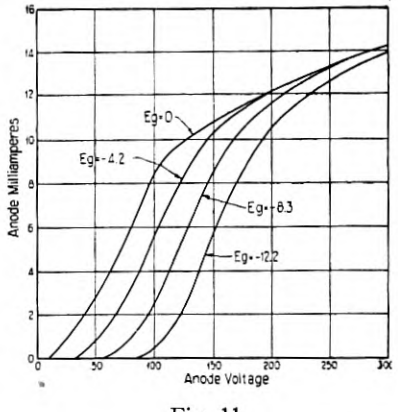


Fig. 11

The abruptness with which the curves of Fig. 10 meet the potential axis is important in certain uses to which tubes are put. The value of E_g which reduces I_p to zero is called the *cutoff voltage*. To have a sharp cutoff a tube should have a fairly large μ and its grid should be sufficiently large with respect to the filament to effectively screen all parts of the filament from the plate.

9. *Grid Current.* For certain purposes, a consideration of the grid current I_g, is necessary. Fig. 12 represents a characteristic relation between I_g and E_g for various E_p. Note that E_g in excess of about 10 volts results in *secondary emission* of electrons from the grid. These secondary electrons flow to the plate and, as shown, their number may actually exceed the total number of primary electrons striking the grid. The character of the grid surface plays an important part

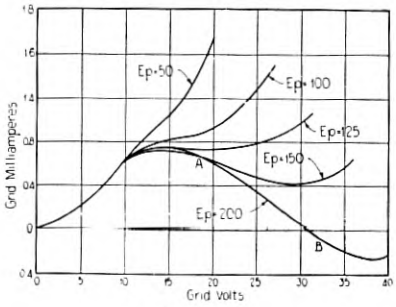


Fig. 12

in determining the amount of secondary emission. The secondary emission from the grid of a tube containing a pure tungsten filament is, in general, less than that from the grid of a tube with an oxide-coated filament. At high temperature a coated filament appears to evaporate a minute amount of its coating, some of which is deposited upon the grid presumably augmenting the secondary activity.[23]

10. *Vacuum Tube Constants.* The two most important constants of the three electrode tube are μ and its internal resistance r_p. The determination of μ and r_p from the characteristic curves (Figs. 10, 11) is obvious. For general design purposes these curves give the best insight into the behavior of a tube and furnish the most instructive means of determining μ and r_p.

11. *Dynamic Methods of Measuring Vacuum Tube Constants.* However, in cases where many tubes, all practically alike, have to be tested, certain "dynamic" methods are timesavers. Several such methods have been devised, but all are modifications of a scheme first published by Miller.[24]

[23] A. W. Hull has designed two types of tube known as the dynatron and pliodynatron which utilize the negative resistance characteristic (AB of Fig. 11) resulting from secondary emission. Procd. Inst. Radio Engrs., Vol. 6, p. 5, 1918.

[24] J. M. Miller, Proc. I. R. E., Vol. 6, p. 141, 1918. For variations of Miller's dynamic method the reader is referred to Van der Bijl, l. c., p. 198, Method of G. H. Stevenson.

Miller's method is illustrated in Fig. 13. To determine μ the key K_1 is open and K_2 is closed. The resistance r_1 is adjusted until the sound heard in the telephones T is a minimum, under which circumstances it is clear that $\mu=\frac{r_1}{r_2}$. To determine r_p, some definite relation between r_1 and r_2 is established. Then, with key K_1 closed, the resistance r_o is adjusted until the telephone response is a minimum. With this adjustment it may be shown that

$$r_p = r_o\left(\mu\frac{r_2}{r_1} - 1\right). \tag{3}$$

This measurement of r_p may be simplified as follows: suppose we adjust r_1 for a minimum tone in T when K_1 is open. Then $\mu=\frac{r_1}{r_2}$, and it is seen from Equation 3 that with this relation between r_1 and r_2 it would not be possible to obtain a balance with K_1 closed; but if

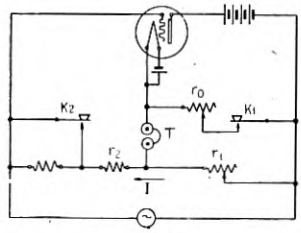


Fig. 13

r_2 be doubled, which can be done by opening K_2 thus adding a resistance equal to r_2, and r_o be now adjusted with K_1 closed to give a minimum tone in T, then $r_p=r_o$.

12. *Dynamic Characteristics of Vacuum Tubes.* In a circuit such as that shown in Fig. 13, the space current causes a fall of potential along any resistance r, and the difference in potential between filament and plate is therefore less than the potential difference across the battery by the amount $I_p r$. If I_p is increased in any way, as for instance, by an increase in E_g, the drop $I_p r$ increases and with a fixed battery e.m.f. the potential difference between the filament and plate diminishes somewhat. It follows, therefore, that a given change in E_g will cause a smaller change in space current when the plate circuit includes an external resistance r than when it does not.

This important fact supplies a simple means of straightening the characteristic of a vacuum tube to such an extent that it may become practically a distortionless amplifier.

To a first approximation,[26] the alternating component J of the space current, when a voltage $e = e_o \cos pt$ is applied to the grid is given by

$$J = \frac{\mu}{r + r_p} e_o \cos pt - \frac{r_p r_p' e_o^2}{2!(r + r_p)^3}(1 + \cos 2pt). \qquad (4)$$

In this equation r_p is the internal resistance of the tube and r_p' is its derivative with respect to the effective voltage $\left(\frac{E_p}{\mu} + E_g\right)$. It will be noted that the second term on the right side of Equation 4, which gives the first harmonic, diminishes rapidly with r as was to be expected from the preceding paragraph. The more or less straightened

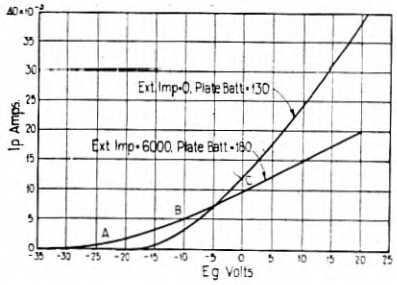


Fig. 14—Average I_p, E_g Characteristics for Five Tubes. Average $\mu = 5.92$; average internal impedance = 6,000 ohms for $E_b = 130$ volts and $E_g = -9$ volts. Plate battery connected to +end of filament and grid battery to −end. Western Electric type 101-B tubes.

characteristic resulting from the effect of r is known as the *dynamic* characteristic; see Fig. 14, which curves fit a tube whose $r = 6{,}000$ ohms. As will be pointed out in the section on amplifier circuits, the dynamic characteristic is a useful guide in selecting tubes as amplifiers.

Equation 4 also expresses the important fact that the application of the voltage e_o to the grid is equivalent, so far as current in the plate circuit is concerned, to the application of the voltage μe_o in the plate circuit.

13. *Internal Capacities and Effect on Input Impedances.* In certain uses to which the vacuum tube may be put, a knowledge of the influence which the internal electrostatic capacities have on the input impedance is important. The equivalent circuit of the tube[27] is

[26] J. R. Carson, Proc. Inst. Radio Engrs., Vol. 7, page, 187, 1919. In case the output circuit of a tube contains reactance as well as resistance, Eq. 4 becomes much more complicated as Carson shows.

[27] H. W. Nichols, *Phys. Rev.*, Vol. 13, p. 405, 1919; J. M. Miller, Bureau of Standards, Bulletin No. 351.

shown in Fig. 15 in which C_1 is the capacity between filament and grid, C_2 capacity between filament and plate, C_3 capacity between grid and plate, and r_1, r_3, are leakage resistances. As the action of the tube is such as to produce an equivalent voltage μe_o between filament and

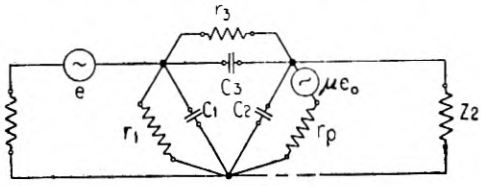


Fig. 15

plate, a generator of voltage μe_o is shown in series with the internal resistance r_p of the tube. Calling Y_g the input admittance of the tube, that is,

$$\frac{1}{Y_g} = Z_g = \frac{e_g}{i_1 + i_3},$$

in which e_g is the alternating voltage between filament and grid, the solution of the above circuit gives for Y_g the value,

$$Y_g = \frac{1}{r_1} + jC_1p + \frac{\left(\frac{1}{r_3} + jC_3p\right)[jC_2pr_pZ_2 + r_p + Z_2(\mu+1)]}{(jC_2pr_pZ_2 + r_p + Z_2) + \left(\frac{1}{r_3} + jC_3p\right)r_pZ_2} \qquad (5)$$

Case 1, *Low Frequencies.* For low frequencies the admittance of the condenser C_2 is negligible compared with that of r_p. Dropping the terms containing C_2 gives the equation,

$$Y_g = \frac{1}{r_1} + jC_1p + \left(\frac{1}{r_3} + jC_3p\right)\frac{r_p + Z_2(\mu+1)}{(r_p + Z_2) + r_pZ_2\left(\frac{1}{r_3} + jC_3p\right)},$$

which yields important interpretations. In case the load impedance Z_2 is a pure resistance r_2, the admittance of the filament-grid branch of the tube may be much greater than the admittance which would result from R_1 and C_1 alone. This is due to the influence which the alternating component of the plate voltage exerts upon the input circuit through the condenser C_2. Figs. 16 and 17 show respectively the effective capacity and effective conductance between filament and grid as a function of the external resistance. For the particular tube studied (W. E. Co. 102-A) Fig. 17 shows that, if $r_2 = 40{,}000$ ohms, the effective capacity between filament and grid is not the capacity C_1

(about 10×10^{-12} farads) but is approximately 120×10^{-12} farads. Fig. 16 shows that the effective conductance is also greatly increased above that due to r_1. This increase in conductance means an increased absorption of input energy by the tube which, of course, is

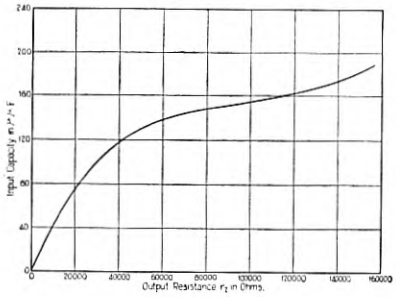


Fig. 16

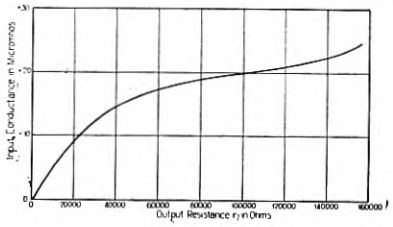


Fig. 17

not dissipated in the grid circuit but passes through the path supplied by the condenser C_3 to be wasted in the plate circuit.

In case Z_2 is a pure inductance L_2, the effective input conductance of the tube is negative and not positive as in the preceding case.

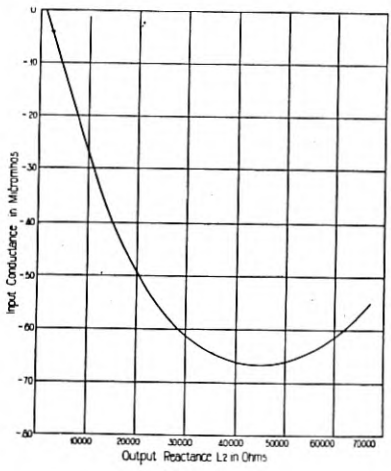


Fig. 18

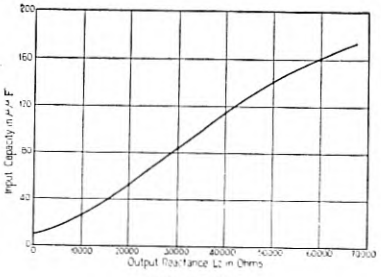


Fig. 19

Fig. 18 shows the variation of this negative conductance with L_2, and Fig. 19 the variation of effective input capacity with L_2. A negative input conductance means that the grid circuit draws power from the plate circuit; if the negative conductance is large enough, a tube in

such a circuit will oscillate steadily or "sing" with no coupling but that provided by its internal capacities. This phenomenon is frequently encountered in vacuum tube amplifiers and at times proves quite troublesome.

Tubes can readily be constructed in which r_1 and r_3 are so large as to exert no influence on the behavior of the tube and may be ignored in the above equations. However, even in such tubes there is an effective input conductance, either positive or negative, depending upon the character [28] of Z_2.

Case 2, *High Frequencies.* For very high frequencies terms of the first order in p are negligible compared to terms of the second order, and Eq. 5, becomes,

$$Y_g = \frac{p(C_1C_2 + C_1C_3 + C_2C_3)}{C_2 + C_3},$$

indicating that as the frequency is raised the effective input impedance approaches that due to the condensers alone. Under these circumstances the grid absorbs very little power, but the amplification is lowered because the input is to an extent short-circuited by the electrode capacities. Fig. 20 shows the variation in voltage am-

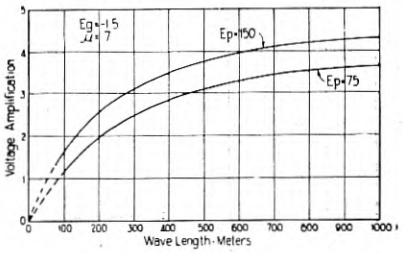


Fig. 20

plification against wave lengths in meters for high frequencies. The two curves are for different E_p, the higher E_p giving a larger amplification because r_p of the tube is lower. It is seen that the amplification at 1,000 meters is about three times as large as the amplification at 100 meters and the amplification at both values of E_p tends to approach zero as the frequency becomes infinite.

Nichols suggests [29] that the reduction in amplification for a given frequency can be avoided by shunting the grid-plate capacity C_3 with

[28] For cases in which Z_2 is neither pure resistance nor reactance, see Van der Bijl, l. c., p. 210.

[29] H. W. Nichols, *Phys. Rev.*, Vol. 13, p. 411, 1919.

an inductance of such a value as to make the impedance between grid and plate infinite at this frequency.

IV. Thermionic Amplifiers

Equation 4, given in Sec. 12 is of fundamental importance in the design of vacuum tube amplifier circuits. Neglecting the second term on the right hand side which, as previously pointed out, expresses distortional effects and should therefore be very small in amplifier circuits, the equation can be written

$$J=\frac{\mu e_o}{Z+r_p}, \tag{6}$$

in which the impedance $Z=r+j\,x$ is substituted for r. From Equation 6 both the voltage and power amplification of a tube for any particular circuit can readily be calculated.

14. *Voltage Amplification.* Assuming as above that the tube works into an output impedance Z, it follows that the voltage amplification (i.e., the ratio of the output to the input voltage) is

$$\frac{\mu Z}{Z+r_p}.$$

This expression shows that the voltage amplification increases as Z increases. Considering separately the two cases in which Z is a pure resistance and pure reactance, typical values of the voltage amplification are plotted in Fig. 21. Curve *a* corresponds to reactance

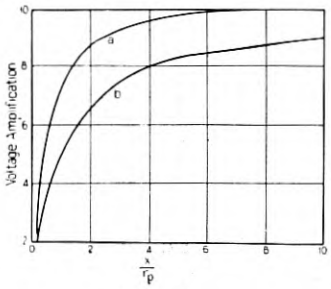


Fig. 21

in the output and *b* to resistance. These curves show that the voltage amplification rises much more rapidly when reactance is used, reaching 90% of its maximum value when $\frac{x}{r_p}=2$.

If the resistance component of Z is made as small as possible, E_p becomes practically equal to the potential across the plate battery,

making possible a given voltage amplification with smaller plate battery than could be obtained if the output circuit contained an appreciable resistance.

15. *Power Amplification.* From Equation 6, which neglects harmonic terms, it is readily seen that the power output is

$$\frac{\mu^2 e_o^2 r}{(r+r_p)^2+x^2},$$

where $r+j\,x$ has been substituted for Z. This output is a maximum when $r^2=r_p^2+x^2$. The case in which $x=0$ is particularly important; evidently for maximum power output, a tube should work into a resistance equal to its internal resistance.

As pointed out in Sec. 13 the input impedance of a tube is not always readily determinable; however, calling the input resistance [30] r_g, the power amplification produced by a tube is given by the expression,

$$\frac{\mu^2 r\, r_g}{(r+r_p)^2+x^2}. \tag{7}$$

This expression has been obtained on the assumption that the grid draws no electron current from the space charge, which in turn requires that the grid remain at a negative potential at all times. Since the power amplification falls rapidly as the grid becomes positive, it is customary in most amplifier circuits to supply means of maintaining the grid at a negative potential.

Expressions 6 and 7 are of fundamental importance in the design of amplifier circuits.

16. *Selection of Tubes.* When selecting tubes for an amplifier, curves such as those shown in Fig. 14 are very useful. By their means it is readily possible to select the tube, the plate potential and the average grid potential which will give satisfactory results for any pre-assigned value of the input voltage. In order to obtain amplification as free from distortion as possible, it is necessary that the grid potential in its excursions neither become positive nor strike the lower end of the characteristic. To a sufficient approximation it is evident that when the variable grid or input voltage e_o is given, we should choose E_g and E_p such that

$$e_o \gtrless -E_g \gtrless \frac{E_p}{2\mu}.$$

[30] Where many tubes of the same design are to be interchanged in a given circuit, and where the conditions of manufacture are such that the insulation resistance between filament and grid is not always of the best, it may be found desirable to shunt the input with a fixed resistance, e.g., ½ megohm.

Referring to Fig. 14, for an input potential of 10 volts (peak value) and $r_p = 6{,}000$ ohms, the point B (but neither A nor C) is evidently a satisfactory mean position about which to operate.

Since both voltage and power amplification increase as E_p is increased (because r_p is decreased) it is frequently desirable to have the value of E_p considerably larger than the lower limit just indicated.

17. *Amplifier Circuits.* The fundamentals of thermionic amplifier circuits may be gathered from Fig. 22. The variable input voltage

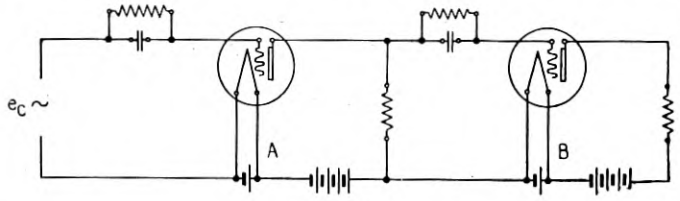


Fig. 22

e_o between grid and filament of the first tube A modulates the plate current of this tube. The circuit is evidently such that the variable I_p of tube A varies the grid potential of B and, due to the properties of the three electrode tube, not only is the power applied to the grid of B greater than that applied to A, but the potential variation may be many times as large as e_o. Hence the variations in I_p of B will be larger than those of I_p in A, thus yielding an amplifying action. Tubes A and B and their associated circuits are known as the first and second stage respectively.

Amplifier circuits may for convenience be divided into six general classes:

1. Resistance coupled circuits (Fig. 23).
2. Resistance-condenser coupled circuits (Fig. 24).
3. Retard-condenser coupled circuits (Fig. 25).
4. Transformer coupled circuits (Fig. 26).
5. Feed-back circuits (Fig. 31).
6. Push-pull circuits (Fig. 32).

The Sections immediately following will point out the advantages of each type of circuit and general design considerations. Equations 6 and 7 show that the amplification of which a single tube (or "stage") is capable is definitely limited. For greater amplification than a single stage can produce, it is necessary to arrange two or more stages in cascade. Multistage amplifiers frequently consist of combinations of certain of the above types of circuits as will be pointed out in the following paragraphs.

18. *Resistance Coupled Amplifier.* This simplest of all amplifier circuits (Fig. 23) is particularly useful where a wide range of frequencies is to be amplified without selective amplification of any particular frequencies. For this reason, it is often used as an amplifier in con-

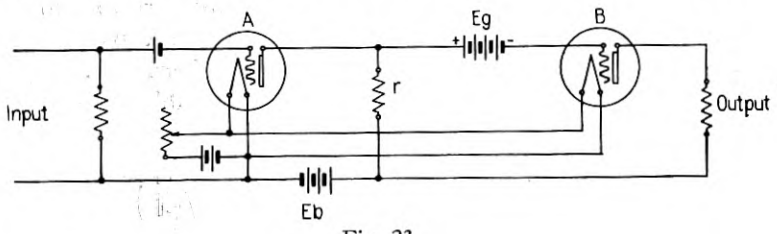


Fig. 23

nection with an oscillograph. It is also one of the few types of circuit which can be used for direct current amplification. However, as will be pointed out later, a special type of push-pull circuit makes a more satisfactory d.c. amplifier for many purposes.

One or more stages of the resistance coupled circuit may be substituted for transformers in voltage amplification. Voltage amplifier tubes having an amplification constant $\mu = 30$ are common and it follows from Equation 6 that such a tube can readily produce a voltage amplification of from 20 to 25. It can, therefore, take the place of an input transformer in one of the other types of amplifier circuit. Unless special considerations require another adjustment, it is customary to arrange all but the last stage of an amplifier for voltage amplification, the last stage being designed for power amplification. (See Secs. 14 and 15.)

Since in resistance coupled amplifiers there is a d.c. path between the plate of one tube and the grid of the following tube, a negative grid battery large enough to counterbalance the plate battery must be used in every stage in order to supply the necessary negative grid potential. As shown in Fig. 23, a common plate battery can be used for two or more stages. A more complete discussion of battery requirements is given under Power Supply.

19. *Resistance-Condenser Coupled Amplifier.* This type of circuit (Fig. 24) is similar to the preceding in all respects except that condensers are inserted between the plates and grids of adjacent stages. This makes the employment of large negative grid batteries unnecessary although it is still important that steady negative potentials be applied to the grid of each tube sufficiently large to prevent their being carried positive by the variations. For example, in Fig. 24, r_g' may be two meghoms and the grid battery emf 2 to 3 volts. This

type of circuit is in general the easiest to handle. Due to the insertion of condensers it will, of course, not serve as a direct current amplifier,

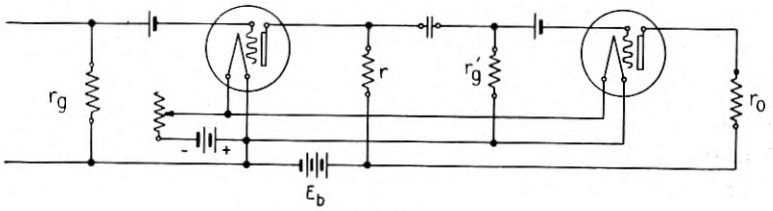


Fig. 24

but with sufficiently large condensers it can handle low frequencies with little or no distortion.

20. *Retard-Condenser Coupled Circuit.* The substitution of retard coils for resistances (see Fig. 25) in the circuit last described affects

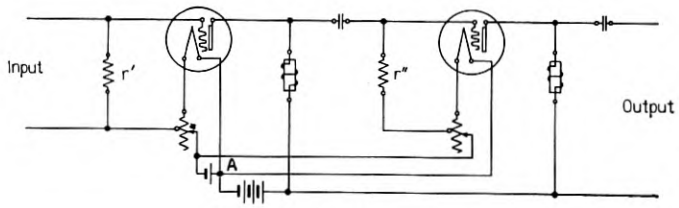


Fig. 25

the behavior of the circuit in several ways. An advantage in the change lies in the fact that a given plate potential can be secured by a smaller plate battery (Sec. 14). Since the tubes in all but the last stage of an amplifier generally act as voltage amplifiers, it is desirable that the inductance of the retard coils be made large. By employing retard coils of the proper inductance and resistance and shunting them with condensers, such an amplifier may be tuned for any particular frequency.

It follows that the width of the frequency band which can be amplified without distortion is likely to be less than for the resistance coupled amplifier. Since the impedance of the retard coils increases with increase of frequency the higher frequencies will, in general, be amplified more than the low. However, it is impossible to make retard coils without a certain amount of distributed capacity, the shunting effect of which tends to limit the amplification of the higher frequencies. By the proper design of coils it is possible to construct a retard-coupled amplifier which will give practically uniform amplification, e.g. throughout the speech range of 200 to 3,000 cycles. It

is customary to make the retard and choke coils of the toroid or closed core type.

21. *Transformer Coupled Amplifiers.* From a theoretical point of view the transformer coupled amplifier (Fig. 26) should be the ideal type. By the proper choice of transformers it should be possible to match stages with respect to one another in such a way as to obtain the greatest efficiency from tubes and batteries. The chief advantage of transformer coupling lies in the fact that the input voltage to the second stage may be made greater than the voltage

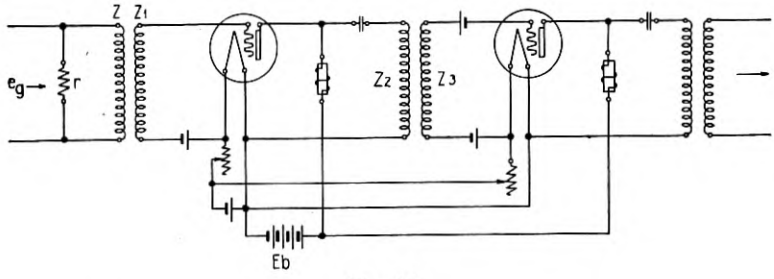


Fig. 26

output of the first, and so on, at the same time that each tube operates into a properly matched impedance to give maximum power output. When uniform amplification over a relatively wide band of frequencies is not required, the interstage transformer may be designed to step up the voltage as many as 30 to 40 times. Other advantages are the economical use of plate batteries (Sec. 14) and the elimination of grid condenser and grid leak or high voltage grid battery.

However, the difficulties attendant upon the design and making of transformers are such that to realize the apparent advantages of this type of circuit will require very careful planning. This will be illus-

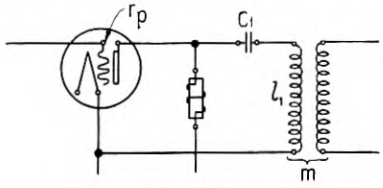


Fig. 26a

trated by the following example of a transformer to handle the frequencies of speech. Given an interstage transformer as in Fig. 26a, we will assume that the transformer works out of a tube impedance r_p and into a grid circuit impedance which has infinite resistance.

Then, imaging for the moment that the condenser C_1 has been removed, the output voltage e_2 of the transformer is

$$e_2 = \frac{emp}{r_p + jl_1p}, \tag{8}$$

in which e is the input voltage, l_1 is the inductance of the primary winding, m is the mutual between the windings, and p is 2π times the frequency. This neglects resistance of the winding and also capacity effects. Inspection of Equation 8 shows that e_2 varies with the fre-

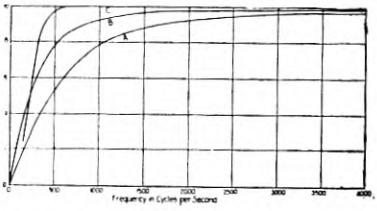


Fig. 26b—Curve A corresponds to $l=1$ henry. Curve C to $l=2$ henries, $C_1=.141$ f.

quency or with p in the manner shown in curve A, Fig. 26b, from which it is seen that the transformer tends to suppress the lower frequencies.

Curve A shows the performance of a transformer as calculated from Equation 8 assuming $r_p=5{,}000$ ohms and $l_1=1$ henry. Such a transformer would be quite unsuited for a speech frequency amplifier as it introduces very serious distortion below 1,000 cycles. Curve B is calculated on the assumption that $l_2=2$ henries and shows marked improvement over A for the lower frequencies.

In input transformer design it is ordinarily necessary to limit the inductance of the two windings not only because of the limited winding space but also because of the need of keeping down the capacity between windings and the capacity within each winding. Curve C shows the performance of the same transformer as Curve B when the capacity C_1 (Fig. 26a) is put in the primary circuit, C_1 having a value of .141 m.f. and being so chosen as to tune l_1 to 300 cycles. With the capacity present Equation 8 becomes

$$e_2 = \frac{emp}{r_p + j\left(l_1p - \frac{1}{C_1p}\right)}. \tag{8A}$$

Use of the capacity improves the transformer characteristic for all frequencies above about 200 cycles and the combination therefore gives better results in a speech amplifier than the transformer alone.

The effect of distributed capacity in the windings (present especially in the secondary because of its greater number of turns) is, more or less, to shunt the high frequencies. This may be counteracted either by the inductance in the primary winding or, if this is not sufficient, by insertion of the proper inductance in series with the primary. It may be said, in a general way, that the lower the ratio of a transformer the better suited its frequency characteristic will be to a wide band of frequencies such as occurs in speech, and transformers with a ratio of 1 to 4 are made which require no correcting provided they are properly chosen with respect to the impedance characteristics of associated tubes.

The selective amplification of an amplifier for particular frequencies may be increased by tuning one or more of the secondaries of the interstage transformers with condensers. (See Equation 8A.)

Due to the fact that there is an appreciable distributed capacity between the primary and the secondary windings, an interstage transformer supplies capacity coupling as well as inductive coupling between adjacent stages, the phase of the capacity coupling being independent of the direction of winding while the inductive coupling is not. Therefore, the transformer may be so placed in the circuit that these two effects either aid or oppose one another. In order to secure the greatest amplification they should aid.

The transformer used in speech frequency work is, in general, made with an iron core; therefore, care should be taken to prevent the d.c. component of I_p magnetizing the core and reducing its efficiency. One method of accomplishing this is shown in Fig. 26 in which the d.c. component of I_p is bypassed by a choke coil. In circuits in which two or more interstage transformers are used, attention should be paid to the danger of magnetic feed-back. This can be largely eliminated by using transformers with closed magnetic circuits. Both the toroid type core and the shell type core (commonly employed in power transformers) have been found satisfactory, and especially the latter.

22. *Amplification of Higher Frequencies.* The transformer coupled amplifier is the type perhaps best suited to use at frequencies higher than those of speech. As a special case the amplifier circuit of Fig. 27 will be considered first. This circuit contains a tuned output and should be used only in case a single frequency or very narrow band of frequencies is to be amplified but in this case will be found very satisfactory. The inductance L may consist of two parallel windings, insulated from each other to avoid any conductive connection between the plate battery and the output. Such an arrangement would be desirable if the output went to a detecting tube or another

amplifying tube. By using a variable condenser C, the frequency of maximum amplification can be readily shifted but for any one setting the amplification will be as shown by curve A of Fig. 28. For maxi-

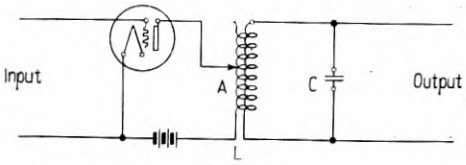


Fig. 27

mum output, the position of the tap A should be set so that the impedance of the tuned circuit LC as seen from the tube is equal to the output impedance of the tube. Tuned circuit amplifiers are especially adapted for amplification at very high frequencies (above 2,000,000 cycles), where the effect of the capacities between the elements of the tubes makes other types of amplifiers very inefficient.

The amplification curve A will be broadened when more and more turns are added to the inductance L and the capacity of the tuning condenser C is diminished due to the effect of the distributed capacity

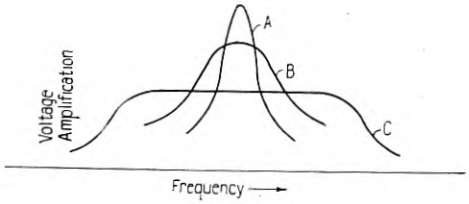


Fig. 28

of the coil. Curve B gives the amplification for a circuit where the condenser C is omitted in which case it becomes a retard coupled amplifier. Maximum amplification occurs at the natural frequency of the coil including the capacity effects of the leads and elements of the tubes.

In case a relatively wide band of frequencies is to be amplified, transformer coupling is usually resorted to, and given suitably designed transformers one to two octaves can be amplified with very fair uniformity at frequencies between 100,000 to 2,000,000 and four to five octaves at frequencies below 100,000 cycles. Use of transformer coupling will broaden the characteristic of the amplifier as shown in curve C, Fig. 28, the exact shape of this curve being largely dependent upon the design of the transformers employed.

For frequencies up to about 100,000 cycles, transformers with iron cores of the ring type are suitable and are preferably enclosed in metal covers which are grounded. A transformer suitable for frequencies higher than 100,000 cycles may consist of two choke coils (one to two inches in diameter) of very fine wire, these coils being mounted close together on a suitable form. The natural frequency of the coils will approximately determine the middle of the band of frequencies which are amplified and the coupling between the two coils will determine the width of the band, closer coupling resulting in a wider band. The coupling is generally a combination of electrostatic and electromagnetic coupling and therefore, in connecting all transformers for high frequency uses, it is essential to establish the proper phase relations between them (see paragraph 21). Each stage of the amplifier should be shielded as shown in Fig. 29 although in certain cases it may be dispensed with. The shielding should

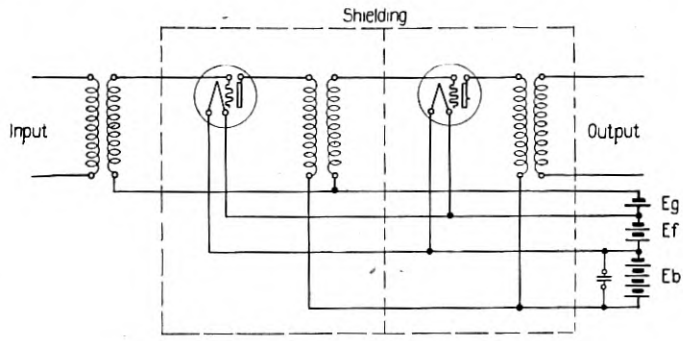


Fig. 29

consist of brass or copper sheeting made into boxes with well-soldered joints and tightly fitting covers. Holes through the shielding should be just large enough to pass the insulation of the wires.

A common plate battery may safely be used for four or more stages provided a condenser is placed across the terminals of the battery as shown in Fig. 29. This condenser should have a capacity large enough to offer practically no impedance to the high frequency currents, and its use may be desirable although the plate battery is common to but two stages. Use of a common grid battery, as shown, introduces a small feed-back from the second stage to the first. This feed-back may be either positive or negative, depending upon the phase relations in the intermediate transformer and may be eliminated by placing a condenser across the grid battery terminals.

It has been the practice in high frequency amplification to use tubes with μ's between 6 and 10, interstage transformers being selected to step up the voltage as much as is possible consistent with the desired flatness of the amplifier characteristic. In general, the larger the ratio of the transformer, the more pronounced is the peak of the characteristic. Other things being equal, the most suitable tubes are those with the smallest internal electrostatic capacities. The largest of these capacities, in general, is that between grid and plate and tubes have been produced in which this does not exceed 5 $\mu\mu$ f. and in which the internal plate resistance is about 20,000 ohms.

In amplifying the higher frequencies the feed-back which occurs through the tube may require attention. In section 13, it was pointed out that an inductive output for a tube gives rise to a negative resistance characteristic in the input which means that feed-back is occurring. To eliminate the possibility of singing and also to eliminate unequal amplification of different frequencies which feed-back introduces, various means of neutralizing it have been proposed.[29] One

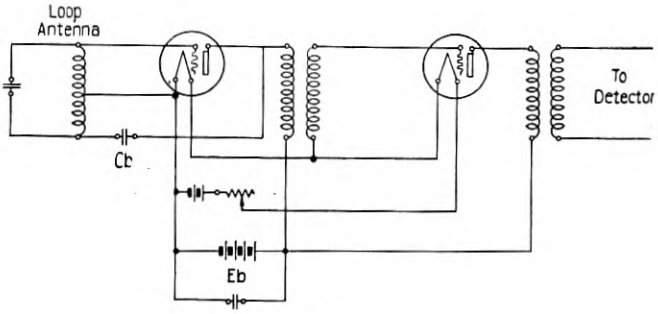


Fig. 30

such means is illustrated in Fig. 30 which is drawn to show radio reception with a loop antenna. Note that the grid of the first tube is joined to one end of the loop and the plate is joined to the other end through the balancing condenser C_b, the filament being joined to the midpoint of the loop. When C_b is chosen equal to C_3 the capacity between grid and plate, it is evident that the feed-back occurring through the tube is just balanced by that occurring through C_b. By adjusting the condenser C_b so as to permit of feed-back, very large amplification may be obtained at a single frequency but at the expense of flatness of characteristic.

[29] See Patent No. 1,183,875 issued to R. V. L. Hartley, and Patent No. 1,334,118 issued to C. W. Rice.

23. *Feed-Back Amplifiers.* This amplifier may be either resistance or inductive coupled, a typical resistance coupled circuit being shown in Fig. 31. In a feed-back circuit, attention must be paid to phase relations. In Fig. 31, let the arrow along the resistance R_1 represent an increase in electron current to the grid of tube A. This corresponds to an increase in the potential of this grid. In phase with this increase in potential is an increase in electron current in R_2 as shown by the arrow. This, in turn, corresponds to a fall in potential of the grid of tube B and therefore to a reduction of the I_p in B, as indicated by the arrow at R_3, which produces an increase in I_p in C. Therefore, in this particular circuit the correct phase relations require the output of one tube to be returned to the input of the second preceding

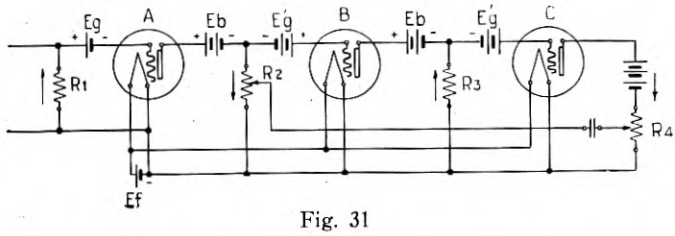


Fig. 31

tube or one of its alternate preceding tubes. The amount of energy fed back can readily be controlled by varying the portion of R_2 through which the feed-back current flows.

24. *Push-Pull Amplifier.* See Fig. 32. This type of circuit is particularly useful as a terminating stage since it makes possible the use of a low impedance in the output circuit without serious distortion. If the tubes A and B have identical characteristics, it is readily seen that the coils of the output transformer may be so connected that the fundamental and odd harmonics will aid one another, while all even harmonics will oppose. Since the third and higher harmonics (counting the fundamental as first) are very small compared to the second, this circuit gives very nearly distortionless amplification. In speech amplifiers it permits of considerable overloading without this being very apparent in the quality of the output.

By reversing the transformer connections it is possible to cause the circuit to add the even harmonics and give the differences of the odd.

An additional use for this circuit will be pointed out in the section dealing with modulation.

Fig. 33 shows a special type of push-pull circuit which is particularly adapted to the amplification of steady and low frequency volt-

ages. It consists of a Wheatstone bridge in which two similar tubes form one pair of arms. The output circuit is the branch in which the galvanometer is ordinarily placed. When a voltage is applied to the two tubes in such manner that the potential of one grid is raised by the same amount as the grid of the other is lowered, the bridge becomes unbalanced and current flows through the output

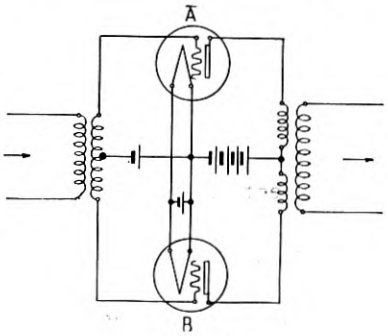


Fig. 32

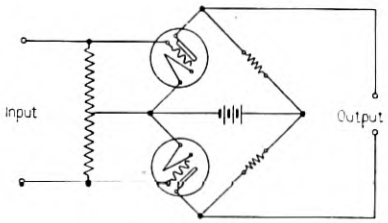


Fig. 33

branch. For small applied voltages the amplification of the circuit is very nearly distortionless. The circuit has the obvious disadvantage of requiring a close balance between the tubes and is therefore liable to require careful adjustment during use.

The push-pull circuit possesses one marked advantage over the resistance-coupled d.c. amplifier described in Sec. 18 for in it current flows through the output branch only when voltage is applied to the input. For the same reason it is also useful for amplifying low frequency alternating voltages.

V. Amplifier Power Supply

The proper power supply for amplifiers is an item of prime importance.

25. *Plate Voltage Supply.* The principal requirement placed on plate voltage is that it be steady. For this reason storage batteries are usually best, but good dry cells are more often used and when fresh, prove very satisfactory. The principal trouble encountered in the use of dry cells arises from an attempt to use partially rundown cells. A dry battery should be tested periodically for voltage, the reading being taken while the battery is delivering a current at least as large as that drawn by the amplifier. Whether dry cells or storage cells are used for plate voltage, in general not more than four stages

should be operated from a single battery. In the case of a retard coupled amplifier whose stages are tuned with condensers, each stage should preferably have a separate plate battery to reduce the tendency to "sing."

A generator as a source of plate voltage is frequently used for power amplifiers. In case a direct current generator is used, a filter is generally necessary in the plate circuit to remove commutator ripples.

26. *Filament Voltage Supply.* A source of constant filament voltage is not necessary in order to insure constant space current within the tubes at temperature saturation, but in general any variation in filament current will affect the relative potential difference between filament and grid, and is, therefore, equivalent to a variation in the input voltage. This possible source of trouble must be particularly guarded against in such a circuit as that shown in Fig. 25, in which a portion of the adjustable resistance of the filament circuit is included between the filament and the "common point" A. If storage batteries are available, they form the best source of filament current; generators have been used satisfactorily however.

27. *Sources of Grid Potential.* A flow of electrons to the grid of a tube is liable to result either in distortion or a loss of amplification or both (see Secs. 15 and 16). The steady negative grid potential required to prevent the input voltage carrying the grid to a positive potential may be obtained from either one of two sources: by a grid battery or by an IR drop in some resistance in the circuit. The requirements for a grid battery are very light since it is called upon to give no appreciable current. As was pointed out in Sec. 26, use of an IR drop for grid voltage pre-supposes steady filament or plate battery according to circumstances, and since the proper grid battery is readily obtainable the use of an IR drop is likely to prove desirable only in very unusual circumstances.

VI. Troubles in Amplifier Circuits

28. *Noise.* The noise in amplifier circuits is due to several causes which may, in general, be grouped into two classes. Certain noises originate within the tubes and other noises find their origin in the circuit. The amount of noise in any amplifier limits the minimum input voltages which it will handle satisfactorily, for obviously input voltages which produce output currents of the same order of magnitude as the currents giving rise to noise will not be satisfactorily amplified.

29. *Tube Noises.* Tubes may be responsible for three distinct kinds of noises. (a) Ringing or rattling is due to the vibration of the tube elements and may be eliminated by proper tube construction or by some form of vibration proof suspension for the early stages of the amplifier. (b) Crackling may be produced by high resistance films on the inner surface of the bulb, forming conducting paths between the leads. Faulty electrical contact between the plate, grid and filament and their respective leads is also a frequent source of crackling. Furthermore, in tubes which are well constructed in regard to the points just mentioned, but which contain tungsten filament, crackling may be observed. This trouble is not found in all tungsten filament tubes but, when present, is sufficiently marked to become apparent in a two stage amplifier. (c) In carefully constructed amplifiers of more than three or four stages a noise which can best be described as a hissing or sighing is certain to be present. It appears to be related to an unavoidable statistical variation in the escape of electrons through the grid to the plate. Its magnitude has been found to correspond approximately to an output voltage from the first stage of between 5×10^{-7} volts and 5×10^{-6} volts. Between these limits the noise is found to increase as the output impedance of the first stage is increased, and also to increase as the resistance across the terminals of the input increases. Its components, above 300 cycles, appear to be of about equal magnitude and uniformly distributed. It is, therefore, impossible at the present time to build amplifiers to handle voltages of less than this order of magnitude, at any rate when the frequencies involved are in the audible range.

30. *Circuit Noises.* In general, circuit noises in amplifiers are due to one or more of the following causes: variations in grid and plate batteries, loose contacts and variations in resistances, leakage of condensers and leakage across the insulating mounting upon which the amplifier parts are fastened, and external electric or magnetic fields acting inductively on the circuit. The remedy in each case is obvious once the exact cause has been found. To eliminate inductive effects in the wiring it is usually sufficient to run wires in pairs and to shield them electrically, the shielding being grounded. In laying out the various parts of an amplifier it is well to place the bulky pieces at points in the circuit at which they will have as near zero potential as possible.

31. *Singing.* Singing, which is one of the most serious troubles in amplifiers, is always due to some form of feed-back. This may be magnetic, electrostatic, or in the form of mechanical vibrations as in an amplifier having a microphone attached to the input and a receiver

to the output. Mechanical feed-back can also occur in the case of tubes whose parts can easily be set into vibration and a cure is usually found in some form of vibration-proof mounting. The coupling which is responsible for feed-back may be difficult to locate, but when found can usually be removed. Both retard coils and transformers may afford an easy method of coupling due to stray fields. If the coupling induces voltages which are in phase with the input voltages, it may cause singing, and if out of phase, the amplification may be seriously reduced. Closed core coils and magnetic shields are the usual remedies for this condition, although a rearrangement of the circuit parts may be necessary.

Certain kinds of electrostatic feed-back may be removed by enclosing each stage in a separate grounded metal cage or box. The electrostatic coupling due to tube capacities (Sec. 13) cannot be eliminated but it is possible to so design circuits that trouble from this source will not present itself. Thus an inductive impedance in the output circuit may prove troublesome because it induces a *negative* resistance back in the input circuit; a non-inductive output can never do this. Feed-back through tubes increases with frequency, and in the case of high frequencies, it may sometimes be necessary to use resistance coupled rather than reactance coupled circuits.

32. *Blocking.* Two entirely different types of blocking may occur in an amplifier. They both result from the grid of one or more tubes having been carried to a positive potential by the input voltage. While positive, the grid picks up a negative charge of electrons which is removed more or less rapidly by the grid leak. In case the leak resistance is high, a residual charge may remain upon the grid for an appreciable length of time, depressing its mean potential to so low a value that the output of the tube is cut to zero or very nearly zero. The remedy is obviously to reduce the input voltage or to increase the voltage of the negative grid battery. In certain cases, a readjustment of the resistance of the grid leak may be desirable.

The second type of blocking involves secondary emission from the grid as discussed in Sec. 9. It can occur only when the input is sufficient to force the grid potential of some tube positive by as much as 10 or 15 volts, and then if the grid leak resistance is large enough, secondary emission will hold the grid at about this positive potential and entirely prevent proper functioning of the amplifier. In eliminating this type of blocking, the first step should be to note the effect of increasing the filament currents as secondary emission is less likely to occur when the filament yields a copious supply of electrons. If this does not remove the trouble, the negative grid batteries in the

stages at fault may be increased and lower grid leaks may be desirable. The volume of input to each stage should also be considered.

33. *Distortion.* Distortion in an amplifier circuit may result either from a failure to amplify all frequencies by the same amount or from the generation of overtones of the fundamental frequencies in the input.

The unequal amplification of various frequencies arises from the presence of resonant characteristics in the circuit. This may take the form of a feed-back which discriminates in favor of certain frequencies, the feed-back not being pronounced enough to cause singing. A negative feed-back may also occur, causing a loss of efficiency over some particular frequency range.

The distortion which arises from the generation of overtones is due to non-linear voltage-current characteristics in one or more branches of the circuit. The usual sources of this trouble are curvature of the plate and grid characteristics (See Equation 4) and the variable permeability of the iron used as cores. With properly chosen coils, practically distortionless amplification can be secured by the method indicated in Sec. 12. In general, to accomplish this, the output impedance need not be more than two or three times r_p. In case it is necessary to use a low output impedance in the final stage, distortion may be reduced by using the push-pull circuit of Sec. 23.

In using an amplifier under circumstances such that distortionless output is desired, care should be taken that no tube by itself is overloaded or caused to work in such fashion that its dynamic characteristic is curved. Distortion which arises from curvature of this characteristic can be detected by inserting an ammeter in the plate circuit of each tube. When each characteristic is straight, or nearly so, there should be no change in ammeter reading as the source of input voltage is thrown on and off, and in the case of a variable input such as that arising from speech, the ammeter readings should remain constant while the amplifier is in operation. This test will not detect distortion which arises from selective amplification with respect to frequency.

34. *Calculation and Measurement of Amplification.* Provided all parts of an amplifier circuit are functioning properly and its constants are known, its amplification can be calculated quite accurately. The following example will illustrate the procedure to be followed in any case. Referring to the transformer coupled amplifier of Fig. 26, assume that the ratio of the first input transformer is $z{:}z_1$ and that the ratio of the second input transformer is $z_2{:}z_3$; assume also that z_2 is numerically equal to r_p, the plate circuit resistance of the first tube.

Then, calling e_g the input voltage, the voltage across the first tube is

$$e_g \sqrt{\frac{z_1}{z}};$$

the voltage across the primary of the second input transformer is

$$e_g \frac{\mu}{2} \sqrt{\frac{z_1}{z}}$$

since z_2 is numerically equal to r_p; and across the secondary is

$$e_g \frac{\mu}{2} \sqrt{\frac{z_1}{z}} \sqrt{\frac{z_3}{z_2}}.$$

Hence the voltage amplification of this portion of the amplifier is

$$\frac{\mu}{2} \sqrt{\frac{z_1}{z}} \sqrt{\frac{z_3}{z_2}},$$

and a similar argument applies to the following stages.

The measurement of amplification can be accomplished by the obvious procedure of determining the magnitudes or the relative magnitudes of the input and output current. This can be done for either a single stage or for several stages at once.

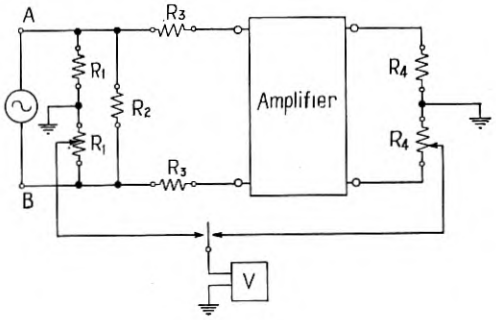


Fig. 34

A very satisfactory circuit for measuring amplification is illustrated in Fig. 34 and through use of a vacuum tube voltmeter (see Sec. 56) as a comparison means, it is capable of an accuracy of about 2%. The conditions which the resistances R_1 and R_2, etc., should satisfy are very simple. The network connected between the oscillator and amplifier input should present an impedance, looking into it from the right, equal to the input impedance of the amplifier; like-

wise the output impedance of the amplifier should equal $2R_4$. The input impedance of the vacuum tube voltmeter is so high as not to shunt the resistances across which it is connected appreciably. The grounds at the mid points eliminate the disturbing effects of capacities to ground. Under these conditions the voltage amplification a is given by the equation:

$$a=\frac{\alpha}{\beta}\left[\frac{4R_3(2R_1+R_2)+2R_1R_2}{2R_3(2R_1+R_2)+2R_1R_2}\right],$$

in which α, β are the fractions of R_1 and R_4 respectively, across which the voltmeter is connected to obtain equal readings when the switch W is thrown from one position to the other. In case R_2 is made quite small with respect to R_1, R_3, the expression for a reduces approximately to $a=\frac{2\alpha}{\beta}$. An expression for current amplification can readily be derived.

Another simple measuring circuit is shown in Fig. 34a in which O is an oscillator of the desired frequency, F is a filter to remove harmonics from the oscillator current, R_1R_2 and R_3R_4 are attenuating networks consisting of resistances, WWW are switches by which the telephone receiver T can be joined either to the output of the amplifier

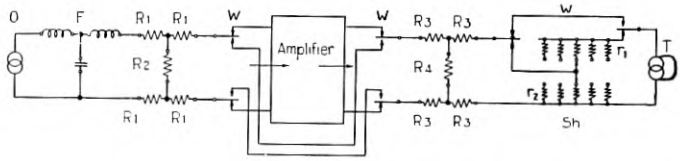


Fig. 34a

or directly to the oscillator, and also provide means for removing the known attenuation in the shunt Sh (receiver shunt) at the same time the amplifier is removed. By the proper design of the receiver shunt, which will be discussed presently, the attenuation required to give the same volume of sound in the receiver whether the amplifier is in or out may be read directly.

In setting up the circuit of Fig. 34a, special attention must be given to the networks R_1R_2 and R_3R_4. In addition to reducing the input to the amplifier to a value safely below the overload point, R_1R_2 should be designed to present an impedance (when seen from the amplifier) equal to that out of which the amplifier is to operate in service. Otherwise the measurements of amplification may be without significance.

The network R_3R_4 serves two important purposes. It is designed to present toward the amplifier the same impedance as the amplifier is to work into in service, and this in turn requires that the input and output impedances of the amplifier be practically equal (or if not, then small with respect to R_1) for otherwise the network R_3R_4 when joined to R_1R_2 will not draw the same fraction of current as the amplifier, thereby upsetting the comparison upon which the measurements are based. Furthermore, the attenuation in R_3R_4 is to be sufficiently large that variations in the impedance of the receiver and its shunt as seen from R_3R_4 will not appreciably affect the impedance into which the amplifier works as the receiver shunt setting is changed. A simple calculation will show how great the attenuation must be in any given case to satisfy these conditions.

Proper values for the steps of the receiver shunt may be calculated as follows, reference being made to Fig. 35 in which the currents and potentials indicated are in accordance with the assumptions made

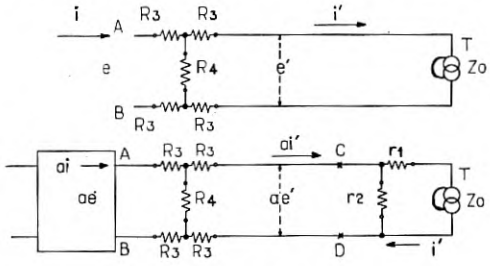


Fig. 35

regarding the attenuation in the various portions of the circuit. Calling a the amplification to be measured it may readily be shown that

$$\frac{(a-1)^2}{a}=\frac{r_1}{r_2}. \tag{9}$$

Or if R is the impedance of the network R_3R_4 as seen from the receiver, and the shunt Sh is proportioned so that it also presents the impedance R to the receiver, whence

$$R=r_1+\frac{r_2R}{r_2+R},$$

then Equation (9) gives

$$a=\frac{R+r_2}{r_2}.$$

Taking account of the necessary approximations it is readily possible to measure current amplification to within 5%, for a range of

frequencies extending from 200 to 3,000 cycles. Receiver shunts are made which, in 10 to 15 steps, will reach a maximum reduction ratio in current of 25:1 which corresponds to an energy reduction of 625:1, and this does not represent the greatest range possible.

In case a rougher approximation of the amplifying power is sufficient, the circuit of Fig. 34*a* may be simplified by omission of the network R_3R_4 and reversal of the receiver shunt to present a constant impedance (except for variation of impedance with frequency and phase angle) toward the amplifier. The network R_1R_2 should preferably be retained and should be so proportioned that the current through the right hand R_1 branches is practically the same whether connected with the amplifier or directly to the receiver.

In measuring the over-all amplification of a multistage circuit, it will probably be desirable to add fixed but known attenuation units similar to R_3R_4 to the receiver shunt which may be cut in or out as required. These units may be given an attenuation equal to and twice the total attenuation of the shunt, etc., after the fashion of the ordinary resistance box. In constructing attenuation networks the arrangement indicated in Fig. 34*a* will be found desirable in that the symmetrical placing of the branches tends materially to eliminate errors which might otherwise arise due to capacities to ground in the oscillator and amplifier. Pairing of lead wires and shielding of leads and resistance coils will be found desirable for accurate work.

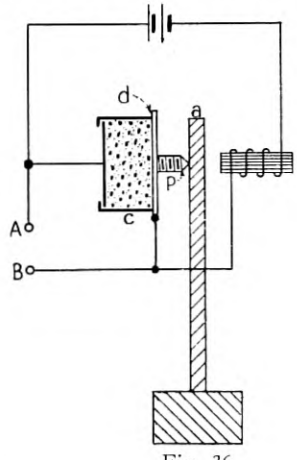


Fig. 36

The filter F should be used in case the amplifier tends, because of the limited range of frequencies which it passes or because of some other kind of distortion, to modify the quality of the note given by

the oscillator, it being very difficult to match sounds for intensity which differ in quality.

A very satisfactory type of audio-frequency generator is shown in Fig. 36; it is a buzzer which operates, not by making and breaking current, but by varying it periodically with a microphonic button. The vibrating parts of this generator may be tuned to any audio-frequency, e.g. 800 cycles, and it gives quite accurately a sinusoidal variation of current, although it is customary to insert a filter (Sec. 6) to insure the input energy being accurately of one frequency.

VII. Thermionic Modulators

In discussing modulation the terminology which has been developed in connection with radio and carrier-current signaling will be used.

By the term "modulation" is meant the varying of the amplitude of a relatively high frequency wave, so that its envelope represents a particular low frequency wave or combination of such waves. (See Curves A, B and C, Fig. 37). The combination of low frequency

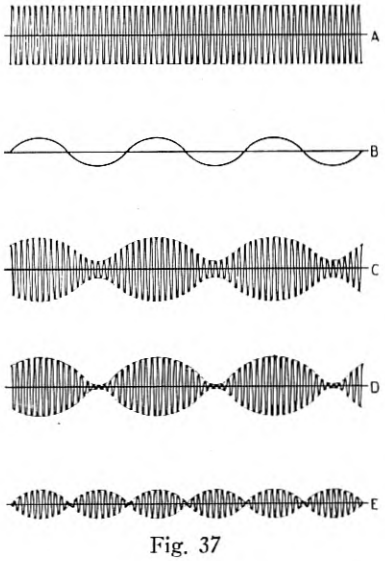


Fig. 37

modulating waves may be very complicated, as in the case of speech, but the principle involved is common to all cases of modulation and can be clearly brought out by the consideration of a single low frequency.

Let Fig. 37, C, represent a high frequency wave modulated by a sinusoidal low frequency. The wave C can be represented by

$$y = a(b + \cos qt) \cos pt, \tag{10}$$

in which $\frac{p}{2\pi}$ is the high (carrier) frequency and $\frac{q}{2\pi}$ the low (signal) frequency. Equation 8 can be rewritten in the form

$$y = ab \cos pt + \frac{a}{2}[\cos (p-q)t + \cos (p+q)t], \tag{11}$$

which brings out the fact that the modulated wave C contains, in general, three distinct frequencies—the carrier frequency $\frac{p}{2\pi}$, a difference frequency $\frac{p-q}{2\pi}$, and a summation frequency $\frac{p+q}{2\pi}$. These latter frequencies represent the so-called "side bands" of the modulated wave.

Two special cases of the wave represented by Equations 10 and 11 are represented graphically at D and E of Fig. 37 and correspond to $b=1$ and $b=o$ respectively. When $b=1$ it is evident that the amplitude of each side band is half the amplitude of the carrier frequency; such a wave is said to be "completely modulated"; when $b=0$ the carrier frequency $\frac{p}{2\pi}$ is absent altogether.[31]

35. *Means for Producing Modulation.* Perhaps the simplest case of modulation is that illustrated by continuous-wave radio telegraphy, in which the intermittent radiation of a uniform wave is accomplished by means of a telegraph key. In most cases, however, modulation requires a gradual change in the amplitude of the high frequency wave. For effecting this the vacuum tube possesses two properties which make it particularly useful—(a) the E_g, I_p characteristic is very nearly parabolic (Sec. 8); (b) the current in the plate circuit is a function of the grid potential (Fig. 10).

Circuits, by means of which modulation may be effected by each of these properties, are described in the following paragraphs.

36. *Modulation by Curved Characteristic.* Considering the circuit of the type illustrated in Fig. 38, let it be assumed that a voltage

$$e = A \cos pt + B \cos qt$$

is applied to the input of the tube. The result is shown graphically in Fig. 39. When this value of e is substituted in Equation 4 we

[31] For a more complete discussion of modulation and the nature of the side bands, see R. V. L. Hartley, Procd. Inst. of Radio Engrs., Feb., 1923, or *Bell System Technical Journal*, Apr., 1923.

obtain for the modulated output (i.e., terms whose frequencies are of the order $\frac{p}{2\pi}$).

$$J_m = A\left[\frac{\mu}{r+r_p} + \frac{\mu^2 r_p r_p'}{(r+r_p)^3} B \cos qt\right] \cos pt. \tag{12}$$

As pointed out above, the first term gives the carrier wave and the second term, the two side bands. It will be remembered that this

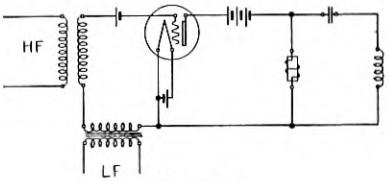


Fig. 38

equation neglects terms of higher order than the second, which is permissible, so long as the tube characteristic is approximately parabolic.

Certain points regarding Equation 12 should be noted. In the first place, the amplitude of the side bands is proportional to the

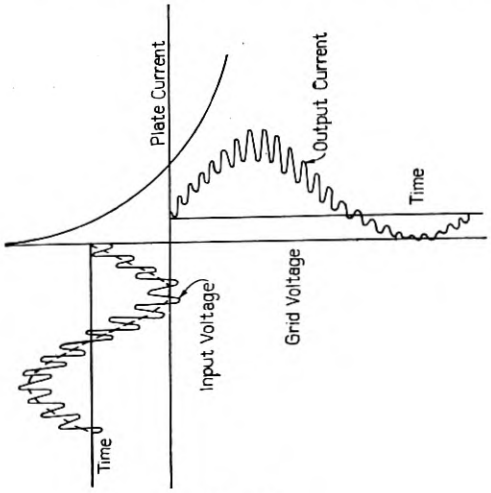


Fig. 39

product AB and is therefore, independent of the relative amplitudes of the original carrier and modulating frequencies. Also the modulated current is proportional to the first power of B, the amplitude

of the low frequency wave; i.e., although the modulation is effected by the curvature of the tube characteristic, the modulated output is free from distortion. Furthermore, the modulated output voltage is proportional to $\frac{r}{(r+r_p)^3}$ which is a maximum when $r=\frac{1}{2}\ r_p$; and the modulated output energy is proportional to $\frac{r}{(r+r_p)^6}$ which is a maximum when $r=\frac{1}{5}\ r_p$.

Another type of circuit in which the modulation is dependent upon the curvature of the E_g,I_p characteristic is shown in Fig. 40. The two tubes are supposed to be alike; so long as no low frequency is impressed on the grids the high frequency space currents are equal,

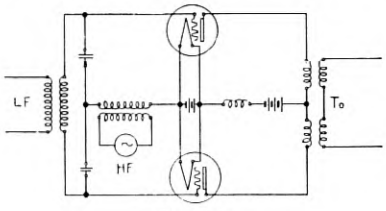


Fig. 40

each passing through one of the primary coils of the transformer T_0, the order of winding these coils being such that this condition gives zero current in the secondary. However, the presence of a low frequency voltage (*L.F.*, Fig. 40) raises one grid potential at the same time that it lowers the other, with the result that the high frequency currents in the two primary coils are no longer equal, and a high frequency current therefore flows in the secondary of T_0, the amplitude of which is determined by the degree to which the two tubes are unbalanced by the low frequency input. It is apparent that the output of this modulator circuit contains only the two side bands and none (or very little, if the tubes are not exactly identical) of the carrier frequency and therefore corresponds to curve *E* in Fig. 37. It is particularly useful in communication circuits where several telephone or telegraph channels are desired on the same pair of wires. Since only the side bands are transmitted the total current which must be handled by repeaters and other line apparatus is materially reduced. By the use of the proper wave-filter it is also possible to suppress one side band, thereby approximately cutting to one half, the width of the frequency band to be transmitted. As will be pointed out under homodyne detection, the suppressed carrier frequency must be supplied locally before detection can occur.

37. *Modulation Effected by Controlling Plate Current with Grid Potential.* Numerous circuits have been developed for modulation, making use of the fact that the grid potential affects the resistance of the plate circuit. Two circuits of this type are shown in Figs.

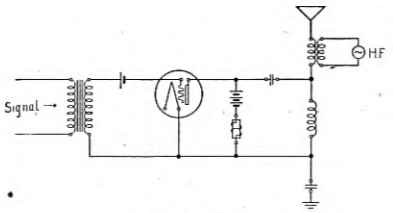


Fig. 41

41 and 42. In the first, the plate circuit of the tube forms a shunt across a portion of the antenna inductance. As the grid potential is varied, the antenna is, therefore, thrown more or less out of tune, with the consequent radiation of a variable amount of high frequency energy.

Fig. 42 shows one of the most efficient modulating schemes thus far developed. As the grid potential of the modulator tube *A* varies, causing a change in plate current through this tube, the plate voltage

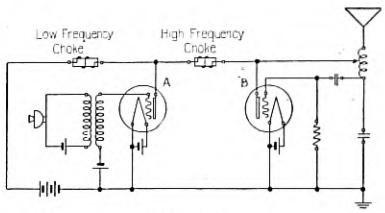


Fig. 42

applied to the oscillator tube, *B*, fluctuates, because of the presence of the low frequency choke coil. Under this condition of variable plate-voltage the oscillator gives a variable amount of high frequency energy to the antenna. When used for speech modulation this circuit is very efficient and gives good quality. The tubes *A* and *B* are ordinarily of the same type.

Many other modulating circuits have been designed, and those given above are to be considered merely as illustrative of the general manner in which the properties of the vacuum tube may be applied to the problem of modulation.[32]

[32] For other types of modulator circuits see a paper by R. A. Heising, Procd. Inst. Radio Engrs., Aug., 1921.

VIII. Thermionic Detectors

Like the modulator the detector is a device for the production and separation of difference frequencies. The object of modulation is, in general, to transform a high frequency $\frac{p}{2\pi}$ and a low frequency $\frac{q}{2\pi}$ into two high frequency side bands, $\frac{p \pm q}{2\pi}$. Detection accomplishes the inverse operation of forming from a carrier frequency $\frac{p}{2\pi}$ and either or both side bands the original low frequency $\frac{q}{2\pi}$, detection often being referred to as demodulation. Detection, like modulation, can be most readily described by the consideration of a single pair of frequencies.

When carried out by means of a vacuum tube it results from rectification in either the grid circuit or the plate circuit. This rectification

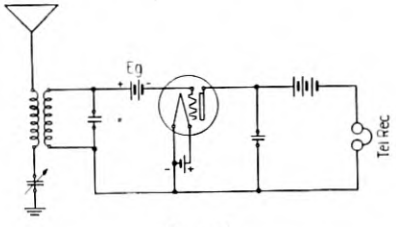


Fig. 43

may arise either from unilateral conductivity or a curved current-voltage characteristic as pointed out in the following paragraphs.

38. *Detection by Curved Plate Characteristic.* Considering the circuit shown in Fig. 43 and assuming an input voltage $e = A(B + \cos qt) \cos pt$, it follows from Equation 4 that the output current, considering only those terms whose frequencies are of the order $\frac{q}{2\pi}$, is

$$J_d = \frac{1}{2!} \frac{\mu^2 r_p r_p'}{(r + r_p)^3} A^2 (B \cos qt + \frac{1}{4} \cos 2qt). \qquad (13)$$

The current J_d, known as the "detected current," therefore, consists of a term whose frequency is $\frac{q}{2\pi}$ and another term whose frequency is twice this. The presence of these two frequencies is readily understood. The detected current of frequency $\frac{q}{2\pi}$ corresponds to the

difference frequency (See Equations 10 and 11) of the carrier of amplitude AB and each side band of amplitude $\frac{A}{2}$ and is therefore proportional to $2 \cdot \frac{A}{2} \cdot AB$. The second term of the detected current represents the difference frequency $\frac{q}{\pi}$ between the two side bands themselves and, as is to be expected, its amplitude is proportional to $\frac{A^2}{4}$. In case one of the side bands is suppressed before detection, this term of double frequency is entirely absent in the detected current. Furthermore, the amplitude of the detected current of frequency $\frac{q}{2\pi}$ is independent of the *relative* amplitudes of the carrier wave and the side bands. In general, AB is large compared to $\frac{A}{2}$ with the result that the term of double frequency in the detecting current is negligible. It follows, therefore, as in the case of modulation that the detecting current is practically free from distortion.

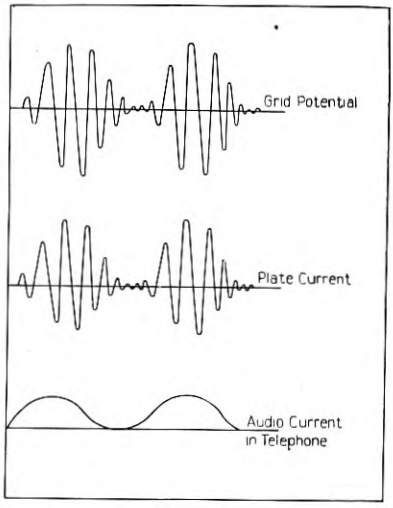


Fig. 44

The detecting action resulting from the curved plate characteristic is shown in Fig. 44.

Equation 13 leads to the result that the output voltage of a detector tube, when working as above, is a maximum when $r = \frac{1}{6} r_p$. In using these relations note that r represents the value of the output resistance

for the high frequency and not the low frequency. When using an amplifier on the output of a detector (see Fig. 45), it is important to choose r to give the maximum detecting *voltage*. As in amplifier cir-

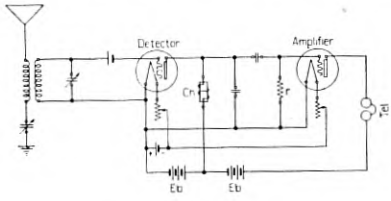


Fig. 45

cuits it is essential in this type of detector that the grid remain always negative. See also the sections on amplifiers.

39. *Detection by Rectification in Grid Circuit.* This type of circuit (see Fig. 46) is now in very general use for radio purposes, and is characterized by the grid blocking condenser C_s. Contrary to the preceding type of detector, the present requires the flow of electrons to the grid and works best when the grid is held permanently at a

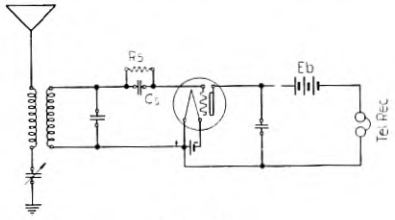


Fig. 46

small positive potential. The action of the high frequency input causes the periodic accumulation of a negative charge upon the grid and the blocking condenser, thus lowering E_g and diminishing I_p. This action is clearly illustrated in Fig. 47. This circuit is most effective when the carrier frequency is much greater than the signal frequency, and not as efficient as the circuit described in Sec. 38 when the carrier is say only four or five times as great as the signal frequency.

Attempts to deduce a quantitative relation for the detecting current in this type of detector have as yet met with little success, one of the principal reasons being that very little is known about the "dynamic" grid current characteristic.[33] Experiments show, however, that the

[33] For a discussion of this topic see Hulbert & Breit, *Phys. Rev.*, Nov., 1920, pp. 408–419; Oct., 1920, pp. 274–281.

detecting current is practically proportional to the square of the input voltage, provided that this is small, thus establishing the relation,

$$J_d = ae^2,$$

which corresponds in form to Equation 13 above.

In designing this type of detector circuit, attention must be paid to the value of the blocking condenser C_s and its leak R_s. It is clear that the capacity of C_s should be sufficiently small to cause the grid to undergo the maximum potential change as a result of the relatively

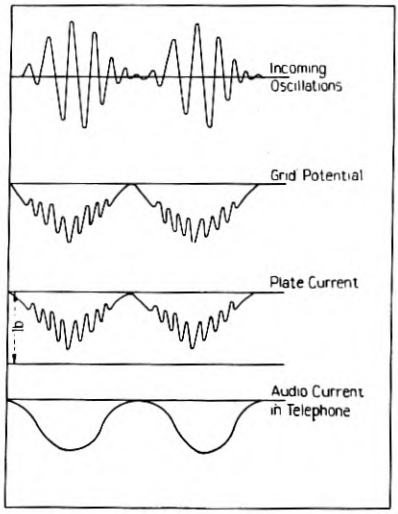


Fig. 47

small electron charge picked up, and yet it must be several times larger than the tube capacity between grid and filament. Furthermore, the time constant of C_s and R_s should approximately match the frequency of the detected current. With the more common detector tubes and radio frequencies, capacities of the order of 200 $\mu\mu$f. are satisfactory.

Detector circuits with grid blocking condensers may be coupled to amplifiers as readily as the other type of detector, and, in general, a higher output resistance for the detector can be used, thus making possible more efficient coupling between the detector and the first stage of amplification. In increasing the output resistance of the detector it should, however, be borne in mind (see Sec. 13) that a secondary result is to reduce the input impedance of the detector, which may entail a reduced input voltage.

40. *Heterodyne and Homodyne Detection.* In continuous-wave radio telegraphy, the dots and dashes of the code are transmitted by a continuous carrier wave of a single frequency. *Heterodyne* reception consists in supplying a slightly different frequency at the receiving station, the transmitted and locally generated frequencies when applied to the detector acting exactly as the carrier and side band frequency described above. The useful output of the detector is the difference frequency which, of course, is chosen in the audible range.

It follows from Sec. 38 that the heterodyne detecting current is proportional to the product of the amplitudes of the transmitted and locally generated waves. Because of this fact a feedback circuit may be used to advantage as a means of increasing the strength of both high frequency terms. In the usual type of feed-back detector, the detector tube is also used as the source of local high frequency. Such

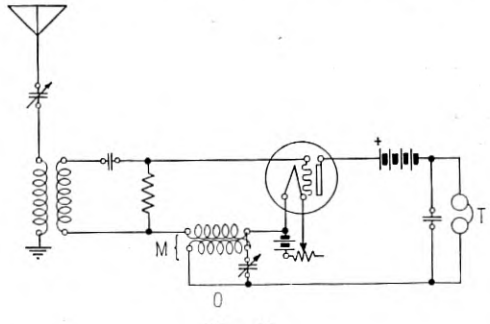


Fig. 48

a circuit is shown in Fig. 48. The oscillatory circuit O is tuned to differ in frequency from the incoming signal by an amount which will give a satisfactory difference frequency in the telephone receiver T. By varying the coupling at M, the intensity of the beat note can be readily changed. It must, however, be sufficient to cause the circuit as a whole to oscillate at the natural frequency of O. A feedback arrangement is particularly applicable to those cases in which the carrier frequency is much higher than the signal frequency and is therefore generally used with a blocking condenser.

In telephone systems, whether radio or carrier current, it is frequently desirable to suppress the carrier frequency and transmit only one[34] or both of the side bands. Detection with only the side bands present would result in a double frequency detecting current which obviously would not be permissible in a telephone circuit. Whenever

[34] For a discussion of the advantages of *single* side band transmission, see reference given in footnote 13.

the side bands alone are transmitted, a locally generated high frequency exactly equal in frequency and phase to the original carrier frequency must be supplied.[35] This is known as *homodyne* reception.

41. *Measurement of Detection Coefficient.* The constant a in the relation $J_d = ae^2$ is called the "detection coefficient." Its measurement by direct means is not difficult but as seen from Equation 13 it involves so many factors that no satisfactory indirect methods of determination have been developed. The requirements of the direct method are quite obvious, and for circuit details reference is made to the *Thermionic Vacuum Tube* by van der Bijl.

42. *Detecting Efficiency.* A knowledge of the detecting coefficient a tells very little about the detecting efficiency of a tube, the efficiency being defined as the ratio between the low frequency energy in the output and the high frequency energy in the input. The efficiency involves the input impedance of the tube which is a function of the circuit constants as well as the tube. It is therefore impossible to specify the detecting efficiency of a tube without certain data concerning the circuit in which it is to be used; a is therefore without much significance.

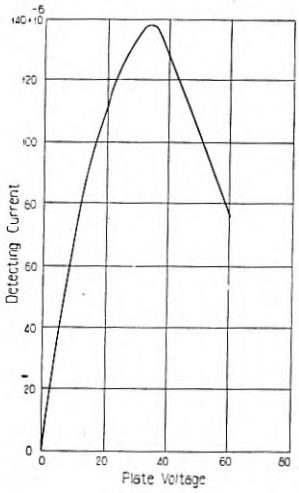


Fig. 49

43. *Detecting Coefficient and Plate Voltage.* The variation of the detecting coefficient with plate voltage depends upon the type of circuit. If detection is accomplished by a curved plate characteristic, experiment shows (see Fig. 49) that the operation is best when the effective voltage is about equal to the potential drop in the filament,

[35] See J. R. Carson, Procd. Institute of Radio Engineers, Vol. II, p. 271, 1923.

it being presupposed (see Sec. 38) that E_g is enough less than zero to keep the grid negative at all times. If detection is accomplished by means of a blocking condenser, the variation is as shown in Fig. 50, no sharply defined maximum being present in the curve.

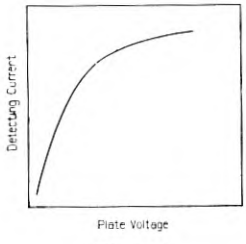


Fig. 50

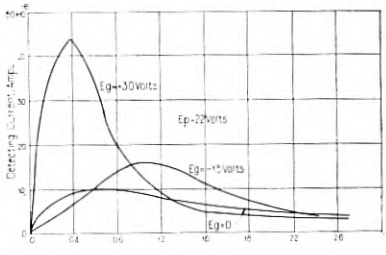


Fig. 51

The variations of detected current under heterodyne operation are shown in Figs. 51 and 52 which refer respectively to detection with and without a blocking condenser. The abscissa, e_1e_2, gives the product of the two high frequency amplitudes in the input. As is to be expected when no grid condenser is used (see Equation 13), the

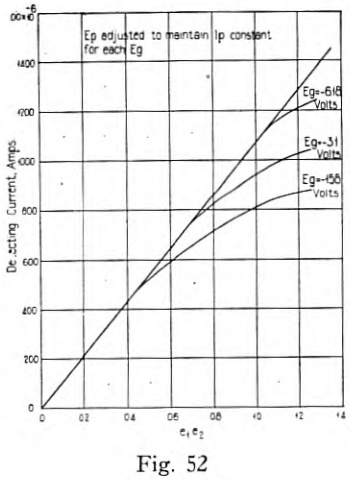


Fig. 52

variation of detected current with e_1e_2 is very nearly linear provided E_g is sufficiently negative. Fig. 51 referring to detection with grid condenser shows no linear relation however. The data for Figs. 51 and 52 are taken from Van der Bijl.

44. *Comparison of Tubes as Detectors.* If a tube is available whose detecting coefficient is known, other tubes may be calibrated in terms

of this standard. The comparison of detectors can be very readily carried out by means of such a circuit as shown in Fig. 53. This circuit makes use of a grid blocking condenser but could readily be rearranged not to employ it. In use, switches S and K are operated together in such manner that the receiver shunt is cut out when the

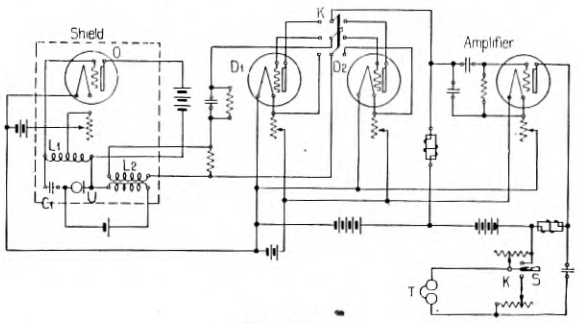


Fig. 53

receiver is connected to the tube of lower detecting power. By proper adjustment of the shunt the two tubes can then be brought to apparent equality, the difference being read from the calibration of the shunt.

IX. Vacuum Tube Oscillators

As pointed out in the section on amplifiers, it is easy to design feed-back circuits which will sing, i.e., will generate continuous oscillations. The necessary requirements which an oscillating circuit must meet are two in number and are readily understood. Any small alternating voltage when applied to the input generates a current in the output, and by virtue of the feed-back a portion of this energy is returned to the input. For continuous oscillations, the energy returned must be in phase with the original input supply. Furthermore, letting e represent the initial input voltage, the feed-back coupling must be sufficient to return to the input a voltage greater than e. If it is less than e the circuit will amplify but will not oscillate.

The circuit requirements necessary for any given tube to return a voltage greater than e may readily be stated in mathematical form, but so far as the practical design of oscillators is concerned, this statement has no particular value. The design of circuits is still very largely an empirical matter, and the problem is not so much to make the circuit oscillate as to make it oscillate with the proper frequency, efficiency and output power. These requirements can usually best be

met by trial and adjustment taking into account such general theoretical considerations as follows.

Vacuum tube oscillators make a convenient way of obtaining large high frequency currents at small voltages and large a.c. voltages at small currents for testing purposes.[36] A special oscillator circuit for giving a very pure sine wave output of constant frequency is discussed in Sec. 53.

45. *Equivalent Resistance of the Oscillator Circuit.* Oscillator circuits are of many types, but the fundamental action of all of them can be reduced to common terms, and to simplify the discussion the type of circuit illustrated in Fig. 54 will be discussed.

It will be noted that a d.c. voltage is applied between filament and plate by means of a battery in series with a choke coil. This choke

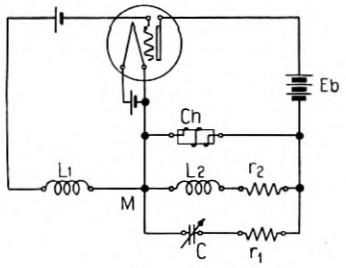


Fig. 54

may be considered as having zero d.c. resistance and virtually an infinite resistance to the a.c. generated by the oscillator. An oscillating circuit consisting of inductance, resistance and capacity is also connected between filament and plate, and by means of another inductance joining filament and grid a portion of the output energy is fed back to the input by virtue of the inductive coupling M.

In circuits as usually constructed the ohmic resistances r_1 and r_2 of the oscillating circuit are very small compared to the impedances of L_2 and C. It follows that the frequency of oscillation differs but slightly from the natural frequency $1/2\pi\sqrt{L_2C}$, and the oscillating circuit may therefore, be looked upon as introducing nothing more than a pure resistance (so far as the fundamental component of I_p is concerned) into the output circuit of the tube. Except for considerations of feed-back, we may therefore imagine the oscillating circuit L_2, C, replaced by an equivalent resistance whose value is given by the equation $R=\frac{r_1r_2+L_2/C}{r_1+r_2}$. This "equivalent resistance" is important in dealing with oscillating circuits.

[36] See W. C. White, Gen. Elect. Rev., Vol. 20, p. 635, 1917.

46. *Phase Relations.* Typical phase relations between the various currents and voltages in the oscillating circuit are illustrated by the oscillograms in Fig. 55. Note that the oscillation current I_o, and also E_p and E_g show practically sinusoidal variations. Such variations

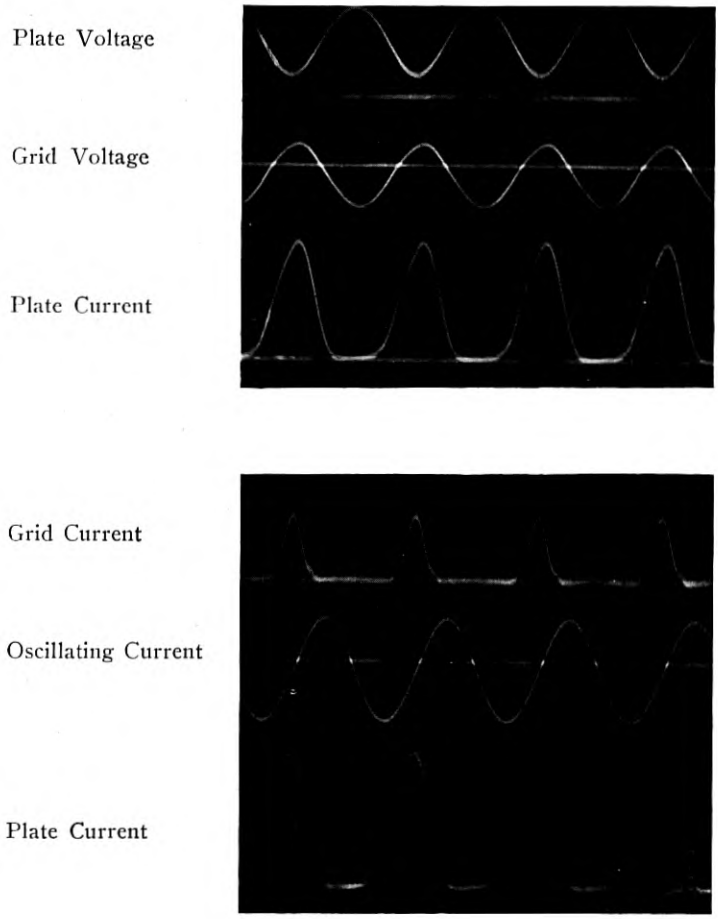


Fig. 55

will be found over very wide ranges of adjustment. Although I_p does not show a sinusoidal variation, the fundamental component (whose frequency is the same as that of I_o) is relatively much larger than the higher components.

Note that the variations of E_p and E_g are practically 180° out of phase, and also that E_p and the fundamental of I_p are 180° out of

phase. This latter condition is obviously required for maximum power output by an a.c. generator. In order that I_p be a maximum when E_p is a minimum, E_g must be a maximum at this time.

47. *Dynamic Characteristic.* A tube when operating into an output resistance follows a dynamic characteristic (also called derived characteristic)[37] whose slope is somewhat less than that of the static (see Sec. 12). The dynamic characteristic is flatter than the static characteristic since the oscillating circuit L_2, C acts as an equivalent resistance. In fact, the dynamic characteristic of the oscillating tube may differ in one important respect from that of Sec. 12, for in an oscillator the grid potential has a relatively very high positive value for a portion of the cycle. Consequently the dynamic plate characteristic very frequently turns downward at its upper end (point B, Fig. 56). As will be pointed out later, this feature of the dynamic

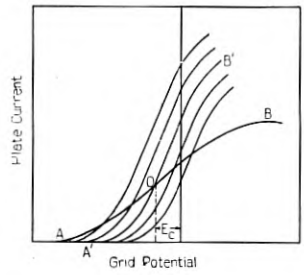


Fig. 56

characteristic is likely to be one of the factors tending to limit the amplitude of oscillation.

48. *Amplitude of Oscillation.* As yet no very comprehensive formula has been derived, either theoretically or empirically, to express the amplitude of oscillation in terms of the tube and circuit constants and the applied E_p. However, certain general statements can be made which will serve as useful guides.

We shall not consider further the circuit shown in Fig. 54; it shows a negative grid battery in the grid filament branch, which, if sufficiently large to control the oscillator when in operation, usually makes the starting difficult and uncertain, and is therefore undesirable. To eliminate the grid battery and yet supply sufficient negative potential, once oscillations have started, circuits are usually supplied with a grid blocking condenser and high resistance leak (see Figs. 64 and 65). Since for a portion of each oscillation the grid is positive, it

[37] See L. A. Hazeltine, Procd. Inst. of Radio Engrs., April, 1918.

picks up a charge of electrons which in flowing off through the leak creates an average negative grid potential. It is apparent that as the amplitude of oscillation increases, the charge picked up at each positive swing of the grid potential increases, with the result that this average negative potential tends to sink lower. The importance of this control feature will be brought out presently.

It is generally found that the oscillations build up to such a value that the greatest positive potential of the grid, which will be represented by $E_{g\max}$, becomes practically equal to the lowest plate potential, $E_{p\min}$. It is not difficult to understand why this condition should represent a sort of limit. It is usually found that the dynamic characteristic, as shown at B, Fig. 56, tends to bend rapidly downward as $E_{g\max}$ becomes greater than $E_{p\min}$. When $E_{g\max}$ tends to become greater than $E_{p\min}$ the electron current to the grid rises suddenly (see cathode ray oscillogram, Fig. 57) with the result that the average

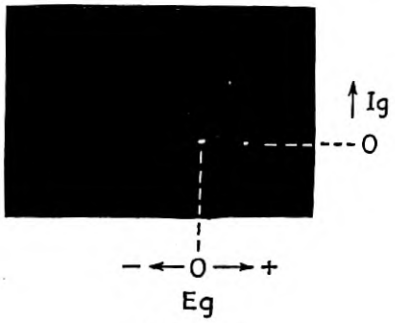


Fig. 57

current flowing through the grid leak increases very rapidly, which in turn results in a marked depression of the average grid potential.

The increased electron current to the grid as E_g tends to become greater than $E_{p\min}$ also represents a greater dissipation of energy upon the grid which is equivalent to an increase in the effective resistance R.

It is quite possible that any one of these three factors, in the absence of the others, would be sufficient to limit the oscillations; but as each springs into importance when the amplitude of the oscillation has reached about the same value, we shall not discuss the exact combination of the three which actually determines the amplitude.

Adopting the relation $E_{g\max}=E_{p\min}$ and making the additional assumption that the dynamic characteristic is straight (which is very nearly true) an expression for the amplitude of oscillation can readily

be obtained. Introducing certain approximations this expression may be written,

$$i_{p1}(R+R_o)=\frac{1}{\sqrt{2}}\,\bar{E}_p, \tag{14}$$

in which terms of the order $\frac{1}{\mu}$ are neglected in comparison to unity. i_{p1} is the fundamental component of the space current, $\bar{E}_p$ is the mean plate voltage and the term R_o is largely determined by the tube, and while not generally equal to r_p, is apparently not much different from it. This equation indicates that the amplitude of oscillation is practically independent of the value of μ.

The remarkably simple relation given by Equation 14 has been tested for a wide variety of circuits and tubes and has been found to hold with a very fair degree of accuracy. It may safely be taken as indicating quite approximately what response may be expected from any tube and circuit when operated at a particular applied E_p. The equation is likely to be more closely followed the more carefully the adjustment of circuit for maximum efficiency has been made.

The condition $E_{g\max}=E_{p\min}$ should not be considered as invariable. Adjustments can readily be made for which $E_{g\max}$ will either be ap-

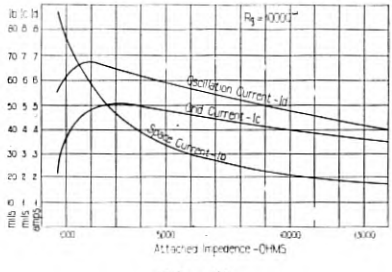


Fig. 58

preciably less or greater than $E_{p\min}$. It is generally found however that these adjustments do not give the most efficient operation.

Additional information as to the relation between the amplitude of oscillation and certain circuit constants are given in Figs. 58, 59, 60, 61. Fig. 59 shows that an oscillator tube may present a well-defined condition of temperature saturation. Figs. 60 and 61 show that the value of the feed-back voltage and grid leak resistance r_g may be varied within wide limits without affecting the output markedly.

49. *Efficiency.* The efficiency of an oscillator may be defined as the ratio between the energy of oscillatory current and the d.c. energy supplied to the plate circuit. This leaves out of account the energy

required to actuate the filament. The efficiency of oscillator circuits ranges all the way from a few per cent. to as high as 90% or better. The principal factors determining the efficiency are those which determine the amount of energy dissipated upon the plate of the tube. Inspection of the oscillogram, Fig. 55, shows that the sharper and

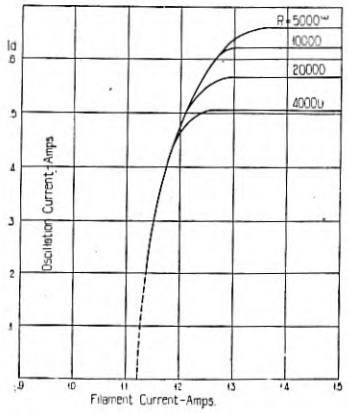


Fig. 59

narrower the plate current wave and the more nearly the plate voltage approaches zero, the higher will be the efficiency. In an extreme case such as the hypothetical one illustrated in Fig. 62 it is evident that the efficiency would be very large indeed. The μ of the tube

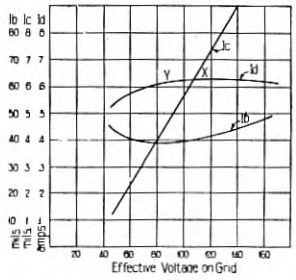


Fig. 60

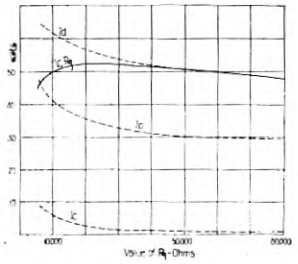


Fig. 61

largely determines the sharpness of the plate current wave and experience shows that for efficiencies of about 50% and better, μ should be at least ten; an increase above this value does not result in any very large improvement. It is also generally true that for the highest efficiencies the circuit constants should be so arranged that R is at

least four or five times as great as r_p, and it may advantageously be made 10 to 15 times as great. In these latter cases R_0 is relatively

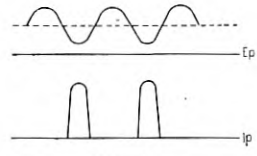


Fig. 62

negligible compared to R in Equation 14 with the result that i_{p1} can be quite accurately calculated although the exact value of R_0 may be unknown.

50. *Types of Oscillating Circuits.* There are many different types of oscillating circuits and as they do not lend themselves readily to classification, only a few of the more common types will be described.

One of the simplest oscillating circuits is that shown in Fig. 63 which is characterized by a tuned grid circuit inductively coupled

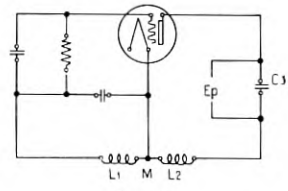


Fig. 63

to a coil in the output. This type is satisfactory for low plate voltages and small powers, but is not to be recommended where large amounts of power and high efficiencies are desired.

The condenser C_0 is inserted to prevent short circuiting of the plate battery or generator and may be made so large as to have no effect on the frequency.

A circuit of similar properties is shown in Fig. 64, the output being tuned instead of the input.

51. *Colpitts and Hartley Circuits.* Two very similar types of circuits which have proved satisfactory for a wide range of frequencies, voltages and powers, and which yield very high efficiencies, are shown in Figs. 64 and 65, the former being known as the Colpitts and the latter as the Hartley circuit. In both circuits, as illustrated, the mean grid potential is secured by a grid leak. The blocking condenser C_g should be large enough to offer very little impedance to the flow of the alternating current which causes the variation of the grid potential,

and as seen in Fig. 61 the resistance of the leak can be varied within wide limits without an appreciable effect upon the performance of the circuit. The source of plate potential may be either battery or generator. The choke in the generator supply should be sufficiently large to prevent an appreciable amount of oscillating current flowing

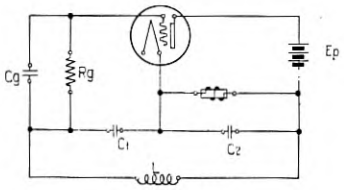


Fig. 64

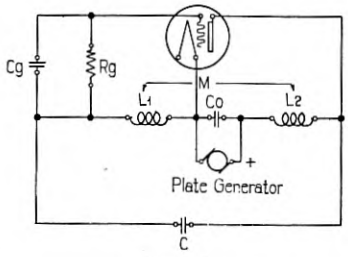


Fig. 65

through the generator. In order to avoid short circuiting the generator in the Hartley circuit, a large condenser C_o is interposed as shown in Fig. 65. For high frequencies it is customary to make this condenser so large that it has very little effect upon the frequency of oscillation.

The power of the oscillating current may be removed either by capacity or magnetic coupling, or it may be dissipated by resistance within the oscillating circuit. Magnetic coupling is very satisfactory for radio frequencies, but for the audio frequencies unless iron core transformers are used, capacity coupling usually proves more convenient. In regard to the use of iron core inductance and transformers, the reader should note the following Section.

52. *Frequency of Oscillation.* It is frequently found that the tube causes an appreciable deviation from the calculated frequency of oscillation. This is not so much because the tube influences the normal operation of the oscillating circuit as it is due to the fact that the impedances of inductance coils (particularly when iron cored) and of the condensers are liable to vary in value with the current amplitude. Hence any change in applied plate potential, filament current, or output resistance, because of its effect upon the amplitude of the oscillating current, will give rise to changes in frequency. In designing circuits to operate at a constant frequency regardless of slight changes in tube constants, plate potential, filament current, etc., the most important requirement is to provide inductances and condensers whose impedances are independent of the currents they carry, and whose resistance components are very small.

53. *Oscillator for A.C. Measurement Purposes.* For many a.c. measuring purposes an oscillator whose output is both free from harmonics and constant in frequency is desirable. Such a circuit is shown in Fig. 66. the design of which is radically different from the oscillator circuits already discussed. It possesses a tuned input LC and coupling is supplied by the resistance R. R is usually given a

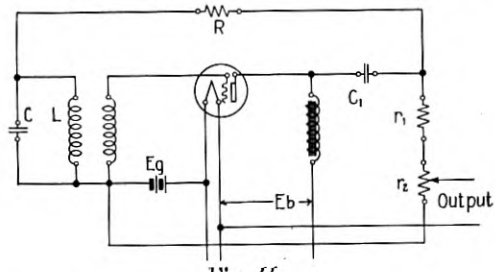


Fig. 66

value between 100,000 and 400,000 ohms. A negative grid battery fixes the average grid voltage, the emf, of this battery being about 8 to 10 volts. It is customary to make the resistances r_1 amd r_2 several thousands ohms apiece, r_1 being perhaps 5 times r_2. The condenser C_1 is merely a blocking condenser and should offer little impedance to the a.c.

It is not difficult to make the oscillator of Fig. 68 maintain a frequency that is constant within 3/10 of 1% and when the feed-back is not too large the harmonics in the output will comprise only 5% or even less of the total a.c. output.

54. *Range of Frequencies Obtainable with Vacuum Tube Oscillators.* Circuits have been constructed whose frequency is but a fraction

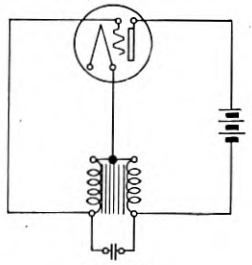


Fig. 67

of a cycle per second. The requirements of such a circuit are large inductance and capacity and very close coupling between input and output. A satisfactory circuit for low frequencies is that shown in Fig. 67; the two inductances taking the form of an iron core transformer.

At the other extreme, frequencies as high as 5×10^7 cycles per second can be obtained by means of tuned vacuum tube circuits of the Hartley or Colpitt type. At this point the coupling reactance of the tube becomes appreciable with that of the circuit.

Circuits capable of considerably higher frequencies have been described by Van der Pol,[37] Southworth,[38] Gutton and Touly,[39] and Holborn.[40] In all of these cases the oscillatory circuit is made up of distributed inductance and capacity connected to the tube in such a way as to utilize the capacity between the elements of the tube as a means of coupling.

The circuit shown in Fig. 68, when properly arranged, is as efficient as those used for lower frequencies and will give frequencies as high as 3×10^8 cycles per second. The oscillatory circuit is indicated by the heavy lines. It consists of a rectangle whose dimensions are appreciable with the wave length. Therefore, waves produced by variations in the electron emission through the grid are guided along the rectangle and are reflected at the ends. The reflected waves produce the proper voltage changes on the grid to sustain oscilla-

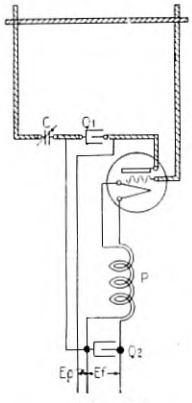


Fig. 68

tions. The ground imposed by the power leads places at least one point of the circuit at earth potential. If the condenser C is properly adjusted relative to the capacity between the grid and the plate the wave front can be made essentially perpendicular to the sides of the rectangle. It has been found that a large part of the power loss in the circuit is due to radiation. This circuit has been used as a basis

[37] B. van der Pol, *Phil. Mag.*, 38, July, 1919.
[38] G. C. Southworth, *Radio Rev.*, 1, Sept., 1920.
[39] Gutton and Touly, Comptes Rendus, 168, Feb. 3, 1919.
[40] F. Holborn, *Zs. fur Physik*, 6, p. 328.

of directive radio in which a metallic mirror was used to reflect the transmitted signals.

Very different means of producing high frequencies have been used by R. Whiddington,[41] Barkhausen and Kunz,[42] and by Gill and Morrell.[43] In some cases they employ tubes having considerable residual gas. The frequencies produced depend on the relative voltages applied to the grid and plate. Probably the best explanation of this phenomenon has been given by Gill and Morrell. Frequencies higher than 3×10^8 cycles per second have been reported.

The most accurate way of measuring these high frequencies is by observing the length of standing waves produced on a parallel wire system. The constancy of the vacuum tube generator, compared with spark ocsillators, combined with the fact that the sharpness of resonance in a parallel wire circuit is comparable with that in ordinary radio circuits, makes it especially adaptable to measurement purposes. It may be used, for example, to measure small inductances and capacities or to determine the dielectric constant of liquids. Many of the corrections necessary when a damped source is used are eliminated.

55. *The Mechanically Coupled Oscillator.* In addition to the types of oscillators described above where the frequency is determined by

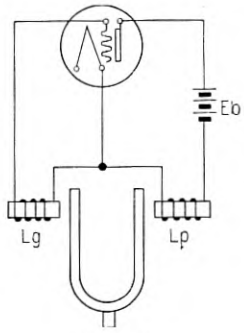


Fig. 68a

inductance and capacity, we may have oscillators in which the frequency is governed by a mechanical system such as a pendulum or a tuning fork.[44] An example is shown in Fig. 68a. The two coils

[41] R. Widdington, *Radio Rev.*, 1, Nov., 1919.

[42] Barkhausen and Kunz, *Phys. Zs.*, Jan, 1, 1920.

[43] Gill and Morrell, *Phil. Mag.*, 44, July, 1922.

[44] See Eckhardt, Karcher and Keiser, *J. O. S. A. & R. S. I.*, Vol. 6, p. 948, 1922; Eccles & Jordan, Phys. Soc. Proc. 31, Aug., 1919 and Phys. Soc. Proc. 32, Aug., 1920; Abraham & Bloch, J. d. Physique, Vol. 9, July, 1920.

L_p L_g are inserted in the plate and grid circuits of the tube. Variations in the plate current through the coil L_p impress forces on the tuning fork which result in its motion. This motion of the fork causes variations in the magnetic field through L_g and induces a varying voltage on the grid. With the proper coupling, sustained oscillations result having a period very nearly that of the tuning fork.

The electrically driven tuning fork described above constitutes a very satisfactory source of either sound or electromotive force. Horton, Ricker and Morrison [45] have made improvements which make it

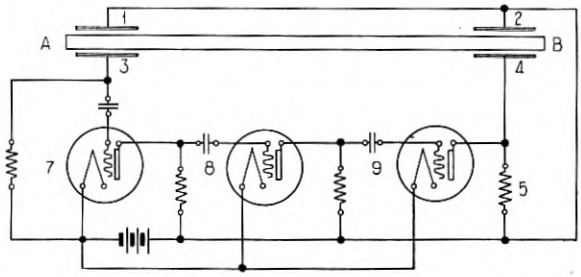


Fig. 68b

constant in amplitude and frequency to six parts in a million over very long periods of time.

An entirely different form of mechanical coupling has been used by Cady.[46] The circuit is shown in Fig. 68b. It makes use of the piezo-electric effect and mechanical vibrations of a crystal. Variations in the plate current in the tube 9 cause a voltage change across the resistance 5. This is communicated to a crystal AB such as quartz by means of the plates 2 and 4. A transverse electric field applied to such a crystal causes a change in its length. If this electric field be periodic, compression waves will travel along the crystal with a velocity depending on its density and elastic properties. These waves will, in turn, cause a varying electric field between plates 1 and 3 which may be communicated to the grid of the tube 7, amplified by 8, and finally transmitted to tube 9. This provides conditions for sustained oscillations having a frequency which is roughly inversely proportional to the length of the crystal.

Cady describes oscillators ranging in frequency from 3×10^4 to 10^6, and states that the frequency is constant to about one part in 10,000. The effect of temperature change is not great.

[45] *Journal of A. I. E. E.*, 1923.
[46] Cady, Proceedings of I. R. E., Vol. 10, No. 2, April, 1922.

X. Miscellaneous Applications of Thermionic Vacuum Tubes

56. *The Tube as a Voltmeter.* The three-element tube may be used for the measurement of either d.c or a.c. voltages. In the case of d.c. voltages it is customary to apply the unknown voltage to the plate, counter-balancing this voltage with a known negative potential applied to the grid. Given the μ of the tube, it is then possible to calculate with a fair degree of accuracy the plate potential. The usual procedure is to adjust the negative grid potential to such a point that the plate current just becomes zero. The tube when used in this manner becomes an electrostatic voltmeter, and it is evident that to give accurate readings the tube should have a well-defined cutoff (see Sec. 8, Fig. 10).

In a somewhat similar fashion a.c. peak voltages may readily be compared with known d.c. voltages. A typical circuit is shown in Fig. 69. In operation a fixed plate voltage is applied to the voltmeter

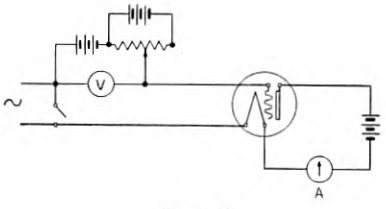


Fig. 69

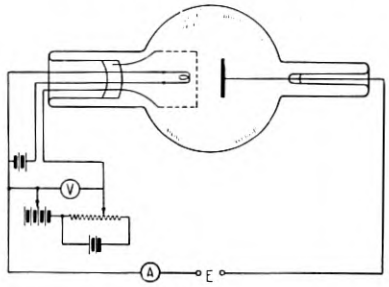


Fig. 70

tube and a steady negative d.c. voltage is applied to the grid which is just sufficient to reduce the plate current to zero. The a.c. voltage is then superimposed in the grid circuit with the result that current flows during the positive halves of the wave. If now the steady grid potential is made more negative until the plate current again just ceases to flow, it is apparent that this change in the steady potential just equals the peak value of the a.c. voltage.

For the measurement of very high voltages a special tube of the design shown in Fig. 70 will be found desirable, the grid being in the form of a screen which surrounds the filament. Such a tube may have a μ as high as 200.

A circuit similar to Fig. 69 may be so employed that the a.c. voltage to be measured causes a change in the space current meter reading, the negative grid potential being preferably so set that the conditions discussed in Sec. 16 are satisfied. The tube, due to its curved char-

acteristic, acts as a detector and as pointed out in Sec. 37, the change in space current is approximately proportional to the square of the a.c. input voltage. For accurate work the circuit requires calibration but the calibration will in general remain good over long periods of time. This method is particularly useful for small voltages.

57. *Power-Limiting Devices.* As pointed out in Sec. 4 the total emission from the filament at a given temperature is fairly sharply defined regardless of the plate voltage, so long as this exceeds the value required to give voltage saturation. The fact that the total emission is limited by the temperature may be used to control the

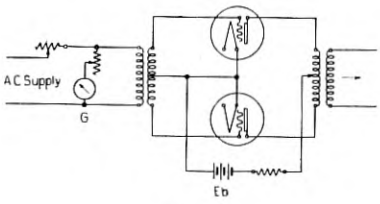


Fig. 71

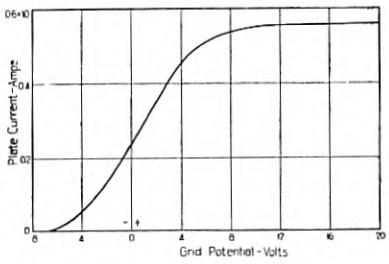


Fig. 72

maximum current in a circuit. As an illustration, Fig. 71 shows its application to an alternating current circuit, the performance of which is illustrated in Fig. 72. The introduction of such a device into an a.c. circuit will, of course, result in the generation of harmonics and may therefore, be objectionable.

There is almost no limit to the number of regulatory circuits which can be devised to employ the three-electrode tube.

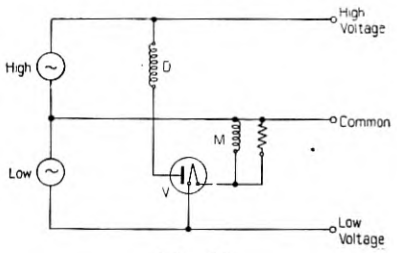


Fig. 73

58. *Voltage and Current Regulation of Generators.* The two-electrode tube with tungsten filament has been used to great advantage as a voltage regulator for a special airplane generator designed to deliver both 28 volts and 300 volts. The circuit arrangement is illustrated

in Fig. 73 in which M represents the main field winding, and D the differential winding which opposes M. The high voltage given by the generator when applied to the plate of the valve is sufficient to produce a condition of voltage saturation. As the speed of the generator increases, the current through the main field winding and

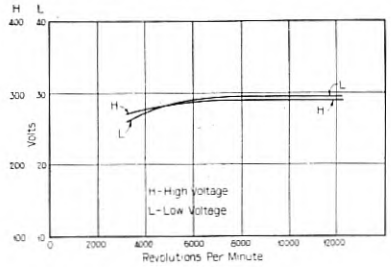


Fig. 74

the filament increases, thereby giving rise to greater emission from the filament and a larger current through the differential winding. It was found possible to so design the valve as to yield the very close regulation illustrated in Fig. 74.[47]

The three-electrode tube can also be used as a voltage regulator for a generator as shown in Fig. 75. It is apparent that an increase

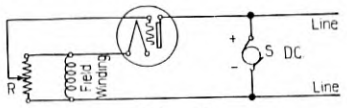


Fig. 75

in the voltage across the line tends to increase the current through the tube and resistance R. This in turn lowers the grid potential and tends to prevent an increase in current through the field winding.

The circuit shown in Fig. 76 illustrates an arrangement for main-

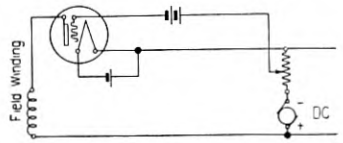


Fig. 76

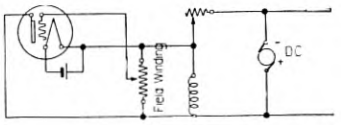


Fig. 77

taining a constant current from a generator. The operation of the device is apparent.

Fig. 77 shows another arrangement for maintaining a constant generator voltage. In this circuit an increase in voltage tends to

[47] Radio Telephony by Craft & Colpitts, Trans. A. I. E. E., Vol. 38, p. 330, 1919

make the grid less negative, thereby reducing the resistance shunted across the field winding.

A somewhat similar arrangement can readily be applied to regulate the voltage delivered by a battery. Fig. 78 illustrates such a circuit. An increase in E_1 raises the grid potential, thereby increasing the current through the tube and the resistance r_3. By a choice of regulating tube and resistances such that

$$r_3 = \frac{r_1 + r_2}{r_1} \frac{dE_g}{dI_p},$$

it may readily be shown that the voltage E_2 remains constant. Since the regulation effected by this circuit is independent of frequency it

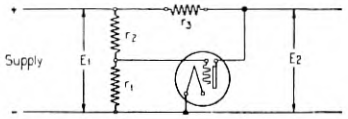


Fig. 78

may also be applied to a generator supply for elimination of commutator noise as well as voltage fluctuations due to changes in speed.

59. *The Ionization Manometer.* When gas is present in a three-electrode tube in quantities not sufficient to seriously affect the activity of the filament, and the plate voltage exceeds a value sufficient to produce ionization by collision, it has been found that the number of ions produced is proportional both to the pressure of the gas and to the electron current passing through the gas to the anode.[48] If now a small negative potential be applied to the grid, a certain fraction of the positive ions will be drawn to it and their number can be accurately measured by the current flowing in the grid circuit. The best arrangement is to apply the positive potential, not to the plate in the usual fashion, but to the grid, and apply the negative potential to the plate making it the collector of the positive ions. Dimensions of a satisfactory tube are given in Fig. 79. The values $E_g = 110$ volts and $E_p = -2$ volts have been found to give very satisfactory results, the electron current being .02 ampere, and K being equal to 0.10 for nitrogen and having approximately this value for air. The gauge equation may be put in the form,

$$P = K \frac{I+}{I-},$$

in which P is the pressure, K a constant depending upon the design

[48] O. E. Buckley, Procd. Nat. Acad., Vol. 2, p. 683, 1916.

of the tube, $I+$ the positive ion current, and $I-$ the electron current.

As it is necessary to know the value of $I-$, and since the emission from the filament is liable to vary somewhat with the kind of gas and its pressure, it will be found advantageous, if many readings are to be made, to place in the $I-$ circuit, the coils of a relay which is adjusted to close at a definite value of $I-$, and which, when closed, cuts in a shunt around the filament which will reduce its heating current.

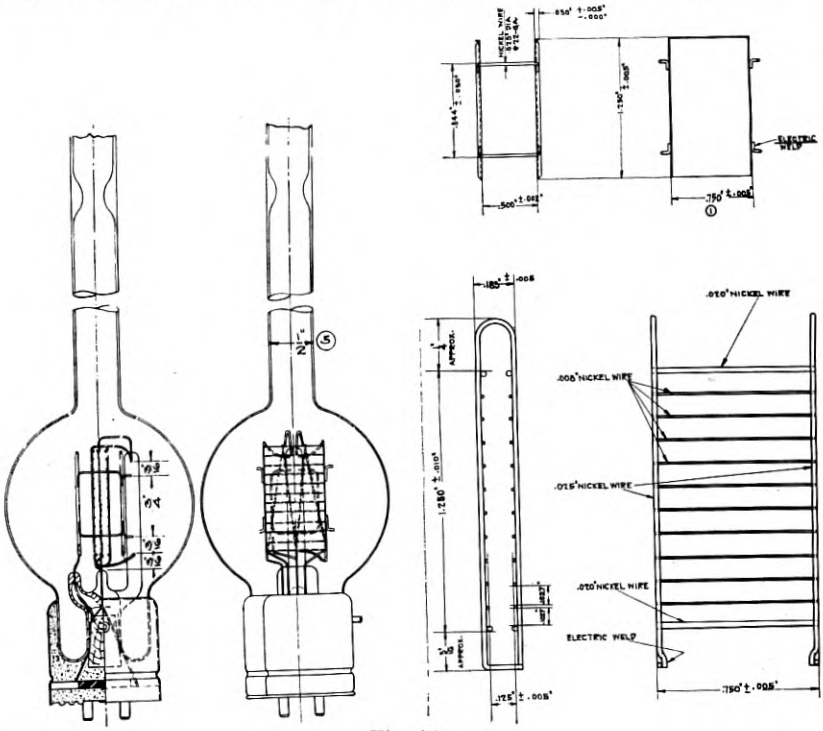


Fig. 79

Automatic regulators of this type have been used with complete success. Experiment shows that the value of K remains constant for pressures as high as 1.5×10^{-3} mm. of Hg. and the lower limit is determined very largely by the sensitivity of the current reading instruments. It follows that an ionization gauge can be calibrated by comparison with a McLeod gauge and, in use, extrapolated to low pressures.[49]

60. *Heterodyne Method of Generating Currents of Very Low Frequencies.* By impressing upon the grid of a detector tube two frequencies which differ by a very small amount (e.g., 99 and 100 cycles),

[49] Dushman, *Phys. Rev.*, Oct., 1920, p. 854.

it is possible to obtain from the output of the detector the difference frequency of one cycle per second. This low frequency may be readily separated by means of a filter. It is apparent, however, that to maintain this difference frequency constant requires that the input frequencies be held within a very narrow range of variation.

61. *Thermionic Valve as a High Tension Switch.* If the plate circuit of a valve is inserted in a high tension circuit, the flow of current in the circuit may readily be stopped by cutting off the filament heating supply, thus making unnecessary the breaking of any contacts in the high tension circuit. In case the transmission of current in both directions is necessary, two valves may be used.

62. *Devices Employing Secondary Emission.* As pointed out in Sec. 9, the grid current in a three-element tube shows a negative resistance characteristic for a certain range of voltage, and various uses of this fact have been pointed out.[50]

63. *Electron Tube Oscillograph.* A special type of thermionic tube designed for oscillographic uses is of great importance as a laboratory instrument. These tubes, using the hot filament as a source of electrons, have certain marked advantages over the Braun tube with its gaseous discharge.[51] One of the very successful thermionic oscillographs has the following properties: anode potential 300–400 volts, sharp focus of electron beam, sensitivity of 1 mm. per volt between deflection plates and 1 mm. per ampere-turn when using magnetic deflection. Photographic recording is possible with relatively short exposures by using suitable fluorescent material.

[49] Dushman, *Phys. Rev.*, Oct., 1920, p. 854.
[50] See footnote 21.
[51] See J. B. Johnson, *J. of Opt. Soc. of Amer.*, Sept., 1922, or *Bell System Technical Journal*, Nov., 1922.

Some Contemporary Advances in Physics

By K. K. DARROW

NOTE: Dr. Darrow, the author of the following article, has made it a practice to prepare abstracts and reviews of such recent researches in physics as appear to him to be of special interest. The results of Dr. Darrow's work have been available to the staffs of the Bell System laboratories for some time and having been very well regarded, it is thought that such a review, published from time to time in the TECHNICAL JOURNAL, might be welcomed by its readers.

The review cannot, of course, cover all the published results of physical research. The author chooses those articles which appear significant to him or instructive to his readers, without attempting to pass judgment on the scientific importance of the different papers published. It is not intended that the review shall always assume the same form; at one time it may cover many articles, at another be devoted to only a few, and it may occasionally treat of but a single piece of work.—*Editor.*

SOME years ago C. T. R. Wilson of Cambridge University developed a beautiful method for making the paths of moving charged atoms and electrons individually visible. The charged particle flies through a gas such as air, mixed with water-vapor; it ionizes many of the molecules near which it passes; the gas is suddenly cooled by expansion and the water-vapor is precipitated upon the ionized molecules, forming a trail of droplets which visibly mark out the path of the ionizing electron or atom. Truly spectacular photographs of such trails, thick straight ones of fast-moving atoms and thin curly ones of electrons, are frequently published in textbooks and in popular articles.

The method is now proving very powerful in the study of collisions and close encounters of electrons with atoms and of atoms with atoms. Rutherford having found by another method that the nuclei of atoms are occasionally broken up by unusually direct blows from fast-moving helium nuclei (alpha-particles), the prospect of actually photographing such an important event becomes alluring. However, it is a very rare event; for W. D. Harkins and R. W. Ryan of the University of Chicago photographed eighty thousand alpha-particle trails in air, and only three of the particles struck molecules so squarely as to be deflected through more than a right angle; and of these only one showed indications of having broken the nucleus it struck. This particular collision is shown in Fig. 1 (two photographs of the same encounter taken from different directions at the same moment). In addition to the tracks of the alpha-particle up to and away from the scene of the encounter, there are two more tracks diverging from it, which are probably the tracks of two fragments of the struck nucleus. Other interpretations, such as two distinct impacts very near together or a stray radioactive atom

happening to disintegrate just as the alpha-particle passed by, are admissible but highly improbable. Another such collision in argon is shown in Fig. 2; this too was the only encounter with four diverging

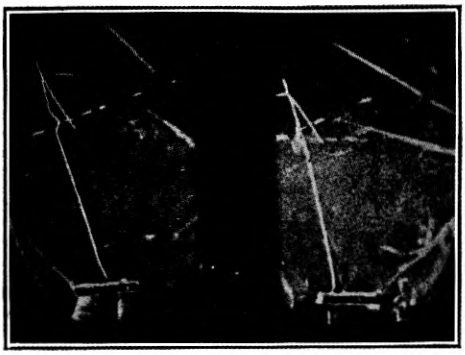


Fig. 1

tracks observed in many thousand photographs with the same gas.[1] A collision in air, in which the struck nucleus was not broken, but knocked to one side while the alpha-particle rebounded in the manner demanded by the principle of conservation of momentum, is shown

Fig. 2

in Fig. 3. These results show how small the atom-nuclei must be, compared to the extension of their electron-systems; for the 80,000 alpha-particles observed in air had traversed the electron-systems of about ten billion molecules altogether.

[1] As Rutherford's experiments indicate that argon atoms are especially stable against disintegration, this may be a case of two consecutive collisions with adjacent atoms.

Fig. 4 shows curious collisions of alpha-particles passing through helium gas, photographed by D. Bose and S. Ghosh of Calcutta. In each of the two left-hand trails the alpha-particle has apparently

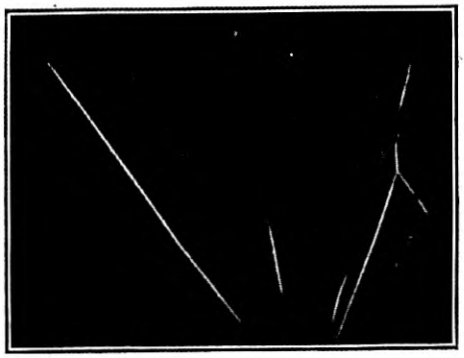


Fig. 3

knocked the nucleus and the two electrons of the atom in three different directions.[2] The alpha-particle of the right-hand trail (*iiib* is a magnification of *iiia*) seems to have produced quite an explosion; this may be the disruption of a nucleus belonging to a stray molecule

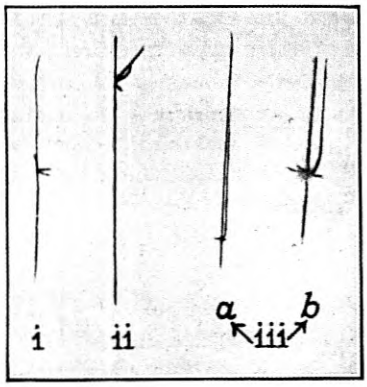


Fig. 4

of nitrogen, but one would not expect the original particle to go on as if unaffected.

[2] In hydrogen they found no cases of two electrons being driven off by the same impact. This agrees with Millikan's conclusion that double ionization is much more frequent with helium atoms than with molecules of any other kind, if indeed, it is not a characteristic of helium alone.

Much work is now being devoted to the spectra of ionized atoms. One instance, that of ionized potassium, will suffice to illustrate the problem. The potassium atom differs from the argon atom in three respects; the weight of its nucleus is slightly different, which is probably inessential; the charge on its nucleus is 19/18 as great; and it has a nineteenth electron outside of the three closed electron-shells, comprising eighteen electrons, which by themselves constitute the whole electron-system of the argon atom. When this outermost electron is removed, we have a system which probably differs from the argon atom in only one essential respect—that the central nucleus has a somewhat larger attractive power and hence the three electron-shells are somewhat more drawn inward. The spectra of ionized potassium and of argon should therefore be very nearly alike. This has been tested by Zeeman and Dik at Amsterdam; the result is very satisfactory, and the difference between the simple and clearly-arranged spectrum of potassium on the one hand, and the rich and intricate spectra of ionized potassium and argon on the other hand, is very striking. At Bonn, the spectrum of ionized rubidium is being compared with that of krypton for the same purpose.

The most extensive results, however, have been obtained by Fowler with silicon. For some reason or other, silicon is a particularly easy element from which to obtain spectra not only of the neutral and the ionized atom, but also of the twice-ionized and thrice-ionized atom—four distinct spectra, one from neutral silicon, the next from an atom resembling aluminium, the next from an atom resembling magnesium, and the last from an atom resembling sodium. These four spectra can be observed in the stars and in the laboratory, some of the important lines from thrice ionized atoms having been photographed by Millikan in the extreme ultra-violet. In their general type, they resemble the spectra of the neutral atoms corresponding in structure to the atoms which emit them.

Data have also been made available for doubly-ionized magnesium (by Paschen) singly-ionized magnesium, and for neutral sodium—three atoms in which the nuclear charge is respectively 13e, 12e, and 11e, while in each of them the nucleus is surrounded by ten electrons and there is an eleventh one much further out. This eleventh electron being responsible for the spectrum and being relatively exempt from perturbations due to the other ten, the spectra of these three atoms are of the simplest and clearest type. The series-lines which in the spectrum of neutral sodium are in the inaccessible infra-red are moved up, in the spectrum of doubly-ionized aluminium, into the visible region. Further study of spectra related to each other in this

manner, and differing by virtue of slight intelligible differences in the atoms which emit them, may be expected to help greatly in making clear the major features of atomic structure.

Two phenomena, first accurately examined by A. H. Compton, afford a striking illustration of the way in which classical electromagnetic theory and quantum theory are alternately successful in explaining the qualities of radiation. On the one hand, Compton has been the first to apply accurate wave-length measurements to scattered X-rays, and finds that they are a mixture of two kinds of X-rays—one having exactly the same wave-length as the primary X-rays, the other a wave-length slightly greater and varying with the angle between the primary and the secondary rays. According to the classical theory, scattered X-rays are simply radiation sent out in all directions by electrons inside the atoms of the scattering substances, vibrating under the influence of the primary X-rays, and hence vibrating necessarily with the same frequency as the primary X-rays. This could account for one of the components of the scattered X-rays, but not the other. The other can be accounted for by assuming that the primary X-ray quanta of frequency n are perfectly elastic spheres which travel with the velocity of light, have momentum hn/c and energy hn, and collide with the atoms just as one elastic sphere collides with another (that is, under conditions of conservation of translatory kinetic energy and of momentum); they depart from the collision with less energy and less momentum than they initially had, and consequently with a diminished frequency. But this does not explain the first-mentioned component, leaving the two theories balanced. On the other hand, in the *Philosophical Magazine* paper, Compton describes the total reflection of X-rays by glass, silver and lacquer—a phenomenon of exactly the type which the classical theory explains far more easily and naturally than the quantum-theory.

In glass and lacquer, the highest natural frequency of any of the electrons in any of the constituent atoms—to speak the language of the classical theory—is far below the frequency of available X-rays; we are, in optical terminology, on the high-frequency or anomalous-dispersion side of the highest-frequency absorption-band; the well-known dispersion formula reduces to a single term,

$$\mu = 1 - Ne^2/2\pi mn^2$$

where N is the total number of electrons able to vibrate in unison with the X-rays, and μ is the index of refraction of the X-rays of

frequency n; e and m have their usual meanings. The index of refraction is less than unity, the X-rays travel faster in glass or in lacquer than in air or in vacuo, and are totally reflected from a glass surface if incident at a sufficiently small angle with the surface. The agreement between experiment and theory is, quantitatively as well as qualitatively, very good. It is equally good for silver, allowance being made for the fact that the frequency of the X-rays used lay between the two absorption-bands of silver. It seems conceivable that this might be refined into a method for determining the numbers of electrons in different orbits of the atom.

The atoms of the inert or "rare" gases argon, krypton, and xenon are almost completely transparent to slow electrons—electrons moving with a speed of one or two equivalent volts. In more exact language, the radius of the effective cross-section of one of these atoms relatively to slow electrons is much smaller than its radius relatively to faster electrons or to other atoms. This almost incredible statement, having been tested by several different experimenters and by at least two entirely distinct methods, now appears to stand beyond doubt. This radius of the effective cross-section of the atom, relatively to an electron, is (by definition) the least distance at which the electron can pass by the centre of the atom without being intercepted or deflected; the radius of the atom relatively to another of the same kind is, naturally, half the least distance at which the centres of the two atoms can pass each other without affecting one another's paths. The concept is not perfectly exact, depending as it does on what we choose to take as the least perceptible alteration of the path of a particle; nevertheless, it is practicable and useful. Years ago the radius relative to other atoms was determined (from the viscosity of the gas). There is no binding reason why it should be identical with the radius relative to electrons, but the first measurements of this latter quantity on such gases as hydrogen, nitrogen, and helium yielded fairly good agreements between the two. Recent measurements on argon disclosed a surprising difference.

The method consists essentially in measuring the fraction of a beam of electrons, projected against a layer of gas, which pass through the layer undeflected. (Another and entirely different method used by Townsend resulted in a valuable confirmation of the result.) If there are N atoms under unit area of the surface of the layer (looking through it in the direction from which the electrons come) and N is not so large that many of the atoms are partly shielded, in the perspective, by others, the fraction of the electrons which go through undeviated is $(1-N\pi r^2)$; r being the radius just defined. The most

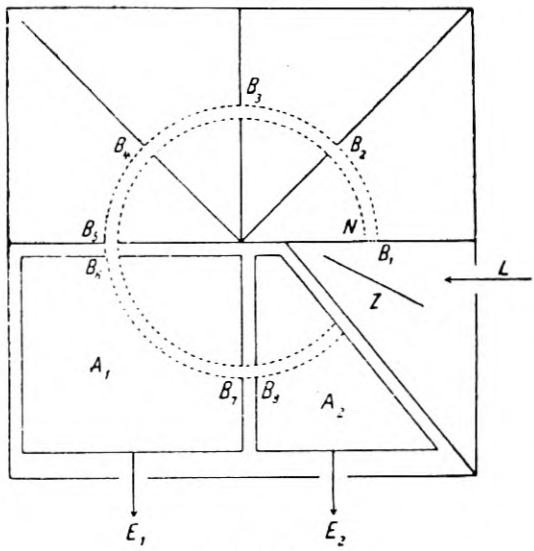

B3
B4
B2
B5
N
B1
B6
L
Z
A1
B7
B8
A2
E1
E2

Fig. 5

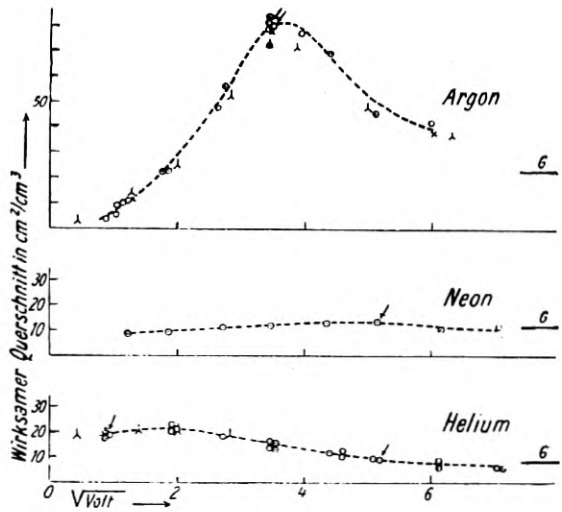

Argon
Neon
Helium
Wirksamer Querschnitt in cm²/cm³
50
30
20
10
6
0
V Volt
2
4
6

Fig. 6

delicate arrangement is that of Ramsauer (Fig. 5); the electrons enter at B_1, and are steered by a magnetic field along a semicircular path through the slits B_2-B_8; electrons deviated even through a very slight angle go against the partitions and are not received by the electrometers connected at E_1 or E_2. Measurements with the two

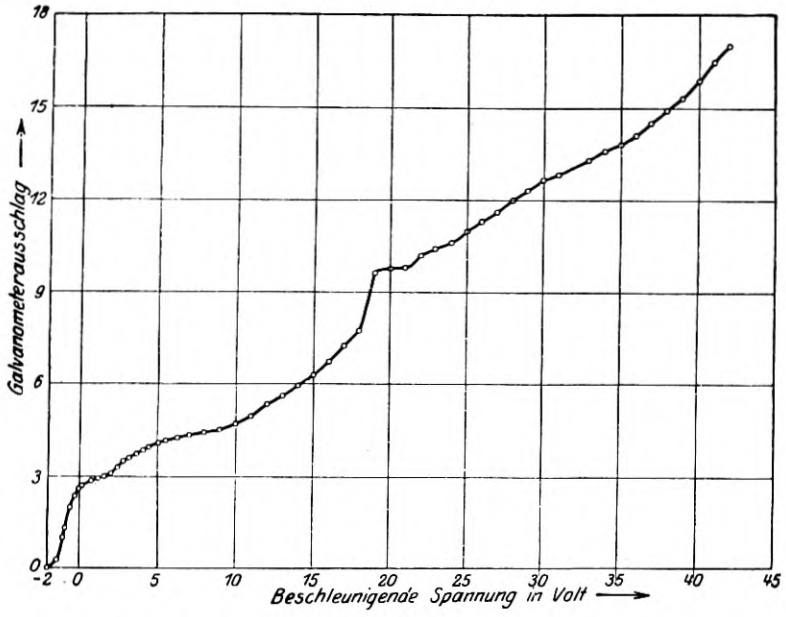


Fig. 7

electrometers, at two pressures of the gas, yield the data required. The values of r^2 thus determined for argon, neon and helium are plotted against the speed of the electrons in Fig. 6. The ordinates of the short straight horizontal lines on the right represent the squares of the radii relatively to other atoms.

The effect shows itself, however, very distinctly in a much simpler and more common device, a cylindrical three-element tube of audion type with the grid very much closer to the filament than is the plate; the plate is maintained at a potential a fraction of a volt higher than that of the grid. Figs. 7 and 8, from a recent article by Minkowski and Sponer, exhibit curves of plate-current versus grid-voltage in helium, which does not show the effect in question, and argon, which does.[3] In helium the current rises steadily as the increasing voltage

[3] The displacement of the curves by about −2 volts along the axis of voltages is probably due in part to drop of potential along the filament, in part to neglected contact-potential-differences.

gradually overcomes the space-charge repulsion, augmented in the gas by the reflection of electrons, for the reflected electrons stay longer in the space between filament and plate than they would if they went straight through. In argon the curve rises at first more swiftly, almost or quite as steeply as in vacuo, for the atoms are almost transparent to the electrons when they are slow; but as their speed is increased and the effective radius of the atom rises, the current sharply declines again. Further on, near 11 volts, there is

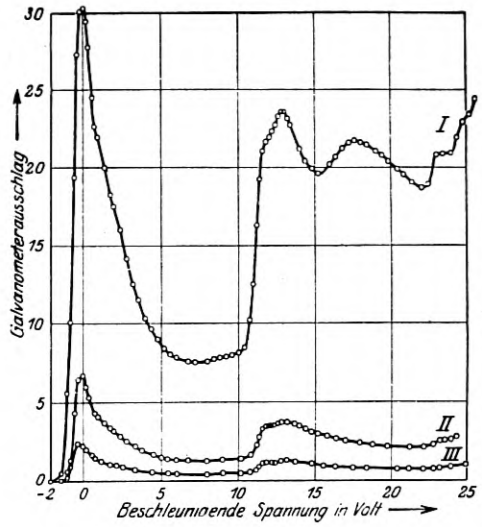


Fig. 8

another peak; near this voltage, the electrons which collide with atoms lose almost all their energy (threshold-speed for inelastic impact at 11.3 equivalent volts) at the first collision, and pass through the rest of the gas-filled region without obstacle. A second peak near 16 volts is ascribed to a second critical speed for inelastic impacts. Krypton and xenon give toothed curves of the same general type. Neon and mercury vapor, however, behave like helium, the curves rising steadily or at most showing slight kinks and inflections which may be indications of a slight effect of the same sort.

The reason for this remarkable effect is still obscure. It may be possible to devise an atom-model adequate to explain it without forfeiting spherical symmetry, which it is desirable to retain if possible, for among all atoms these of the heavy inert gases would be

expected to display the most complete symmetry and smoothness. F. Hund tried to devise an atom such that any one-volt electron passing within a distance r of the centre would be deflected through exactly 360° before coming out; he attained a formal solution of the problem, but the model involved a continuous distribution of negative charge from the nucleus outward to the distance r, which is quite incompatible with all our other knowledge of atomic structure. H. A. Wilson of Rice Institute tried out a well-known and popular model, consisting of a nucleus surrounded by a spherical surface of radius r, over which negative charge equal in amount to the positive charge on the nucleus is uniformly spread. For very slow electrons, the average angle of deflection is 90°; it increases with speed, becoming 180° at a certain critical value v_0 at which every electron is turned back into the direction whence it came; beyond v_0 it decreases indefinitely with increasing speed. At v_0 the oncoming electrons are more radically deflected by the atoms, so to speak, than at any greater or lesser speed; below v_0 the variation of mean deflection with speed is in the proper sense to agree with experiment, but not by any means of a sufficiently great order-of-magnitude. The theory, however, seems to explain the mild variations encountered in such gases as hydrogen and nitrogen, and the explanation of the more striking ones may lie in the same direction.[4]

Important contributions have lately been made to our knowledge of self-sustaining discharges, such as the glow and the arc, which maintain themselves as long as the proper voltage is applied at the electrodes, without requiring the assistance of a separate source of ions such as a hot filament or an outside ionizing agency such as X-rays. The field has perhaps been somewhat neglected, because it is easier to obtain simple clear results with electrons and ions admitted into a very rarefied gas after being generated elsewhere. In a self-sustaining discharge, there is usually a sudden steep potential-drop just in front of the anode, and another just in front of the cathode—the so-called anode-fall and cathode-fall; in the region between, the potential varies gradually. The anode always tends to become very hot, and Gunther-Schulze at Berlin has lately measured the rate at which heat is generated at the anode of a mercury arc; he finds that it agrees wonderfully well with the rate calculated from the assumption that practically the entire current is carried by negative ions

[4] Any competent theory must explain the results obtained by different methods, notably the fact that the value of r measured in an apparatus like Ramsauer's agrees with the value measured in an apparatus in which electrons deflected through considerable angle should yet reach the collector, and so be counted as though they had not been deflected at all.

(probably electrons) which dash against the anode with the entire kinetic energy acquired during unobstructed passage through the anode-fall.[5]

The heat generated at the cathode must arise in the converse way, from the kinetic energy of positive ions pulled violently against the cathode surface. K. T. Compton of Princeton, has made elaborate calculations for the arc in air with a carbon cathode and the arc in hydrogen with a tungsten cathode, and comparing the results with the experimental evidence, concludes that a few per cent of the current at the cathode is carried by positive ions, the remainder by electrons moving away from the cathode. Compton then attacked the same problem in an entirely different manner; he assumed that the region near the cathode, in which the cathode-fall occurs, is a region in which positive ions are moving gradually towards the cathode, accelerated by the field, and retarded by their collisions with neutral molecules and by their mutual space-charge repulsion. The problem is formally similar to that of determining the current-voltage relation in a thermionic vacuum-tube, and the solution is a relation between cathode-fall, current, and width of the region in which the cathode-fall occurs. The first quantity is known; the third is assumed to be the mean distance which an electron travels from the cathode before striking a molecule; the second quantity, the current of positive ions into the cathode, comes out to be a few per cent of the observed total current. These two methods thus support one another in indicating that in the arc-discharge some 90–98% of the current near the cathode is carried by electrons, and the small remainder by positive ions. In the glow-discharge, according to experiments by Gunther-Schulze, the rate at which heat is generated at the cathode is 25% to 75% of what it would be if all the current were carried by positive ions, falling against the cathode with the entire energy derived in passing through the cathode-fall. Expecting that a much larger fraction of the energy of the positive ions would be dissipated in collisions with neutral gas molecules, he concludes that the region of the cathode-fall must be a region in which the gas is abnormally rarefied because abnormally hot; the hotness in turn being due to the collisions between ions and molecules.

In the central region of the arc, the potential-gradient is uniform and consequently the positive and negative charges per unit volume

[5] It is obviously necessary to be very cautious in making deductions of this kind, for the entire energy iV (i representing the current and V the anode-fall or cathode-fall, as the case may be) is dissipated as heat in the region of the anode-fall or cathode-fall; and if this region is very narrow it is hard to distinguish between heat generated within it and heat generated at the anode or cathode surface.

must exactly balance one another. In this region Compton suggests that the gas is in the state of thermal ionization defined and described by Saha, in which at all times a certain constant percentage of the atoms, depending only on their ionizing-potential and on the temperature, is ionized. If the temperature of the central region of the carbon arc is about 4000°, and the ionizing-potential of the gas about 8 volts, the proportion of ionized molecules will be about right.

According to one of the newer and stranger developments of the quantum-theory, an atom possessing magnetic moment and submerged in a magnetic field is not at liberty to orient itself in any direction whatever, not even momentarily; it may set itself only at certain specified inclinations, such that the cosine of the angle between the direction of its magnetic axis and the direction of the field will have one of certain specified values. Imagine for example, an atom consisting of a single electron revolving in a one-quantum orbit (the smallest possible orbit) about a centre which itself is not magnetic; such a centre might be a simple nucleus, or a nucleus surrounded by a number of electrons moving in orbits so inclined to each other that their magnetic moments cancel one another out. The magnetic moment of such an atom is $eh/4\pi m$ (e the charge and m the mass of the electron); its magnetic axis is perpendicular to the plane of the orbit of the electron. According to the theory, the magnetic axis must point exactly with or exactly against the magnetic field; the cosine of the angle must be $+1$ or -1. This was verified last year by Gerlach, who projected a ray of silver atoms (shooting off from a hot rapidly-evaporating silver filament through a small hole) across a magnetic field with an extremely steep field-gradient. The ray divided itself into two, one consisting of atoms with their north magnetic poles pointing directly up the field, the other of atoms turned through 180° relatively to the first set; there was quantitative agreement with the theory. If the outside electron moves in a two-quantum orbit, the magnetic moment of the atom is $2\ eh/4\pi m$, and the cosine of the angle may take the values ± 1 and the values $\pm \frac{1}{2}$; if in a n-quantum orbit, the moment is $neh/4\pi m$ and the permissible values for the cosine are $\pm 1/n,\ \pm 2/n, \ldots\ldots \pm n/n$.[6]

The theory also accounts for the normal Zeeman effect. It remains to be settled whether the magnetic moments of actual paramagnetic substances can be calculated from it. According to the accepted

[6] The condition governing the angle is, that the integrals of (a) the angular momentum of the electron in its orbit, and (b) the projection of the angular momentum on the plane normal to the field, taken around a complete cycle of the orbital motion, must both separately be integer-multiples of the quantum-constant h. The latter integer-multiple cannot be zero, according to Gerlach's experiment and Sommerfeld's theory.

belief, the atoms of a paramagnetic substance all have a given constant magnetic moment, but are oriented in every possible direction so that the resultant magnetic moment of any piece of the substance is zero. If all the atoms could be made to point in the same direction by a powerful magnetic field, the total moment of the piece would be equal to the number of atoms in it multiplied by the moment of each atom, which could then be determined. No attainable magnetic field is strong enough to do this; the persistent effort of the field to twist the atoms into parallelism is almost completely counterbalanced by the thermal agitation. The total moment of the piece when all the atoms are parallel, and therefore the moment of each atom, have therefore to be calculated from the trend of the magnetization-versus-field strength curve in its attainable portion. In making this calculation it has heretofore been assumed that all orientations of the atoms are possible. Replacing this assumption by the contrasting one explained in the foregoing, we find the method of calculation altered;[7] the data heretofore assembled remain valid, but the values of magnetic moment computed from them are replaced by an entirely new set.

The old set of values of magnetic moment, calculated for a number of solid and gaseous substances and of ionized liquids, by Weiss and others, were said to be integer multiples of a fundamental constant, the "Weiss magneton." No one had succeeded in calculating the observed value of this constant from any atomic theory, and it is not compatible with the picture of the atom given above. The new set of values, according to Gerlach and to Pauli, who have worked over the published experimental material, is compatible with the atom-model. The values for solid platinum and palladium; for nickel in its high-temperature non-ferromagnetic "beta" form; and for nitric oxide gas, agree with the simplest model—the electron in a one-quantum orbit revolving around a non-magnetic centre. The value for gaseous oxygen agrees with the model having an electron in a two-quantum orbit; gamma-iron with the three-quantum, Mn_2O with the 4-quantum and MnO with the 5-quantum model. Various ions in solution from Cabrera's data also give values in accordance with the theory. It is implied that these cover all the reliable ob-

[7] In the latter case it is assumed that the number of atoms oriented with their axes in one permissible direction D_1 stands to the number oriented in another permissible direction D_2 in a ratio given by $exp\ (W/kT)$ where T is the temperature, k is Boltzmann's constant, and W is the work required to twist an atom from direction D_1 to direction D_2 against the magnetic field. In the former case all directions are regarded as permissible, and in the assumption just stated, "number of atoms oriented in direction D" is replaced by "density in solid angle of atoms oriented in direction D," a fundamental change.

servations on paramagnetic substances, except for two ions which yield values not reducible to agreement with the new theory.[8] The new method of calculating magnetic moments thus leads to values which confirm the contemporary atom-model. It would not be desirable to dismiss the old method and the old theory too hastily, considering that they lead to values which are claimed to be integer multiples of an apparently fundamental constant; but this constant has proved so intractable to theory that it would be gratifying to be able to discard it.

The arrangement of atoms in two samples of Heusler alloys was investigated with the X-ray method by J. F. T. Young at Toronto. These alloys are mixtures of the metals, copper, manganese, and aluminium in certain proportions; they are strongly ferromagnetic while the component metals are not ferromagnetic at all. Of the two samples, one had a much higher permeability than the other; the atoms of the former sample were arranged in a body-centered cubic lattice, with no trace of the characteristic lattices of the component metals. The atoms of the latter sample were arranged in a face-centred-cubic lattice. Thus these alloys furnish an additional instance of the frequent, though not by any means universal, correlation between body-centred-cubic lattice and strong ferromagnetism. L. W. McKeehan of the Western Electric used the same method to investigate palladium containing great quantities of occluded hydrogen. The space-lattice of the hydrogen-free metal was distended by a certain fixed percentage by saturating it with hydrogen; and it appeared that when the palladium contained a lesser quantity of hydrogen than the maximum or saturation amount, some parts of it were quite saturated and others contained no hydrogen at all, instead of the whole lattice being equally enlarged; it is probable that the individual crystals of the metal are saturated one by one as the hydrogen creeps in.

[8] The value for beta-iron as quoted by Gerlach does not agree with the theory. As for the values assigned by Weiss to the three ferromagnetic metals iron, nickel and cobalt, obtained from direct measurements of the saturation-intensity at the temperature of boiling hydrogen, the first two do not agree with the theory, the last agrees very well (assuming the electron to be in a one-quantum orbit). Of course, it is likely enough that the theory should not be applied to ferromagnetics. It seems fitting to quote a remark of Andrade about theories of magnetism in general " . . . the substances selected for verification of theories are of a very limited class, called of normal behavior rather because they agree with the theories than because they represent a numerical majority."

REFERENCES

D. M. Bose and S. K. Ghosh: *Nature* 111, pp. 463–464; 1923.

A. H. Compton: *Physical Review*, 2d ser., 21, pp. 483–502 and 715; 1923. (X-ray scattering.)

A. H. Compton: *Philosophical Magazine*, 43, pp. 1121–1129; 1923. (Total reflection of X-rays.)
K. T. Compton: *Physical Review*, 2d ser., 21, pp. 266–291; 1923.
W. Gerlach: *ZS. für Physik*, 9, pp. 349–355; 1922. (Magnetic moment of silver atoms.)
W. Gerlach: *Physikalische ZS.*, 24, pp. 275–277; 1923.
H. Gunther-Schulze: *ZS. für Physik*, 15, pp. 8–23; 1923. (Cathode-fall.)
H. Gunther-Schulze: *ZS. für Physik*, 13, pp. 378–391; 1923. (Anode-fall.)
W. D. Harkins and W. H. Ryan: *Journal Am. Chem. Soc.*, 45, pp. 2095–2107; 1923.
F. Hund: *ZS. für Physik*, 14, pp. 241–263; 1923.
L. W. McKeehan: *Physical Review*, 2d ser., 21, pp. 334–342; 1923.
R. Minkowski and H. Sponer: *ZS. für Physik*, 15, pp. 399–409; 1923.
F. Paschen: *Annalen der Physik*, 71, pp. 142–161; 1923.
W. Pauli: *Physikalische ZS.*, 21, pp. 615–617; 1920.
C. Ramsauer: *Annalen der Physik*, 66, pp. 546–558; 1921.
H. A. Wilson: *Proc. Royal Society*, A103, pp. 53–57; 1923.
J. F. T. Young: *Philosophical Magazine*, 46, pp. 291–305; 1923.

Transatlantic Radio Telephony[1]

By H. D. ARNOLD and LLOYD ESPENSCHIED

SYNOPSIS: The first transmission of the human voice across the Atlantic was accomplished by means of radio in 1915. Since that time substantial progress has been made in the art of radio telephony and in January of this year another important step was taken in the accomplishment of transoceanic voice communication. At a prearranged time telephonic messages were received in London from New York clearly and with uniform intensity over a period of about two hours.

These talking tests were part of a series of experiments on transatlantic telephony which are now under way, the results of which to date are reported in this paper.

A new method of transmission, radiating only a single side-band, is being employed for the first time. As compared with the ordinary method of transmission, this system possesses the following important advantages:

The effectiveness of transmission is greatly increased because all of the energy radiated is effective in conveying the message; whereas in the ordinary method, most of the energy is not thus effective.

The stability of transmission is improved.

The frequency band required for transmission is reduced, thus conserving wave length space in the ether and also simplifying the transmitting antenna problem.

An important element of the high-power transmitter is the water-cooled tubes, by means of which the power of the transmitted currents is amplified to the order of 100 kilowatts or more. The direct-current power for these tubes is supplied from a 60-cycle, a-c. source through water-cooled rectifier tubes.

A highly selective and stable type of receiving circuit is employed. Methods and apparatus have been developed for measuring the strength of the electromagnetic field which is delivered to the receiving point and for measuring the interference produced by static.

The transmission tests so far have been conducted on a wave length of 5260 meters (57,000 cycles per second). The results of the measurements during the first quarter of the year on the transmission from the United States to England show large diurnal variations in the strength of the received signal and in the radio noise strength, as is to be expected, and correspondingly large diurnal variations in the ratio of the signal to noise strength and in the resulting reception of spoken words. Also, the measurements, although as yet incomplete, show a large seasonal variation.

The character of the diurnal and seasonal variations is clearly indicated in the figures. The curves present the most accurate and complete data of this kind yet obtained.

ON January 15, of this year, a group of about 60 people gathered in London at a prearranged time and listened to messages spoken by officials of the American Telephone and Telegraph Company from their offices at 195 Broadway, New York City. The transmission was conducted through a period of about two hours, and during this time the words were received in London with as much clearness and uniformity as they would be received over an ordinary wire telephone circuit. During a part of the time a loud speaker

[1]This paper, with the exception of the Appendix, was presented at the Annual Convention of the A. I. E. E., Swampscott, Mass., June 26–29, 1923, and was printed in the Journal for August, 1923.

was used in connection with the receiving set, instead of head receivers. The reporters present easily made a transcription of all the remarks, both with head sets and with the loud speaker.

These tests were made possible by cooperation between the engineers of the American Telephone and Telegraph Company and the Western Electric Company, and the engineers of the Radio Corporation of America and its associated companies. The sending apparatus was installed in the station of the Radio Corporation of America, at Rocky Point, L. I., in order to make use of that company's very efficient multiple-tuned antenna. The receiving apparatus was installed in the buildings of the Western Electric Company, Ltd., at New Southgate, England.

This was not the first time speech had been transmitted from America to Europe. Transatlantic telephony was first accomplished in 1915, when the American Telephone and Telegraph Company transmitted from the Navy station at Arlington, Va., to the Eiffel Tower, Paris. In these earlier tests, however, speech was received in Paris only at occasional moments when transmission conditions were exceptionally favorable. The success of the present tests indicates the large amount of development which has been carried out since this first date.

The recent talking tests were carried out as part of an investigation of transatlantic radio telephony. This investigation is directed at determining (1) the effectiveness of new methods and apparatus which have been developed for telephonically modulating and transmitting the large amounts of power necessary for transoceanic operation, (2) the efficacy of improved methods for the reception of this transmission and for so selecting it as to give an extremely sharp differentiation between the range of frequencies transmitted and all the frequencies outside of this range; and (3) determining the transmission characteristics for transatlantic distances and the variation of the characteristics with the time of day and the season of the year, including the measurement of the amount of static interference.

The tests are being continued, particularly as regards the study of transmission efficiency.

Single Side-Band Eliminated Carrier Method of Transmission

The method of transmission used in these experiments is what we know as the single side-band eliminated carrier method[2]. With this

[2]For a more complete exposition of this method see U. S. patent No. 1449382 issued to John R. Carson to whom belongs the credit for having first suggested it. Also see Carson patents Nos. 1,343,306 and 1,343,307.

method, the narrowest possible band of wave lengths in the ether is used, and all of the energy radiated has maximum effectiveness in transmitting the message.

As had been pointed out in other papers[3], when a carrier is modulated by telephone waves, the power given out is distributed over a frequency range, and may be conveniently considered in three parts: (1) energy at the carrier frequency itself, (2) energy distributed in a frequency band extending from the carrier upward, and having a width equal to the frequencies appearing in the telephone waves, and (3) energy in a band extending from the carrier downward, and having a similar width. The power at the carrier frequency itself makes up somewhat more than two-thirds of the total power, even when modulation is as complete as possible. Furthermore, this energy can, in itself, convey no message, as is self evident. In the present method, therefore, the carrier-frequency component is eliminated, by methods explained in detail below with the result that a large saving in power is effected. Each of the remaining frequency ranges, generally known as the upper and the lower side-band respectively, transmits power representing the complete message. It is therefore unnecessary to transmit both of these side-bands, so that in the present method one of them is eliminated. In this way the transmission of the message uses only half the frequency band required in the usual method of operation. Similarly the frequency-band accepted by the receiving set is narrowed to conform to a single side-band as compared with the usual double side-band reception, and as a result the ratio of signal to interference is improved. Certain other advantages of this method will be brought out in the further discussion.

While these advantages of the single side-band eliminated carrier method hold good for radio telephone transmission generally, they become of the utmost importance in transoceanic work, because of the necessity of conserving power in a system where the transmitting powers are large, and also because the very limited frequency range available for long distance transmission makes it imperative that each part of the range shall be utilized with the greatest of care. Before discussing the method further, the circuits and apparatus which are actually used in the tests will be described.

[3]"Carrier Current Telephony and Telegraphy" by Colpitts and Blackwell. Journal A. I. E. E., April, 1921.

"Application to Radio of Wire Transmission Engineering" by Lloyd Espenschied. Proc. Inst. Radio Engrs., Oct. 1922.

"Relations of Carrier and Side-bands in Radio Transmission" by R. V. L. Hartley. Proc. Inst. Radio Engrs., Feb. 1923.

The Transmitting System

The transmitting system is shown in simplified circuit form in Fig. 1. It is illustrated as grouped into three parts: The low-power modulating and amplifying stages, shown below in light lines; the high-power amplifiers, shown in heavy lines above and to the right; and the rectifier which supplies the power amplifier with high-tension direct current, shown in the upper left-hand portion of the diagram.

Referring first to the low-power portion of the system, it will be seen that the voice currents (from either a telephone line or a local microphone) are fed into a balanced type of modulator circuit and are modulated with a carrier current of a frequency of about 33,000 cycles. The operation of the balanced type of modulator in suppressing the unmodulated carrier component is explained in the Colpitts and Blackwell carrier current paper referred to above. The result of this modulating action is to produce in the output circuit of modulator No. 1, modulated current representing the two side-bands, for example, the upper one extending from 33,300 to 36,000 cycles and the lower one from 32,700 down to 30,000 cycles. These components are impressed upon a band filter circuit which selects the lower side-band to the exclusion of the upper one and of any remaining part of the carrier, with the result that only one side-band is impressed upon the input of the second modulator. This second modulator is provided with an oscillator which supplies a carrier current of 88,500 cycles. The result of modulation between the single side-band and this carrier current is to produce a pair of side-bands which are widely separated in frequency, the upper one, representing the sum of the two frequencies, extending from 118,500 to 121,200 cycles and the lower one, representing the difference between the two frequencies, extending from 58,500 down to 55,800 cycles. In this second stage of modulation there is a relatively wide separation between the two-side bands which facilitates the selection at these higher frequencies of one side-band to the exclusion of the other. Another important advantage is that it allows a range of adjustment of the transmitted frequency without changing filters. This is accomplished by varying the frequency of the oscillator in the second step. In the present case, the frequency desired for transmission is that corresponding to the lower side-band of the second modulator. The lower side-band of from 58,500 to 55,800 is therefore selected by means of the filter indicated. This filter excludes not only the other side-band but also any small residual of 90,000-cycle un-

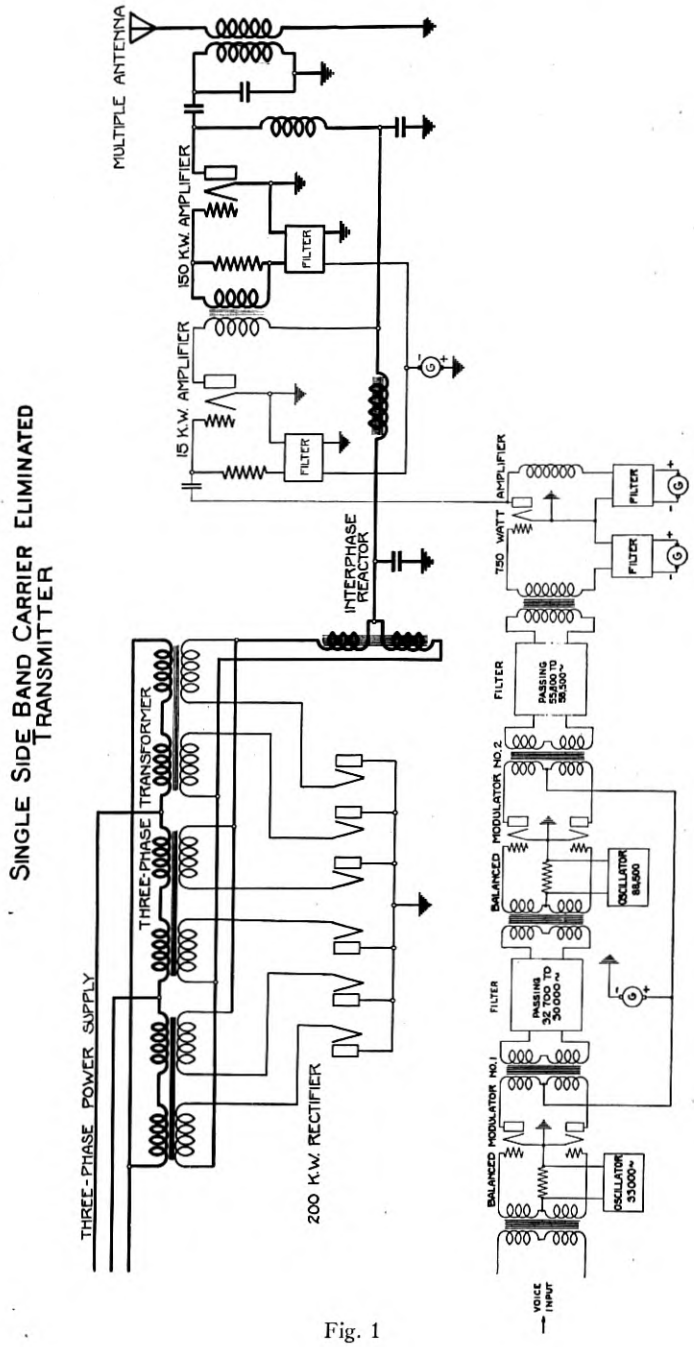

SINGLE SIDE BAND CARRIER ELIMINATED TRANSMITTER
MULTIPLE ANTENNA
150 K.W. AMPLIFIER
15 K.W. AMPLIFIER
FILTER
INTERPHASE REACTOR
750 WATT AMPLIFIER
PASSING 55,500 TO 58,500~
BALANCED MODULATOR NO.2
OSCILLATOR 88,500
PASSING 32,000 TO 35,000~
BALANCED MODULATOR NO.1
OSCILLATOR 35,000~
VOICE INPUT
200 K.W. RECTIFIER
THREE-PHASE TRANSFORMER
THREE-PHASE POWER SUPPLY

Fig. 1

modulated carrier current which may get through the second modulator circuit if it is imperfectly balanced.

Having prepared at low power the side-band currents of desired frequency it is necessary to amplify them to the required magnitude for application to the transmitting antenna. This amplification is carried out in three stages. The first stage increases the power to about 750 watts, and is shown in Fig. 1 together with the modulating circuits. This amplifier employs in its last stage three glass vacuum tubes rated at 250 watts each and operating at 1500 volts.

The output of the 750 watt amplifier is applied to the input of the larger-power amplifying system beginning with the 15-kw. ampli-

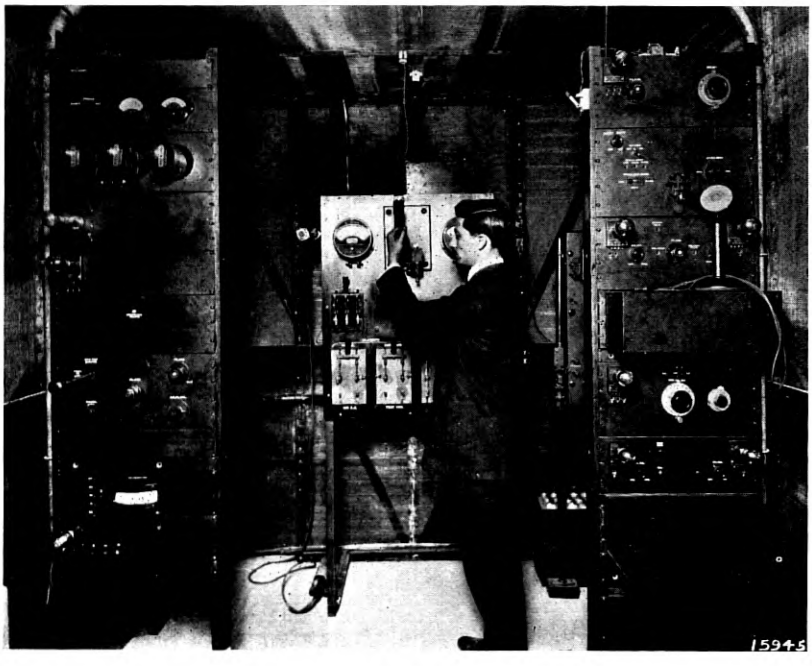


Fig. 2

fier of Fig. 1. This consists of two water-cooled tubes in parallel, operating at approximately 10,000 volts. The output of this amplifier is applied by means of a transformer to the input of the 150-kw. amplifier which consists of two units of ten water-cooled tubes each, all operating in parallel at about 10,000 volts.

The high-voltage, d-c. supply is furnished by a large vacuum tube rectifier unit rated at 200 kw. It employs water-cooled tubes similar

to those used in the power amplifiers except that they are of the two-electrode type. The rectifier operates from a 60-cycle, three-phase supply circuit and utilizes both halves of each wave. The two sets of rectified waves are combined by means of an inter-phase reactor which serves to smooth out the resultant current and by distributing the load between tubes of adjacent phases increases the effective load capacity of the rectifier. The ripple is further reduced by the filtering retardation coil and condensers shown.

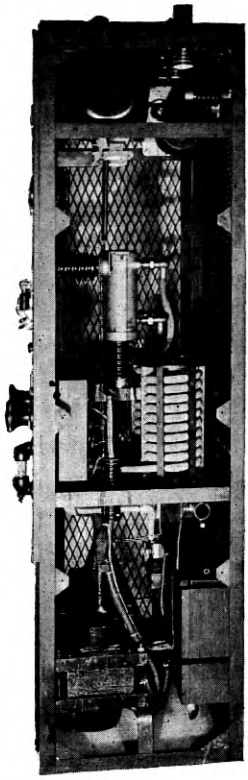


Fig. 3

Reproductions of the apparatus comprising the transmitter system as described above are given in Figs. 2, 3, 4 and 5.

Fig. 2 shows the apparatus comprising the low-power stage of the transmitting system. The right-hand rack contains the two weak-power modulating units and the two single-side-band selecting filters. The left-hand rack is the 750-watt amplifier unit. The three radiation-cooled tubes of 250-watt capacity each will be seen near the top.

Below are the smaller amplifying stages. The power supply board is shown in the center of the photograph.

Fig. 3 is a side view of the 15-kw. amplifier unit. The face of the panel from which the control handles protrude is on the left. Mounted in the cage behind the front panel are two water-jackets for accommodating the water-cooled tubes, also a coiled hose for increasing the electrical resistance of the water supply circuit (the water-cooled anodes of the tubes being operated above ground potential).

The final amplifier of 150-kw. capacity is shown in Fig. 4. It comprises two units each of 75 kw. Each unit contains 10 water cooled tubes which can be seen mounted in their water jackets. To the right of these units is located the 200-kw. rectifier unit shown in Fig. 5. The unit contains actually 12 tubes, there being two tubes

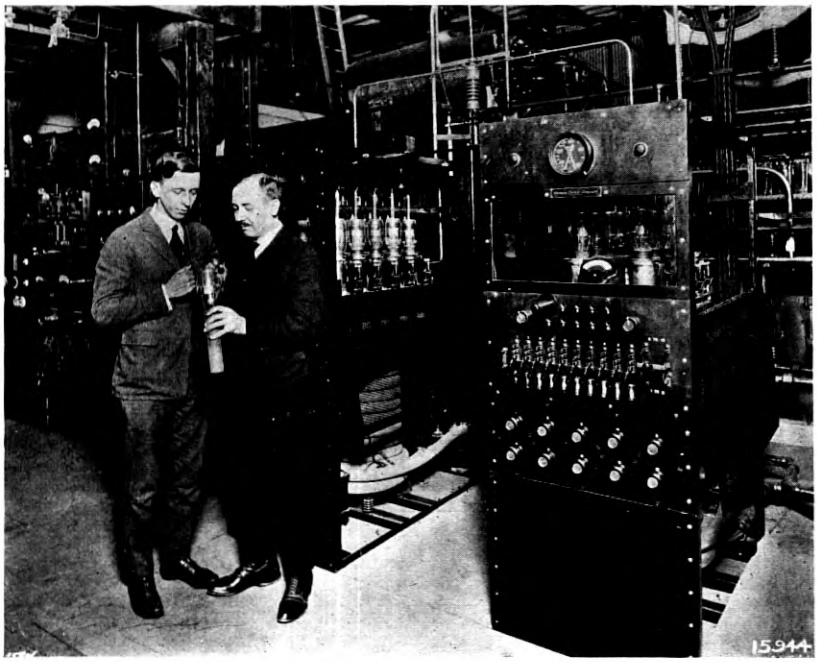


Fig. 4

for each of the six half waves. The pancake coils on the top of the rack are protecting choke coils to guard the transformer secondary winding against steep wave fronts in case of tube failure.

From the above description it will be understood that the transmitting system is one in which the useful side-band is first developed

by modulation and filtration at low power and then its power is built up to a large value in a succession of powerful amplifiers. It

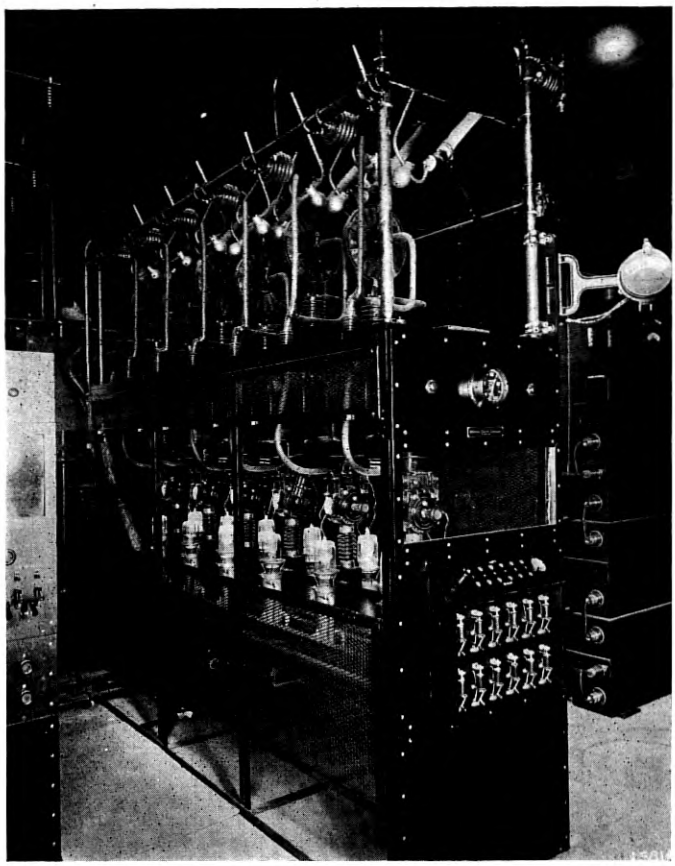


Fig. 5

will be appreciated, therefore, that the large-power amplifiers and in particular the water-cooled tubes which are their essential elements represent one of the major problems of the development.

High-Power Tubes

The development of the high-power tubes is described quite fully in another paper[4]. The present discussion is, therefore, limited to a few of the outstanding features.

[4] *Bell System Technical Journal*, July 1922.

In the design of high-power tubes for use in this system the main problem is to insure the ready disposal of the large amounts of heat generated at the anodes. For the conditions of use in the present type of system where the tube is employed as an amplifier, the power which must be disposed of as heat at the anode is of the same order of magnitude as the power which the tube will deliver to the antenna. In the case of the present equipment, therefore, the tube must be so designed as to operate continuously with a heat dissipation at the anode of more than 10 kw. It is obviously difficult to secure so large a dissipation in a tube enclosed with glass walls, and a tube was therefore designed in which the anode forms a part of the wall of the containing vessel and the heat generated in it is removed by circulating water. The tube used is shown in Fig. 6. The lower cylindrical portion is the anode which is drawn from a sheet of copper. The

Fig. 6

upper portion is of glass and serves both to support and insulate the grid and filament elements.

The three principal difficulties met in the construction of these tubes are the making of a vacuum-tight seal between the copper and the glass, the provision of adequate means for conducting through the glass wall the large currents necessary to heat the filament, and the obtaining of the necessary vacuum for high-power operation.

The first of these problems was solved by the development of a new metal to glass seal. In making this seal the glass and metal parts are brought into contact while hot, the temperature being high enough for the glass to wet the metal. The part of the metal in contact with the glass is made so thin that the stresses which are set up when the seal cools are not great enough to fracture the glass or to break it away from the metal at the surface of contact. Seals made in this way are sufficiently rugged to stand repeated heating and cooling from the temperature of liquid air to that of molten glass without deterioration.

A seal employing the same principle but different in form is also used at the point where the leads carrying the filament current pass through the glass walls of the tube. The lead is made of copper 0.064 in. in diameter and passes through the center of a copper disk, 0.010 in. thick, the joint between the lead and the disk being made vacuum-tight by the use of a high melting solder. The disk is sealed to the end of a glass tube which is in turn sealed into the glass wall of the vacuum tube.

In exhausting the tubes it has been found necessary to subject all the metal parts to a preliminary heat treatment in a vacuum furnace during which the great bulk of the occluded gasses is removed. By this method the time of exhaust can be considerably reduced but the vacuum conditions to be met are so stringent that the final processes of evacuation must be carefully controlled and often occupy as much as twelve hours.

The tubes are operated at a plate voltage of 10,000 volts and are capable of delivering 10 kw. at this voltage in a suitable oscillatory circuit. For this performance an average electron current of 1.35 amperes is required. The total electron current that the filament must be capable of supplying to insure steady operation is about 6 amperes.

When the tubes are used to amplify modulated currents with large peak values such as are characteristic of telephone signals it is essential that the maximum electron current through the tube shall be several times the normal operating current and therefore to insure the necessary high quality of transmission these tubes are operated for telephone purposes with an average output of about 5 kw.

The Receiving System

In the method of transmission ordinarily employed in radio telephony by which the carrier and both side-bands are sent out from the transmitting station and received at the distant end, detection is readily

accomplished merely by permitting all of these components to pass through the detector tube. The detecting action whereby the voice-frequency currents are derived, is accomplished by a remodulation of the carrier with each side-band.

With the present eliminated carrier method of transmission the side-band is unaccompanied by any carrier with which to remodulate in the receiving detector. It is necessary, therefore, to supply the

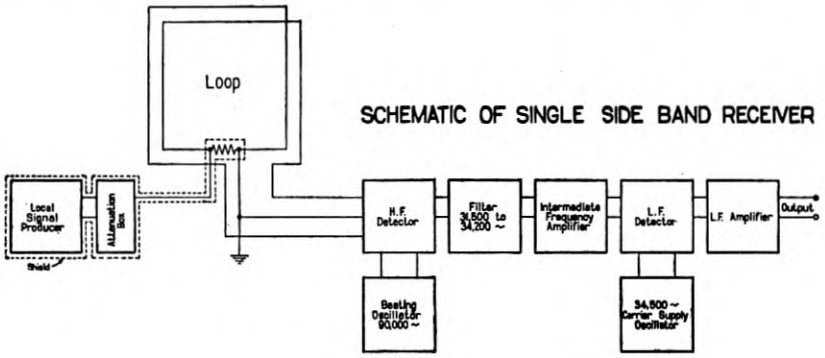


Fig. 7

detector with current of the carrier frequency obtained from a local source. Thus, in the present experiments, if a current of the original carrier frequency, 55,500 cycles, is supplied to the detector it will remodulate or "beat" with the received side-band of, say 55,800 to 58,500 cycles and a difference-frequency band of 300 to 3000 cycles, *i.e.*, the voice frequency band will result.

The arrangement actually used, however, is not quite so simple as this. It is shown schematically in Fig. 7. Reception is carried out in two steps, the received side-band being stepped down to a lower frequency before it is detected. The stepping down action is accomplished by combining in the first detector the incoming band of 55,800 to 58,500 cycles with a locally generated current of about 90,000 cycles. In the output circuit of the detector the difference-frequency band of 34,200 to 31,500 cycles is selected by a band filter and passed through amplifiers and thence to the second detector. This detector is supplied with a carrier of 34,500 cycles which, upon "beating" with the selected band, gives in the output of the detector the original voice-frequency band.

The object of thus stepping down the received frequency is to secure the combination of a high degree of selectivity with flexibility in tuning. The high selectivity is obtained by the use of a band filter.

It is further improved by applying the filter after the frequency is stepped down rather than before. To illustrate this improvement assume that there is present an interfering signal at 60,000 cycles, 1,500 cycles off from the edge of the received telephone band. This is a frequency difference of about 2½ per cent; but after each of these frequencies is subtracted from 90,000 cycles, the difference of 1500 cycles becomes almost 5 per cent. This enables the filter to effect a sharper discrimination against the interfering signal. Furthermore, the filter is not required to be of variable frequency as would be the

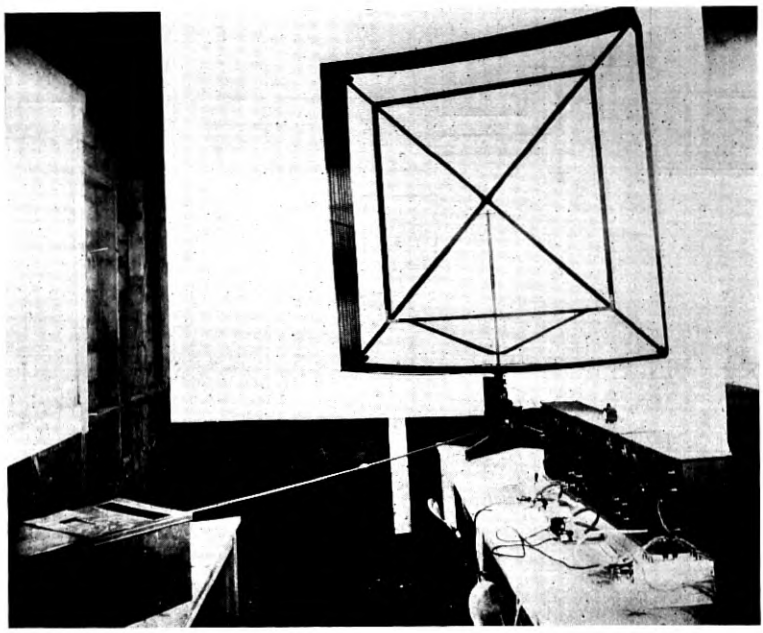


Fig. 8

case were it employed directly at the received frequency since by adjusting the frequency of the beating down oscillator the filter is in effect readily applied anywhere in a wide range of received frequencies. The receiving method, therefore, enables the filter circuit, and indeed also the intermediate frequency amplifiers, to be designed for maximum efficiency at fixed frequency values without sacrificing the frequency flexibility of the receiving set.

A photograph of the receiving set used in the transatlantic measurements is reproduced in Fig. 8. The signals are received on a square

loop six feet on a side and wound with 46 turns of stranded wire. The first box contains the beating oscillator and high-frequency detector, the second box of the filter and amplifying apparatus for the intermediate frequency and the third box the final detector and amplifier. The shielded box at the left of the picture, which is connected to the loop by means of leads in the copper tube, is the apparatus for introducing the comparison signal of known strength into the loop for measuring purposes. This receiving and measuring set is described more in detail in a paper by Bown, Englund, and Friis in the "Proceedings of the Institute of Radio Engineers for April, 1923."

Although it was this very selective and reliable method of intermediate-frequency reception which was used in London, it is quite possible to receive the single-side-band transmission by means of a regular heterodyne receiving set. Even a self-regenerative set will suffice under some conditions. It is necessary, however, to adjust the frequency of the oscillator very carefully to that of the transmitting carrier frequency, otherwise serious distortion of the received speech will result. Also it is, of course, necessary that the tuning be not too sharp if ordinary tuned circuits and not filter circuits are employed. One might expect that some difficulty would be experienced in maintaining the frequency at the receiving end in sufficiently close agreement with the sending frequency. In the tests no particular difficulty was experienced, the oscillators at the two ends being so stable that only an occasional slight readjustment of the receiving oscillator was required. With the development of more stable oscillators, doubtless, the frequency with which readjustments are required, will be further reduced. If serious distortion of the received speech is to be avoided the two frequencies must be within about 50 cycles, an accuracy of 0.1 per cent at 50,000 cycles.

Transmission Advantages of the System

Since the present experiment represents the first use of the single-side-band eliminated carrier type of system some further discussion of the characteristics and advantages of the system is appropriate.

The importance of the system in conserving frequency range will be appreciated when it is realized that the total frequency range available for transatlantic telephony is distinctly limited. Just what the most suitable range is has not been accurately determined but it seems limited to below 60,000 cycles (5000 meters) because of the large attenuation suffered during the daylight hours by frequencies higher than this. On the lower end of the frequency scale, trans-

atlantic telegraphy at present pretty well preempts frequencies below 30,000 cycles (10,000 meters). Therefore, for the present at least transatlantic telephony is limited to a range of some 30,000 cycles. Now transmission of speech requires as a minimum for good quality a single-side-band 3000 cycles wide. Allowing for variations and clearances between channels it is doubtful if the channels could be made to average closer than one every 4000 cycles for single-side-band transmission and one every 7000 cycles for the ordinary double-side-band transmission. This means that even were the whole range from 30,000 to 60,000 cycles devoted to telephony to the exclusion of telegraphy, only about four channels could be obtained by the older methods and some seven by the present one.

It is a rather interesting commentary to note that a somewhat similar situation as to limitation in frequency range exists in the case of carrier-current transmission over wires. The transmission efficiency falls off with increase in frequency and limits the range of frequencies which can be economically used, in much the same way as it is limited in long distance radio transmission. It is because of this limitation in the case of wires and the value which attaches to conserving the frequency range consumed per message that single-side-band transmission was first developed for wire carrier current systems. Its development in wire transmission has been of considerable aid in adapting the method to the present purpose of transatlantic operation.

The second of the outstanding characteristics of the present system resides in the large power economy which it permits. Transatlantic telephony requires hundreds of kilowatts of high-frequency power. Since it is difficult and expensive to produce this power it is important that every effort be made to increase its efficiency or effectiveness in transmitting the voice. To illustrate how the present system effects economies in power, consider the case of a carrier wave completely modulated by a single frequency tone. In such a completely modulated wave, only 1/3 of the total power contains the message, the remaining 2/3 conveying only the carrier frequency which can as well be supplied from an oscillator of small power at the receiving station. It is obvious, therefore, that by eliminating the carrier only 1/3 as much power need be used as would be required were all the elements of the completely modulated wave transmitted. To realize the maximum advantage of this mode of operation, the system eliminates the carrier at low power and, thereby, the high-power apparatus is devoted exclusively to the amplification of the essential part of the signal.

If, after having suppressed the carrier, both side-bands were transmitted, their reception would require perfect synchronism between the carrier resupplied at the receiving end and that eliminated at the sending end, a condition which is practically impossible to meet without transmitting some form of synchronizing channel, which is, indeed, much the same as transmitting the carrier itself. If the receiving carrier is not synchronized, the two side-bands will interfere with each other upon being detected. By eliminating one side-band, this interference is prevented and reception may be carried on, using a locally supplied frequency which is only approximately equal to that of the suppressed carrier. The two may differ by as much as 50 cycles before the quality of the received speech is greatly impaired. The importance to the carrier suppression method of eliminating one side-band will, therefore, be appreciated. The present system eliminates one side-band while still in the low-power stage. While it would be possible to do this selecting after they have both been raised to the full transmitting power, this would require the use of a filter of high-power carrying capacity, which would make the filter very costly and also render the system inflexible to change of wave length. The present system overcomes both of these difficulties by filtering our one side-band at low-power levels and by the use of the double modulation method.

Another very important reason for the transmission of a frequency band as narrow as is possible lies in the difficulty of constructing an antenna to transmit more or less uniformly at these long waves a band of frequencies which is an appreciable fraction of the main carrier frequencies. For example, in the ordinary method of transmission an antenna which was intended to transmit a 30,000-cycle carrier and its two speech side-bands would need to be designed to transmit all the frequencies from 27,000 cycles to 33,000 cycles, a band which is equal to 20 per cent of the carrier frequency. This band is considerably wider than that given by the resonance curve of a highly efficient long wave antenna. To accommodate both side-bands would require flattening out the resonance curve either by damping, which means sacrifice in power efficiency, or by special design of the antenna, possibly throwing it into a series of interacting networks and causing it to become a rather elaborate wave filter. The importance, from the antenna standpoint, of narrowing the frequency band required to be transmitted is, therefore, evident.

It is extremely important that the received signal be affected as little as possible by changes in the transmission efficiency of the medium. The voice frequency currents produced at the receiving

end, after detection, are proportional to the product of the carrier wave and the side-band. If the carrier as well as the side-band is transmitted through the medium, then a given variation in the transmission efficiency of the medium will affect both components and will change the received speech in proportion to the square of the variation, as compared to the first power if only the side-band is transmitted and the carrier is supplied locally. Thus it will be seen that the omission of the carrier from the sending end and the resupplying of it from the constant source at the receiving end gives greater stability of transmission.

Without discussing the system in further detail the advantages of it may be summarized as follows:

1. It conserves the frequency (wave length) band required for radio telephony, which is particularly important at long wave lengths.
2. It conserves power, in that all of the power transmitted is useful signal-producing power. This is particularly important also in long distance transmission which requires the use of large powers.
3. The fact that only a single-band of frequencies is transmitted simplifies the antenna problem at long wave lengths, where the resonance band becomes too narrow to transmit both side-bands.
4. As compared with a system which eliminates the carrier but transmits both side-bands the simple side-band system has the important advantage of not requiring an extreme accuracy of frequency in the carrier which is resupplied at the receiver. Were both side-bands transmitted very perfect synchronism would be required for good quality.
5. It improves the transmission stability of the radio circuit since variations in the ether attenuation affect only one (the side-band) of the two components effective in carrying out the detecting action in the receiver.
6. The receiving part of the overall system has two advantages:

a. It need accept only half of the frequency band which would be required in double side-band transmission, thereby accepting only half of the "static" interfering energy.

b. By stepping down the frequency of the received currents and filtering and amplifying at the low-frequency stage a very sharp cutoff is obtained for frequencies outside of the desired band and a very stable and easily maintained amplifying system is obtained.

Study of Transatlantic Transmission

We come now to a consideration of the second major part of the investigation, namely, that having to do with the transmission of the waves across the Atlantic. It will be evident, from what has been said earlier, that the transmission question is essentially one of how best to deliver, through the variable conditions of the ether to the receiving station, speech-carrying waves sufficiently free from interference to be readily interpretable in the receiving telephone. The transmission efficiency of the medium varies with time of day and year, and is different for different wave lengths. The interference conditions are also influenced by these same factors.

Now we can study this transmission medium in much the same way we would a physical telephone circuit, by putting into it, at the sending end, electromagnetic waves of a known amount of power and measuring the power delivered at the receiving end. The interference at the receiving station likewise may be measured and the ratio of the strength of the signal waves to the interfering waves may be taken as a measure of freedom from interference; this in turn being directly related to the readiness with which the messages are understood. Accordingly, there has been included as an integral part of the investigation of transatlantic radio telephony, the development of suitable methods and apparatus for measuring the strength of the signal waves and of the interfering waves, as they arrive at the receiving station. The apparatus[5] employed in measuring the field strength of the received signals has been outlined above under Receiving System and need not be gone into further. However, a word of explanation about the method which is employed in making the measurement may be helpful. It will be recalled that the specially designed receiving set is provided with a local source of high frequency from which can be originated signals of predetermined strength. A measurement of the field strength of a signal received from the distant transmitter is made by listening first to the distant signal and then to the locally produced signal, shifting back and forth between these signals and adjusting the strength of the local signal until the two are substantially of the same strength. Then, knowing the power delivered by the local source, the power received from the distant station is likewise known. The relation between the power in the input of the radio receiving circuit to the field strength required to deliver that power is known through the geometry of the receiving

[5]It is described in detail in the paper entitled, "Radio Transmission Measurements" by Bown, Englund, and Friis, *Proc. Institute of Radio Engrs.*, April, 1923.

antenna (in this case a loop) and, therefore, the measured power of the signal can be translated directly into the field strength of the received waves.

The measurement tone signal is transmitted from the Rocky Point sending station by substituting for the microphone telephone transmitter a source of weak alternating current of about 1/100 watt at a frequency of approximately 1500 cycles. This tone modulates the radio telephone transmitter in the same way that voice currents would and is radiated from the antenna as a single-frequency wave of 5260 meters (57,000 cycles per second). It, therefore, constitutes a means of sending out a single-frequency continuous wave for measurement purposes. At the receiving end this continuous wave is demodulated to the same tone frequency which it originally had.

For measuring the strength of the received noise, *i.e.*, the radio frequency currents arising from static or other station interference, the method is quite similar. In this case, however, the noise received is so different from that which can be set up artificially in any simple manner that no attempt is made to compare it directly with a local noise standard. Instead the volume of the interfering noise is expressed in terms of its effect in interfering with the audibility of a local tone signal by measuring the local signal which can just be definitely discerned through it. This is a threshold type of measurement which is necessarily difficult to carry out with accuracy. In order to increase the sharpness of definition of the local signal and to make it correspond more closely to speech reception the signal tone is subjected to a continuous frequency fluctuation. The comparison signal has therefore a warbling tone which occupies a frequency band not unlike that of the voice. This method of measuring the interference is discussed in more detail also in the measurement paper referred to above.

Procedure in Making Transmission Measurements. The three quantities which are included in the transmission measurements, namely, the signal strength, the noise strength, and the percentage of words received correctly, are observed one after another in what might termed a unit test period. Although the duration of this test period and the order of making the measurements has been changed somewhat during the course of the experiments, the following program is representative of the conditions under which the data presented below were taken.

A 25-minute test period divided as follows:

5 minutes of tone telegraph identification signals (for receiving adjustment purposes).

10 minutes of disconnected spoken words.

10 minutes of a succession of five-second tone dashes separated by five-second intervals, (for measurement of the received field strength, the intervals between the dashes being used for throwing on the local receiving source and adjusting its strength to equal that of the receive signals by alternately listening to one and then the other).

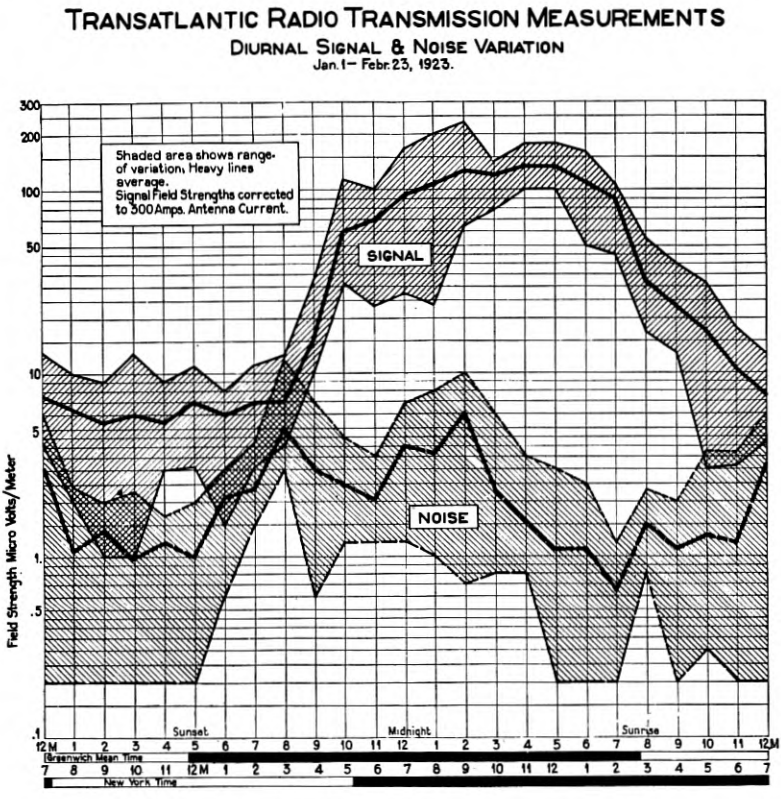


Fig. 9

Immediately following this test period the London observers measured the noise level.

This unit test period was repeated every hour over a period which varied from several hours to as long as two days' duration. Most of the test periods ran for about 28 hours, starting about eleven o'clock Sunday morning and continuing until about three o'clock Monday morning, London time. During this time the telegraph load through

the Rocky Point station of the Radio Corporation was sufficiently light to enable one of the two antennas to be devoted to these experiments. The measurements were started January 1, 1923 and are still in progress.

At the present time (April) the results for the first three months of the tests are available. These results are not yet sufficiently complete nor do they cover a sufficient number of variables in terms

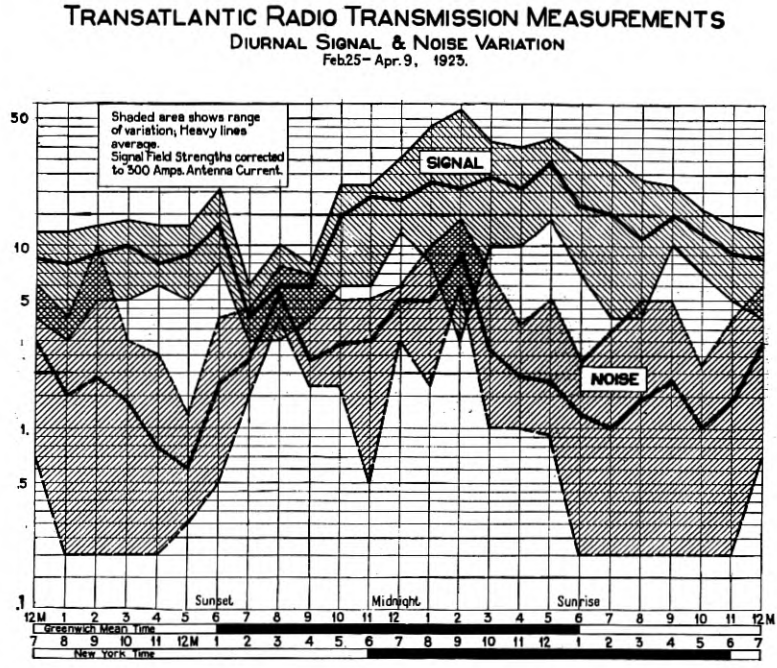


Fig. 10

of time, wave length, etc., to enable any very definite conclusions to be drawn from them. They do illustrate, however, the usefulness of the methods employed, and even in their incomplete state show some factors of considerable interest.

The results of the measurements of received signal, strength and received noise are given in Figs. 9 and 10. The data have been divided and plotted in these two sets of curves because the transmission conditions across the North Atlantic appeared to suffer a rather rapid change about February 23rd. Fig. 9 therefore covers the winter period from January 1 (when the test started) to February 23; and Fig. 10 covers the next period from February 25 to April 9.

The curves are plotted between time of day as abscissas and field strength in microvolts per meter as ordinates. The time during which darkness prevailed at Rocky Point and at London is indicated by the block-fills on the time scales. The overlap of these block-fills indicates the time during which darkness extended over the entire transatlantic path. For Fig. 9 the darkness-belt is as of February 1 and for Fig. 10 as of March 21. The curves show the mean of the results and also the boundaries of the maximum and minimum values observed.

Received Signal Strength. The outstanding factors to be noted concerning the signal strength curves are:

1. The diurnal variations are plainly in evidence. During the first test period covered by Fig. 9, for example, the field strength varied in the ratio of the order of 15 to 1 between day and night conditions, running about 100 microvolts per meter during the night and averaging about 6 microvolts per meter during the day. The diurnal variation is also to be seen in Fig. 10 although the variations between night and day transmission are less marked.

The measured daylight values lend support to the Austin-Cohen absorption coefficient. The average of the observed daylight value for the period of these tests is between 7 and 8 microvolts per meter, while the calculated value is 9.5. Concerning the high field strengths obtaining at night, it should be noted that the maximum observed value of 237 microvolts per meter does not exceed the value of some 340 microvolts which it is estimated should obtain at London were no absorption present in the intervening medium, *i.e.*, were the waves attenuated in accordance with the simple inverse-with-distance law. While no definite conclusions can yet be drawn from these results as to the cause of the diurnal variations, this indication that the upper limit of the variation is the no-absorption condition suggests that the diurnal fluctuations are controlled by the absorption conditions of the medium rather than by reflection or refraction effects.

2. An indication of the seasonal variation which apparently occurs in developing from winter to early spring is found in a comparison of the signal strength curves of Figs. 9 and 10. On the whole the signal strength received in the second test period is considerably less than that received for the first period. This drop in the average of the 24 hours is caused by a large decrease in the night-time transmission efficiency. The daylight transmission does not change much, but what little change there is lies in the direction of an increase as the season advances.

3. A decrease in the transmission efficiency is observed between the time of sundown in London and sundown in New York, that is,

during the period when the sunset condition intervenes in the transmission path. This dip is particularly noticeable in the signal strength curve of Fig. 10. It is not noticeable in Fig. 9, except for the fact that the rise in signal strength corresponding to night conditions in London is delayed until the major part of the transmission path is in darkness.

Strength of Received Noise. The variation in the strength of received noise is shown by the noise curves of Figs. 9 and 10.

1. The diurnal variation of that portion of the noise which is due to atmospheric or "static" disturbances is somewhat obscured by the presence of artificial noise, *i.e.*, noise caused by interference from other stations. The rise in the noise curve at 12 noon is known to be due to artificial interference. In general, however, the large noise values shown to prevail throughout the night in London between about 6 p. m. and 4 a. m. are known to be due to atmospherics. This diurnal variation shows up quite prominently in both figures.

The maximum noise is reached at 2 a. m. London time. Up to this time the night belt extends over London and a sector of the earth considerably to the east and including Europe, Africa and Asia. The noise begins to drop off shortly thereafter and reaches its minimum at sunrise in London. This could be accounted for on the assumption that the major source of the noise lies considerably to the east of London and that transmission of the stray electric waves to London is gradually diminished in efficiency as daylight overtakes the path of transmission.

2. The seasonal variation, as shown by a comparison of the noise curve of Fig. 9 with that of Fig. 10, is not so great as is the case with the transmission efficiency of the signal. However, the noise level is noticeably higher during the second period of the tests, as shown by the average curve of Fig. 10, particularly during the night when the maximum noise obtains.

This indicates that the noise is largely of continental origin lying to the east or south east of London which is in agreement with rough observations made by means of a loop and suggests that the employment of directional antennas would be of considerable advantage. It is expected to include such antennas in the further measurement work.

In connection with these noise curves it should be noted that what they represent is in reality the strength of a local warbling tone-signal, expressed in terms of equivalent field strength in microvolts, which which is just definitely audible through the noise. The actual value of the noise currents, were they measured by an integrating device such as a thermocouple, for example, would be a number of times larger than indicated.

Ratio of Signal to Noise Strength; Words Received. The noise curve of Fig. 9 and that of Fig. 10 can, therefore, be read as "The strength of the signal tone which can just be heard through the noise." It can, therefore, be directly compared with the signal curve itself and the difference between the two curves is a measure of the level of the actual signal strength above that which would just permit of the signals being heard. Actually, the difference between the two curves, as shown in the figures, is proportional to the *ratio* of the signal to the noise strength, because the curves are plotted to a logarithmic scale.

This signal to noise ratio is plotted in Fig. 11 for the test period which corresponds to Fig. 9, and Fig. 12 for the test period which corresponds to Fig. 10. These ratio curves are derived by going back to the original data and taking the ratio for each unit measurement period and spotting it upon the chart as shown by the black points. An average is taken of the points for each hour of the 24-hour period as shown by the circle points. The dash line curves of Figs. 11 and 12, therefore, trace the average diurnal variation of signal to noise ratio.

These curves show:

1. That the signal to noise ratio reaches its minimum during the time when the sunset period intervenes between London and New York.

2. During the night in London the ratio increases more or less continuously and reaches a maximum around the time of sunrise in London.

3. During the course of the daylight period in London the ratio starts out high and drops rather rapidly during the forenoon and assumes a more or less constant intermediate value during the afternoon until sundown. It is during this afternoon period in London that the business hours of the day in London and New York coincide, so that this is the most important period from a telephone communication standpoint.

The drop in the very low ratios obtaining in London in the early evening is due to the fact that an increase in noise occurring at this time is accompanied by a decrease in transmission efficiency from America. This may readily be seen by referring to Fig. 10. The noise increases as the night belt, proceeding westward, envelops England and improves the transmission of atmospherics, which arise possibly in continental Europe, Asia and Africa. As the shadow wall, proceeding westward, intervenes between England and America, the transmission efficiency of the desired signals from America drops and it is not until the night belt extends as far west as America that the transmission efficiency improves sufficiently to overcome the dis-

advantage in London of the large noise values which night there had brought on. Conversely, the high signal to noise ratio, obtaining at about sunrise in London, appears to be due to the fact that as the termination of the night belt, moving westward, intervenes between England and the source of atmospherics to the east, the noise level drops rapidly and has reached low values by the time sunrise arrives in London. At this time, however, darkness still extends to the west

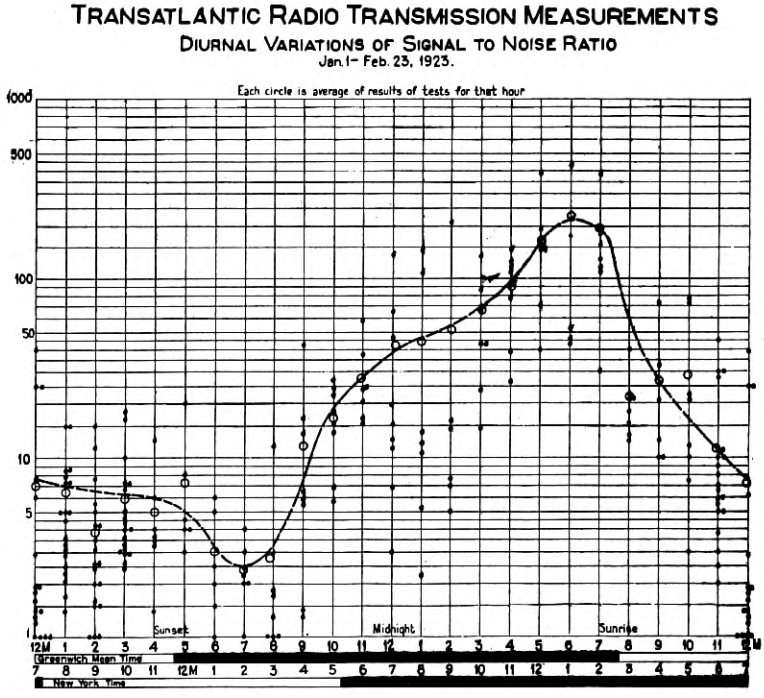


Fig. 11

and the transmission efficiency from America is at its maximum. It is, therefore, due to this interplay between these two factors, signal strength and noise strength, controlled very largely by the transition periods between day and night, that the signal to static ratio varies diurnally in the manner pictured in Figs. 11 and 12.

Concerning seasonal variation, shown by a comparison of Figs. 11 and 12, the following can be said: The dimunition in signal-to-noise ratio in the second test period as compared with the first is caused by the fact that the signal strength has decreased and at the same time the noise has somewhat increased. There is just one other

point that concerns the dip in the ratio occurring at night in London between 12 midnight and 3 a. m. This dip is due to an increase in the noise which occurs around 2 a. m. (A further reduction during this

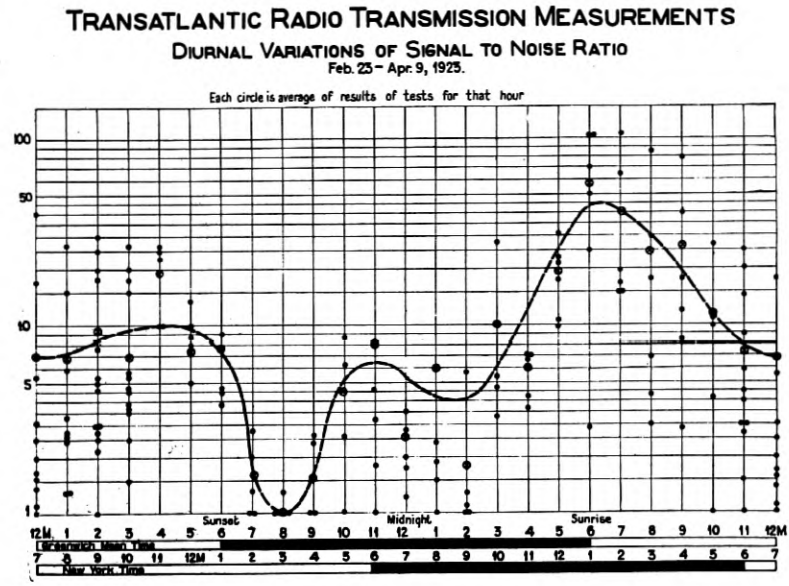


Fig. 12

time, and one which extends the time of minimum ratio from sundown on through the night until 2 a. m. is shown by the April measurements which time has not permitted including in the curves).

During each test period lists of disconnected words were spoken over the systems. As an approximate and easily applied method of indicating the talking efficiency of the circuit, note was made of the percentage of the words which were correctly received.

The curves of Figs. 13 and 14 show the manner in which the percentage of the words which were correctly received varies through the 24 hours. Each point corresponds to the percentage of words correctly received during one unit test period. In many of these tests the interference was noted to be caused by radio telegraph stations, and the data in which the interference is of this character, in so far as identified, are indicated by the triangular dots. It will be seen that most of the poor receptions were due to this cause. Especially is this true of tests at 12 noon at which time severe interference from sources local in London was experienced. The circle points are the

average of results for each hour's tests. The dash line curve is a smoothing out curve of these points.

It is interesting to note that these curves of actual word count conform very well in general shape with those of Figs. 11 and 12 which also really measure receptiveness although in a less direct manner. Reception is best during the late night and early morning,

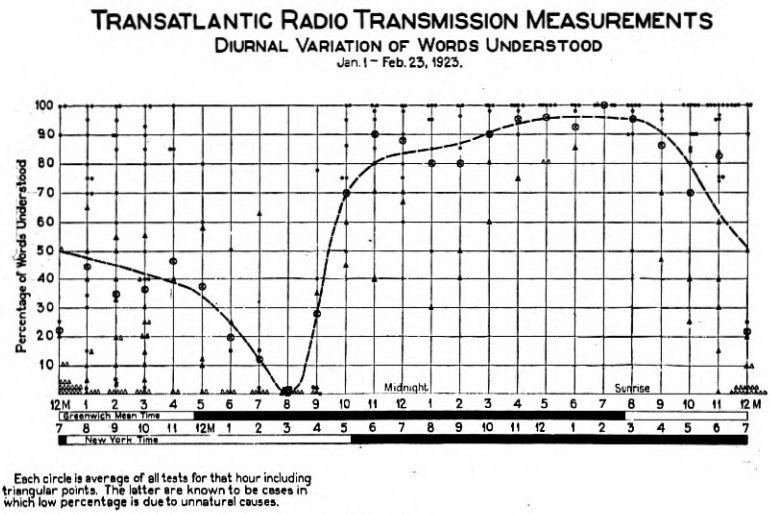


Fig. 13

drops off during the day, reaching a minimum during the evening. Furthermore, the night reception is shown to be considerably better for the January-February period than for the February-March period. The curve of Fig. 14 corresponds quite closely with that of Fig. 12. The curve of Fig. 13 does not show as much of a peak as does that of Fig. 11 which is, of course, due to the fact that above a certain ratio the percentage of words understood is high and cannot rise above 100 per cent.

CONCLUSION

As has been indicated this is a report of work which is still in progress. To date:

A new type of radio telephone system affording important advantages for transatlantic telephony has been developed and put into successful experimental operation across the Atlantic.

Sustained one-way telephonic transmission has been obtained across the Atlantic for the first time by means of this system.

The advantages of this system which had been anticipated, particularly, in respect to economies of power and wave lengths, have been realized. Furthermore, it has been demonstrated that the high-power water-cooled vacuum tubes which have seen their first prolonged operation in this installation are admirably adapted for use in high-power radio installations and particularly for use as high

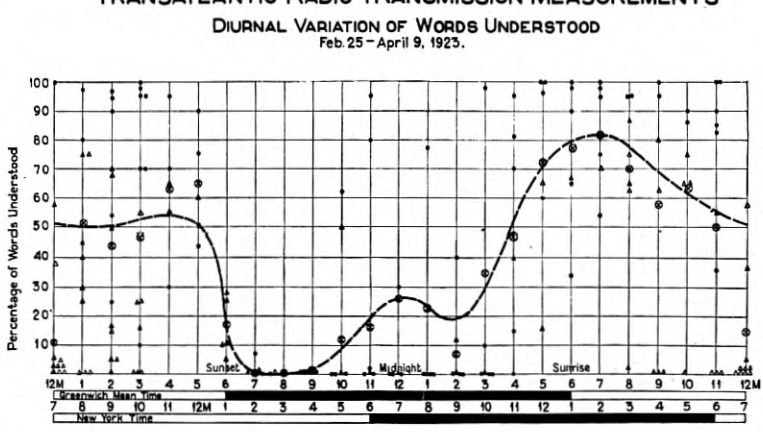


Fig. 14

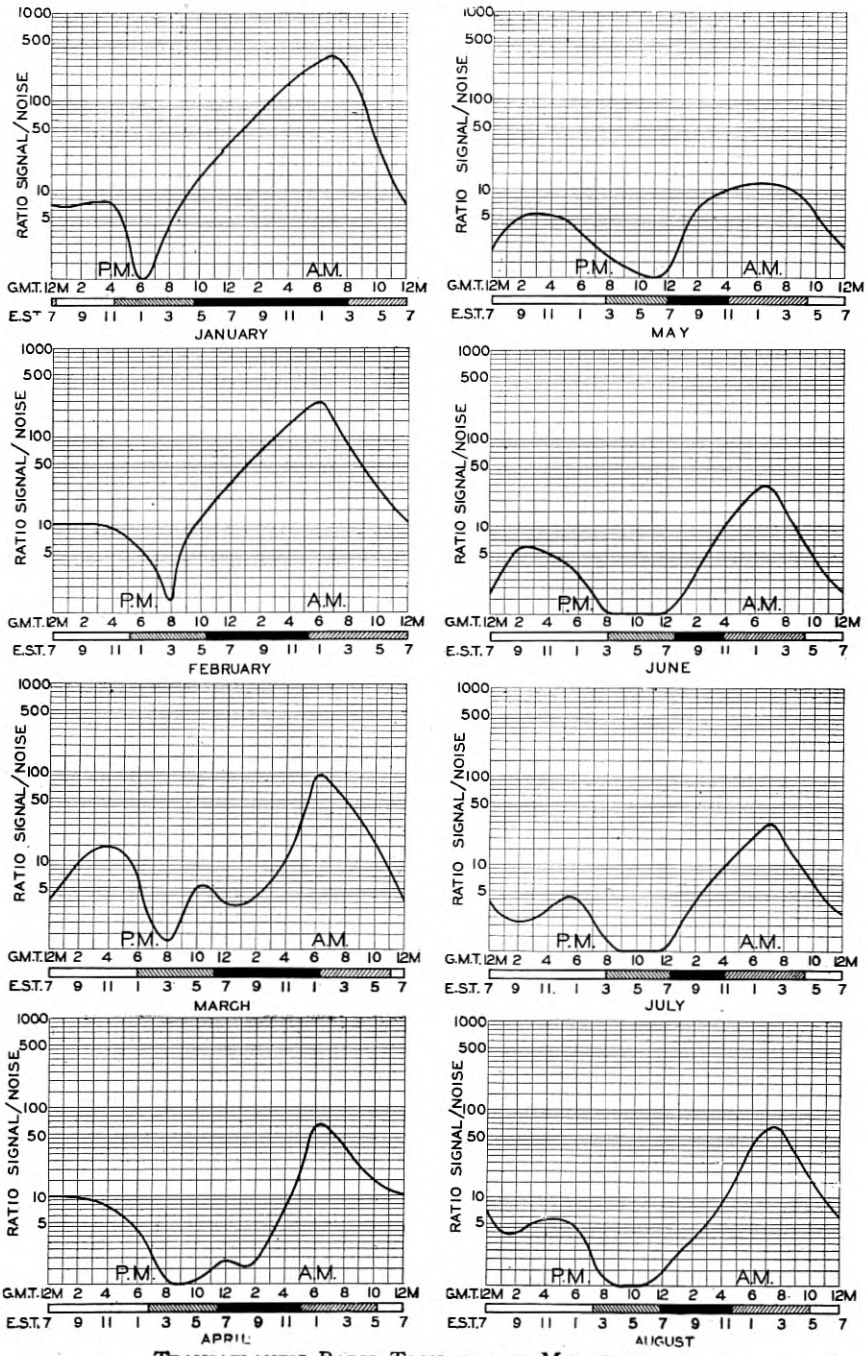
power amplifiers, in the type of system we have described. Also, the method of reception has proved itself to be eminently satisfactory for use with the single side-band type of transmission and to possess important advantages for radio telephony in respect to selectivity and amplification.

Methods have been developed for measuring the strength of the received signals and the strength of the received interfering noise and these methods have been successfully applied in the initiation of a study of the variations to which transatlantic transmission is subject.

The results of the transmission measurements show that, at 5000 meters, the diurnal variations are large, as was to be expected, and give evidences of a large seasonal variation which was, indeed, also to be expected. The results indicate that it will probably be desirable to use a wave length longer than 5000 meters. The measurements are now being made to include the longer wave lengths.

APPENDIX ADDED SEPTEMBER 23, 1923

The results of the transmission measurements from January through August are now available and are summarized in the curves following:



Transatlantic Radio Transmission Measurements
Monthly Averages of Diurnal Variations in Signal to Noise Ratio for 1923. Transmission from Rocky Point to London on 57,000 Cycles (5,260 Meters). Measurements on Loop Reception. Curves Corrected to 300 Amperes Antenna Current

Physical Measurements of Audition and Their Bearing on the Theory of Hearing*

By HARVEY FLETCHER

SYNOPSIS: The author states his purpose to be the presentation of certain facts of audition which have been determined recently with considerable accuracy and the discussion of the theory which best explains these facts.

Making use of data of Knudsen's as well as his own measurements of the auditory sensation area, the author estimates that the normal ear can perceive approximately 300,000 different pure tones. This is taking account of all possible variations in both pitch and intensity. Knudsen's data show that for considerable ranges the minimum perceptible difference in intensity bears a constant ratio to the intensity and the minimum perceptible difference in frequency bears a constant ratio to the frequency. These relations have been termed by psychologists "The Law of Weber and Fechner."

A loudness scale is proposed such that the difference in loudness between two tones is equal to ten times the common logarithm of their intensity ratio. A pitch scale is proposed such that the difference in pitch is equal to one hundred times the logarithm to the base two of the frequency ratio. A method for measuring the loudness of complex sounds is mentioned but is to be discussed in a later paper. A method is proposed for expressing quantitatively different degrees of deafness.

Reference is made to data obtained by the author on the masking of one pure tone by another. The minimum audible intensity of a pure tone depends upon the presence of another tone of different frequency. A low pitched note will, in general, exert a surprisingly large masking effect upon notes of higher frequency. The masking of a low note by a higher is not nearly as pronounced. From his observations, the author draws certain interesting conclusions. For example, given a complex tone consisting of three frequencies 400, 300 and 200 cycles with relative loudness values of 50, 10 and 10, respectively, the ear would hear only the 400 cycle tone and the 200 cycle tone. If the sound is now increased 30 loudness units, without distortion, the 400 cycle tone and the 300 cycle tone only, will be heard.

Binaural masking in which each ear receives one of the two sounds is considered and the conclusion reached that the masking effect noted results from conduction of the masking tone through the bones of the head to the ear receiving the masked tone.

It is stated on the basis of data obtained by Wegel and Lane that the oscillatory system of the ear, comprised by the membranes and little bones of the middle and inner ears, does not obey Hooke's Law regarding the proportionality of stress and strain. Consequently, the ear, when stimulated by a pure tone, introduces harmonics and the workers cited have observed harmonics as high as the 4th order. The non-linear transmission characteristic of the vibratory system of the ear is held to account for the greater masking of a high frequency by a lower.

A theory of hearing is advanced which pictures the basilar membrane as being caused to vibrate by incident sound waves. In the case of a pure tone, the membrane is supposed not to vibrate uniformly throughout its length but the region of maximum amplitude determines the pitch of the tone as interpreted by the ear and the maximum amplitude determines the intensity.—*Editor.*

THE question of how we hear has been a subject for discussion by scientists and philosophers for a long time. Practically every year during the past fifty years articles have appeared discussing the

* Presented at the meeting of the Section of Physics and Chemistry of The Franklin Institute held Thursday, March 29, 1923, and published in the *Journal of the Franklin Institute* for September, 1923.

pros and cons of various theories of hearing. These discussions have been participated in by men from the various branches of science and particularly by the psychologists, physiologists, otologists, and physicists. During the past two or three years this discussion has been particularly acute. It is not uncommon to pick up an article and read in the beginning or concluding paragraphs statements such as the Helmholtz theory of audition seems to have sunk beyond recovery, [90, 65] † and at the same time an article written probably a month later will have the conclusion that the Helmholtz theory of audition is definitely established beyond all controversy.[70-75]

There is apparently a great deal of misunderstanding between various writers because of different points of view due to different training. To the physicist it seems that most of the discussions show a profound ignorance of the dynamics of the transmission of sound by the mechanism of the ear. Those discussions by the physicists are frequently open to criticism by the otologist and psychologist, due to his lack of knowledge of the structure of the ear or the mental reaction involved in the process of interpretation. I think it is fortunate that some of these scientists from the different branches are now cooperating in their research work as is evinced by the appearance of several joint papers. (Papers by Dean and Bunch, Minton and Wilson, Wegel and Fowler, Kranz and Pohlman, and others.)

It is not my purpose to discuss the merits of the various theories of hearing, but I desire to present some of the facts of audition which have been recently determined with considerable accuracy, and then discuss the theory of hearing which best explains these facts.

Hearing is one of the five senses. It is that sense that makes us aware of the presence of physical disturbances called sound waves. For my purpose, sounds may be classified into two groups, namely, pure tones and complex sounds. A pure tone is specified psychologically by two properties, namely, the pitch and the loudness. These sensory properties are directly related to the physical properties, frequency and intensity of vibration. Mixtures of pure tones of different loudness, but of the same pitch, fall under the first class, since such mixtures give rise to a pure tone. The complex sounds are varying mixtures of pure tones. It will be noticed that phase has not been taken into account. Except when using the two ears for locating the direction of sources of sound, phase differences are not ordinarily appreciated by the ear.*

† These numbers refer to the bibliography at the end of the paper.

These tones are usually transmitted by means of air waves through the outer ear canal to the drum of the ear. From here the vibrations are transmitted by means of the bones in the middle ear to the mechanism of the inner ear.

Those facts of audition which are familiar to almost everybody are as follows:

1. Pure tones are sensed by the ear and differentiated by means of the properties *pitch* and *loudness*.

2. When two notes, separated by a musical interval, are sounded together, they are sensed as two separate notes. They would never be taken for a tone having the intermediate pitch. In this respect, hearing is radically different from seeing. When a red and a green light are mixed together, the impression received by the eye is that of yellow, an intermediate color between the two.

3. There is a definite limiting difference in pitch that can just be sensed.[1-9]

4. There is a definite limiting difference in intensity that can just be sensed.[10-14]

5. There is a minimum intensity of sound below which there is no sensation.[15-31]

6. There is an upper limit on the pitch scale above which no auditory sensation is produced.[32-44]

7. There is a lower limit on the pitch scale below which there is no auditory sensation produced.[45-51]

8. The ear perceives tones separated by an octave as being very similar sensations.

Another quality of audition which is not so commonly known was pointed out by A. M. Mayer.[52] He stated that high tones can be completely masked by louder lower tones while intense higher tones cannot obliterate lower ones though the latter are very weak. Experiments to be described later in the paper show that this statement must be modified somewhat. Very intense low ones will produce a masking effect upon still lower tones, although the masking effect is very much more pronounced in the opposite case. Many of the opponents of the Helmholtz resonant theory of hearing claim that this fact is fatal to such a theory.[52]

* This statement may require modification when more experimental data are available. As shown later in the paper the middle ear has a non-linear response. Consequently it would be expected that phase differences, especially between tones which are harmonic, would produce spacial differences in nerve stimulation.

Limits of the Field of Audition

The new tools which have made possible more accurate measurements in audition are the vacuum tube, the thermal receiver and the condenser transmitter. When connected in a proper arrangement of circuits, the vacuum tube is capable of generating an oscillating electrical current of any desired frequency. This electrical vibration is translated into a sound vibration by means of the telephone receiver. Between the receiver and the oscillator, a wire network called an attenuator [28] is interposed which makes it possible to regulate the

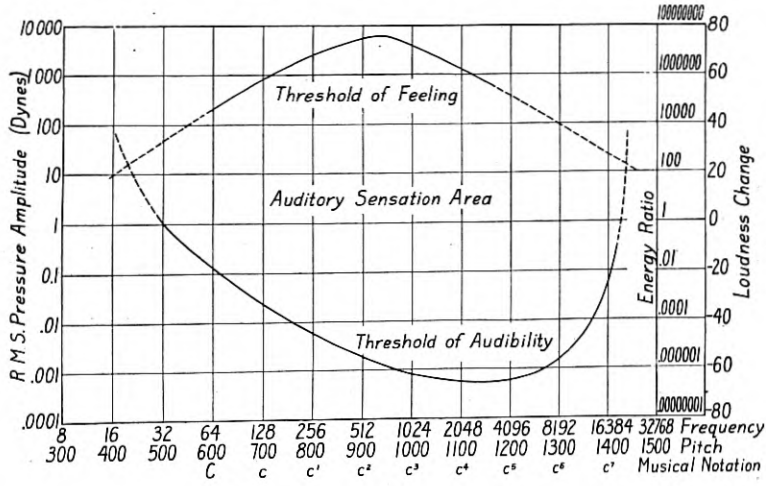


Fig. 1

volume of sound. The theory [54-55] of the thermal receiver has been worked out so that it is possible to calculate its acoustic output from the electrical energy it is absorbing. In this way, it is possible to calculate the pressure variation produced in the outer ear canal when a tone is being perceived. A detailed description of the apparatus and method used in such measurements was given in a paper presented before the National Academy of Science, November 14, 1921.[55] Such a combination of apparatus which has been calibrated is called an audiometer and is suitable for measuring abnormal as well as normal hearing. A receiver more rugged than the thermal may be substituted when its efficiency compared to the thermal receiver is known for all frequencies. By using such an audiometer the average absolute sensitivity for approximately 100 ears which were considered to be normal was determined. The lower curve in

Fig. 1, labelled the threshold of audibility, shows the results of such measurements. The ordinates give the amplitude of the pressure variation in dynes per square centimeter that is just sufficient to cause an auditory sensation and the abscissæ give the frequency of vibration of the tone being perceived. Both are plotted on a logarithmic scale. The experimental difficulties made it impossible to make a very accurate determination for those parts of the curve shown by dotted lines. More work needs to be done on these portions of the curve. In the important speech range, namely, from 500 to 5,000 cycles, it requires approximately .001 of a dyne pressure variation in the air to cause an auditory sensation. This corresponds to a fractional change of about one-billionth in the atmospheric pressure, which shows the extreme sensitiveness of the hearing mechanism.

In order to obtain an idea of the intensity range used in hearing, an attempt was also made to obtain an upper limit for audible intensities. When the intensity of a tone is continually increased, a value is reached where the ear experiences a tickling sensation. Experiments show that the intensity for this sensation is approximately the same for various individuals and the results can be duplicated as accurately as those for the minumum intensity value. It was found that if this same intensity of sound is impressed against the finger, it excites the tactile nerves, In other words, the sensation of feeling for the ear is practically the same as for other parts of the body. When the intensity goes slightly above this feeling point, pain is experienced. Consequently, this intensity for the threshold of feeling was considered to be the maximum intensity that could be used in any practical way for hearing. The two points where these two curves intersect have interesting interpretations. At these two points, the ear both hears and feels the tone. At frequencies above the upper intersecting point, the ear feels the sound before hearing it, and in general would experience pain before exciting the sensation of hearing. Consequently, the intersection point may be considered as the upper limit in pitch which can be sensed. In a similar way, the lower intersection point represents the lowest pitch than can be sensed.

There has been considerable work [32-51] in the past to determine the upper frequency and lower frequency limits of audibility, but it would appear that without the criterion just mentioned, such limiting points apply only to the particular intensity used in the determination. Not enough attention has been paid to the intensity of the tones for such determinations. It is quite evident from this

figure that both the upper and lower limits of audibility which are found in any particular experimental investigation will very largely depend upon the intensity of the tones sounded. For example, if the intensity were along the .01 dyne line, the limits would be 200 and 12,600 cycles.

The area enclosed between the maximum and minimum audibility curves has been called the auditory-sensation area and each point in it represents a pure tone. The question then arises: How many such pure tones can be sensed by the normal ear?

The answer to this question has been made possible by the recent work of Mr. V. O. Knudsen.[14] In this work Knudsen made determinations of the sensibility of the ear for small differences in pitch and intensity. In Fig. 2, the average results of his measurements for

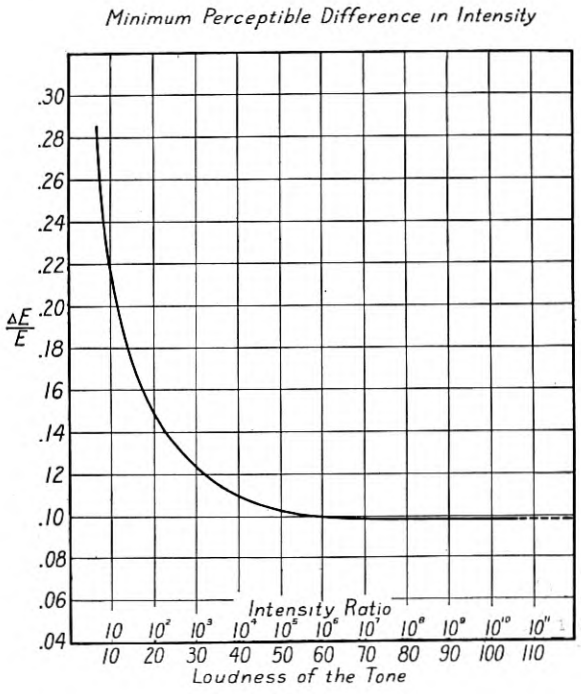


Fig. 2

changes in intensity are shown. Each ordinate gives the fractional change in the sound energy which is just perceptible, this fractional change being called the Fechner ratio. The abscissæ are equal to ten times the logarithm of the ratio of intensities, the zero corre-

sponding to the intensity at the threshold of audibility. For intensities greater than 10^4 times the threshold of audibility, the Fechner ratio has the constant value of approximately one-tenth. It was found that this ratio is approximately the same for all frequencies. In Fig. 3 is shown the results taken from Knudsen's article on the

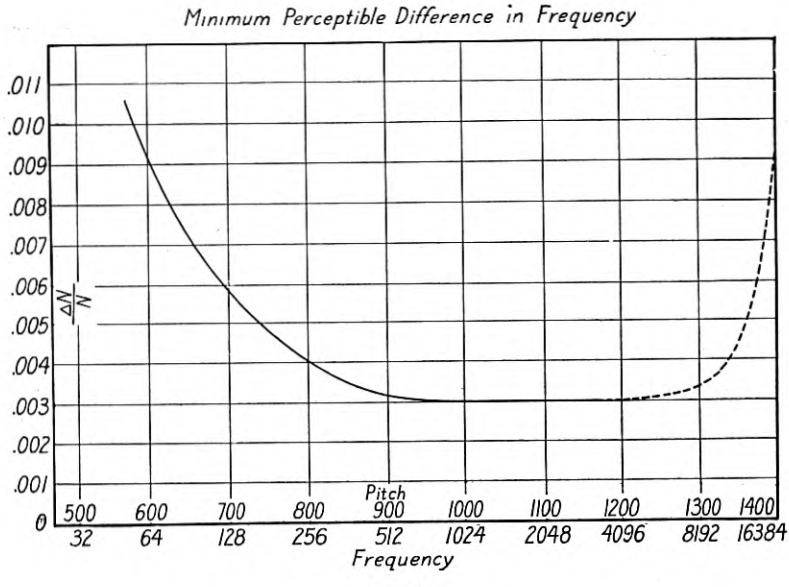


Fig. 3

pitch sensibility. The ordinates give the fractional change in the frequency which is just perceptible and the abscissæ give the frequency on a logarithmic scale. The meaning of the pitch scale at the bottom of this figure will be discussed later. For frequencies above 400 this fractional change is a constant equal to .003. This ratio probably becomes larger again for the very high frequencies. It was found that it varied with intensity in approximately the same way as that given for the energy ratio.

Using these values in connection with the auditory-sensation area, it is possible to calculate the number of pure tones which the ear can perceive as being different. For example, if, starting at the minimum audibility curve, ordinate increments are laid off along a constant pitch line, that are successively equal to the value of ΔE at the intensity position above the threshold, then the number of such increments between the upper and lower curves in Fig. 1 is equal to the number of pure tones of constant pitch that can be per-

ceived as being different in volume. If the minimum and maximum audibility curves were plotted on an energy scale, the increment length ΔE near the maximum audibility curve would be a million million times longer than its length in the minimum audibility curve, whereas when they are plotted on a logarithmic scale, this increment length remains approximately constant, changing by less than a factor 2 for 90 per cent. of the distance across the auditory-sensation area. The calculation shows (see Appendix A) that the number of such increments on the 100-cycle frequency line is 270, that is, 270 tones having a frequency of vibration of 1,000 cycles can be perceived as being different in loudness.

What has been said of the intensity scale applies equally well to the frequency scale. The calculation (see Appendix A) indicates that the number of tones that are perceivable as being different in pitch along the 10-dyne pressure line is approximately 1,300.

If an ordinate increment corresponding to ΔE and an abscissa increment corresponding to ΔN be drawn, a small rectangle will be formed which may be considered as forming the boundary lines for a single pure tone. All tones which lie in this area sound alike to the ear. The number of such small rectangles in the auditory-sensation area corresponds to the number of pure tones which can be perceived as being different. The calculation (see Appendix A) of this number indicates that there are approximately 300,000 such tones.

One might well ask the question: How many complex sounds which are different can be sensed by the ear? At first thought, one might say that this number is represented by all the possible combinations of pure tones. Of course, such a number would be entirely too large, for some of these would sound alike to the ear, since the louder tones would necessarily mask the feebler ones. It is evident, however, that the number of such complex sounds will be very much larger than the number of pure tones.

Scales of Loudness and Pitch

It is seen that the use of the logarithmic scale in Fig. 1 is much more convenient not only on account of the large range of values necessary to represent the auditory-sensation area, but also because of its scientific basis. Psychologists have recognized this since Weber and Fechner formulated the relation between the sensation and the stimulus. Although logarithmic units have been used by various authors in measuring the amount of sensation, the numerical values have been quite different. It seems inevitable that there will be a

greater cooperation in the future between men in the various branches of science working on this subject, so, in order to avoid misunderstanding, it would be very advantageous for all to use, as far as possible, the same units. With this in mind, I am taking the liberty of suggesting for discussion units for both loudness and pitch.

In the telephone business, the commodity being delivered to the customers is reproduced speech. One of the most important qualities of this speech is its loudness, so it is very reasonable to use a sensation scale to define the volume of the speech delivered. At the present time, an endeavor is being made to obtain an agreement of all the telephone companies, both in the United States and abroad, to adopt a standard logarithmic unit for defining the efficiency of telephone circuits and the electrical speech levels at various points along the transmission lines. The *chief interest* in changes in efficiency of transmission apparatus is their effects upon the loudness of the speech delivered by the receiver at the end of the telephone circuit. So it would be very advantageous to use this same logarithmic scale for measuring differences in loudness.

This scale is chosen so that the loudness difference is ten times the common logarithm of the intensity ratio. This means that if the intensity is multiplied by a factor 10, the loudness is increased by ten; if the intensity is multiplied by 100, the loudness is increased by 20; if the intensity is multiplied by 1,000, the loudness is increased by 30, etc. It was seen above that under the most favorable circumstances a change in loudness equal to 1/2 on this scale could just be detected. Knudsen's data indicate, however, that when a silent interval of only two seconds intervenes between the two tones being compared, a loudness change greater than unity on this scale is required before it is noticeable. So the smallest loudness change that is ordinarily appreciated is equivalent to one unit on this scale. It is also convenient because of the decimal relation between loudness change and intensity ratio. This relation is expressed by the formula:

$$\Delta L = L_1 - L_2 = 10 \log_{10} \frac{I_1}{I_2} \text{ or } \frac{I_1}{I_2} = 10^{\frac{\Delta L}{10}}$$

where L_1 and L_2 are the two loudness values corresponding to the intensities I_1 and I_2. Since intensities of sound are proportional to the square of pressure amplitudes this may also be written:

$$\Delta L = 20 \log \frac{p_1}{p_2}$$

The most convenient choice of the intensity or pressure used as a standard for comparison depends upon the problem under consider-

ation. In the sensation area chart of Fig. 1, the intensity line corresponding to one dyne was used as the zero level, that is, p_2 was chosen equal to 1 so that

$$\Delta L = 20 \log p$$

The choice of the base of logarithms for the pitch scale is dictated by the fact mentioned before, that the ear perceives octaves as being very similar sensations. Consequently the base 2 is the most logical choice for expressing pitch changes. If the logarithm of the frequency to the base 2 were used, perceptible changes in pitch would correspond to inconveniently small values of the logarithm. It is better to use the logarithm to the base $\sqrt[100]{2}$ which is 100 times as large. On this scale the smallest perceptible difference in pitch is approximately unity—somewhat more for frequencies greater than 100 cycles or somewhat less for lower frequencies, according to Knudsen's data. The scale on the charts is chosen so that the change in pitch is given by

$$\Delta P = 100 \log_2 N$$

where N is the frequency of vibration.

It is now evident why such pitch and loudness scales were used in Fig. 1. With these scales, the number of units in any area gives approximately the number of tones that can be ordinarily appreciated in that area. For example, there are approximately 2,000 distinguishable tones in each square, there being more near the centre and fewer near the boundary lines than this number.

Experiments have shown that pure tones of different frequencies which are an equal number of units above the threshold value sound equally loud. This statement may require modification when very loud tones are compared, but the data indicated that throughout the most practical range this was true. Consequently, the absolute loudness of any tone can be taken as the number of units above the threshold value.

Loudness of Complex Sounds

In the measurement of the loudness of complex tones, the situation is not so simple. It has been found that if two complex tones are judged equally loud at one intensity level and then each is magnified equal amounts in intensity, they then may or may not sound equally loud. The curves shown in Figs. 4, 5 and 6 will illustrate this. The first (Fig. 4) shows the comparisons at different intensity levels of two sounds whose pressure spectra are shown in the two figures at

the top. The x-axis gives units above threshold for sound A and the y-axis gives the units above threshold of the sound B when the two sound equally loud. In this case, the spectra are somewhat similar

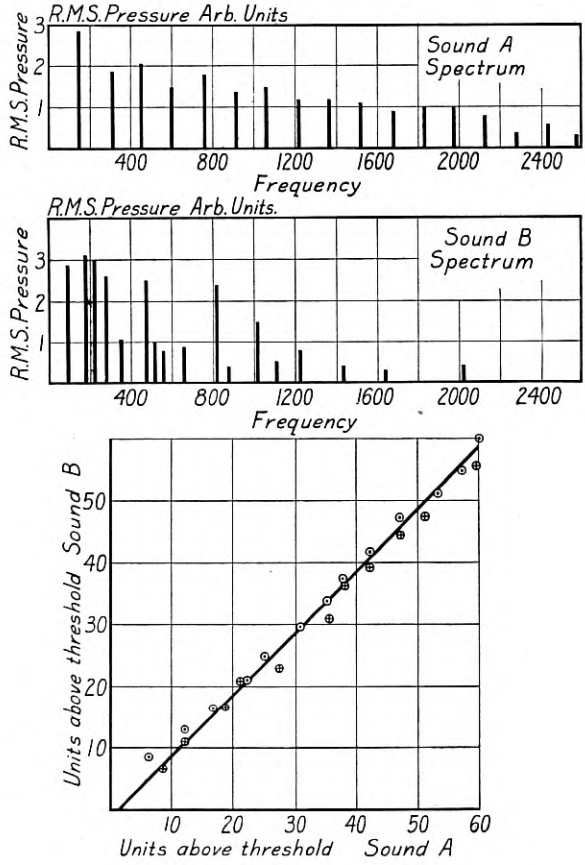


Fig. 4

and we have a straight line of slope 45° passing nearly through the origin. The two sounds are thus of practically equal loudness when they are the same number of units above threshold. In Fig. 5 we have similar data for two sounds which have quite different spectra as is indicated by the two charts at the top. The curve for C means that it was a practically continuous spectrum. It was produced by a device for making the "swishing" type of noises which are usually so prominent in office rooms. The curve representing the relation is not straight, since for values of intensity near the threshold, the

loudness increases faster for the C sound for increments in the intensity than for the A sound. For example, it is seen that when the sound C is 30 units above the threshold, the sound A is 45 units

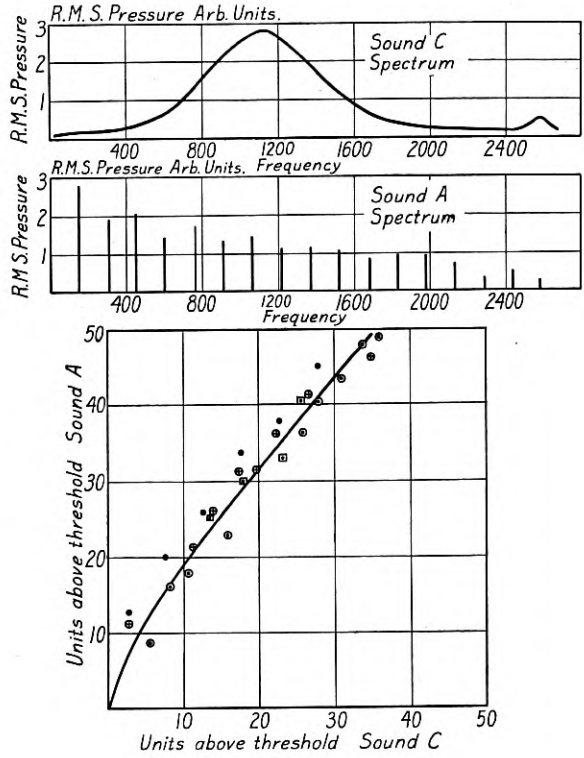


Fig. 5

above the threshold when the two sound equally loud. In Fig. 6 a comparison is given between the loudness of a pure tone of 700 cycles and a complex sound designated by A in the last figure. In this case again the relation is expressed by a curve. The technique of making such loudness measurements is rather difficult and requires a large number of observations before the values are reliable. A paper on this subject which will soon be published will give a detailed account of this work on loudness.

Enough data have been given to show that in order to give loudness a definite meaning for complex sounds, a more precise definition is necessary. It has been found convenient to define the loudness of any complex or pure tone in terms of the loudness of a sound standard.

This standard is a pure tone having a vibration frequency of 700 cycles per second. Its absolute loudness is defined as the change in loudness measured on the scale defined above, from the loudness value corresponding to the threshold pressure for normal ears which for 700 cycles is exactly 0.001 dyne. This frequency was arbitrarily chosen as a standard for measuring loudness because of this particular value of its threshold pressure, and because it is close to the frequency

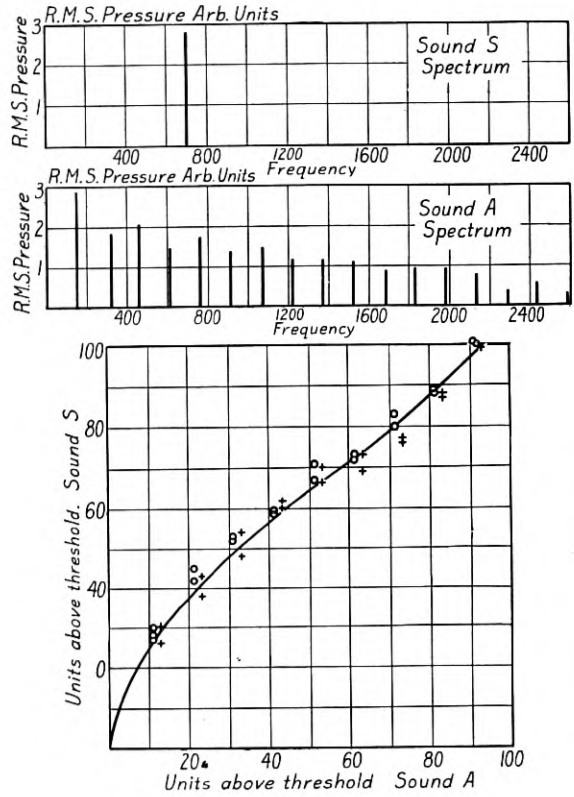


Fig. 6

at which the loudest tones used in conversational speech occur. By this definition, the loudness of a tone of frequency 700, for which p is the pressure variation, expressed as a root mean square value,

$$L = 60 + 20 \log p$$

and the loudness of any other sound, pure or complex, is defined as being equal to that of a tone of frequency 700, seeming equally loud. Such a definition implies that experimental measurements can be

made to determine when any complex sound is equally loud to a 700-cycle tone. Such measurements can be made although the observational error is rather large and the judgment of various individuals is sometimes quite different, which means only that loudness as measured by various individuals is different. For use in engineering work, however, the average of a large number of individuals can be taken and this loudness will have a definite determinable value. For example in Fig. 6, the loudness of the A sound when it is 60 units above the threshold is 72, since it sounds as loud as a 700-cycle tone which is 72 units above its threshold. The loudness of complex sounds usually increases faster with increases in intensity than that of pure tones. This would be expected since the threshold is determined principally by the loudest frequency in the complex sound and as the intensity is increased the other frequencies begin to add to the total loudness.

Since pure tones of different pitches which are the same number of units above the threshold sound equally loud their loudness L can be represented by the formula

$$L = L_0 + 20 \log p$$

where p is the root mean square value of the pressure amplitude produced in the ear by the tone and L_0 is the number of units from the 1-dyne line to the minimum audibility curve. The values of L_0 can be read directly from the chart in Fig. 1.

Measurement of Degree of Deafness

The choice of the loudness and pitch units used above leads to a *rational definition of the degree of deafness.*

The number of possible pure tones that can be sensed by a deaf person is considerably smaller than that mentioned above obtained from the normal auditory-sensation area. A logical way of defining the amount of hearing is: *To give the per cent. of the total number of distinguishable pure tones audible to a person with normal hearing, that can be sensed by the deaf person.*

Some definition of this sort will be very helpful in clearing up the confusion that now exists in court cases involving the degree of deafness. It is well known that there are a number of laws which prevent people who have more than a defined amount of deafness from doing certain classes of work. For example, one cannot operate an automobile if he has a certain per cent. of deafness. At the present time, there is a large variation between the standards set up by the various doctors in different parts of the country.

From the discussion above it was seen that the number of tones corresponding to any region was approximately proportional to the area of that region when the logarithmic units were used. Consequently the per cent.* of hearing can be taken as the fractional part

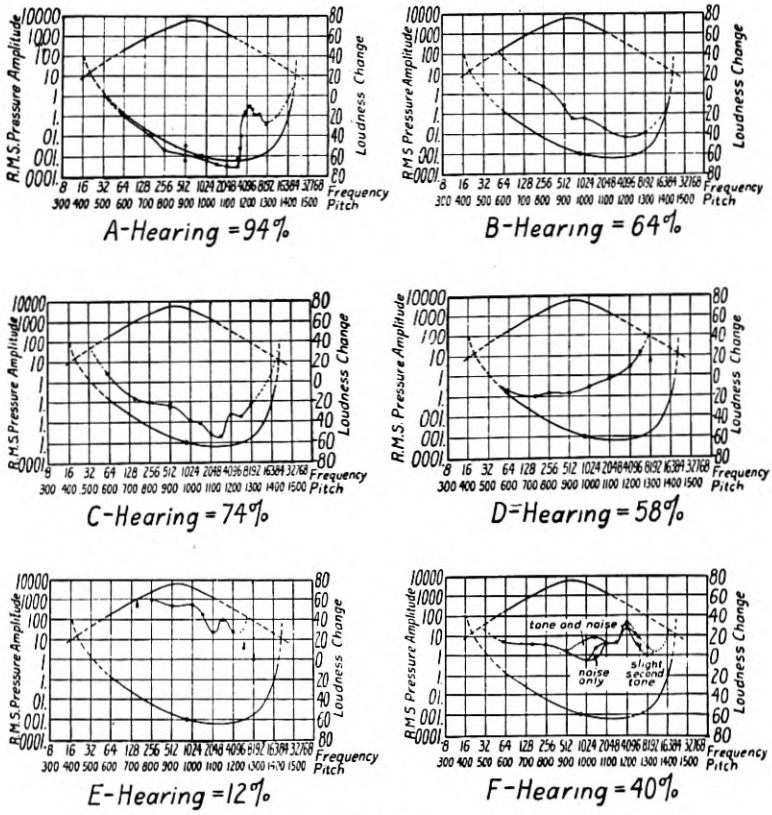


Fig. 7

Audiograms for Typical Cases of Deafness

of the normal auditory-sensation area in which tones can be properly sensed. The per cent. of deafness is, of course, 100 minus the per cent. of hearing.

To emphasize the meaning of this definition, some audiograms, that is minimum audible intensity curves, for some typical cases of deafness will be given. These are shown in Fig. 7. The first chart

* This assumes that the Fechner ratio for pitch and loudness is approximately the same for one having abnormal as for one having normal hearing.

shows a common type of deafness in which the sensitivity to the high frequencies suddenly decreases, as is indicated by the rise in the minimum audible intensity curve when the frequency exceeds 3,000 cycles per second. The sensation area for this person is 94 per cent. of that for the average. Consequently, his per cent. of hearing is 94 per cent. It is also convenient to speak of the per cent. of hearing for each pitch. It is evident that the logical definition for this is the ratio of the widths of the sensation area for the person tested and normal person, measured along the ordinate drawn at the frequency in question.[56] For example, in this audiogram the person had more than 100 per cent. hearing for most of the pitch range. At 4,000 cycles, however, the per cent. hearing was only 60 per cent. This means that for this pitch, the person when compared with one having normal hearing could sense only 60 per cent. as many gradations in tonal volume before reaching the threshold of feeling.

The second chart corresponds to a type of deafness that is not so common. It shows relatively large losses at the lower frequencies. The per cent. hearing in this case is seen to be 64 per cent.

The third type is very common and corresponds to a general lowering of the frequencies throughout the entire pitch range. In these first three cases, the deaf persons could carry on a conversation without any difficulty whatever. In the last two of these, difficulty was experienced in understanding a speaker at any considerable distance. In the first case, the person could not hear the steam issuing from a jet or any other high hissing sound. However, he could hear and understand speech practically as well as anyone with normal hearing.

The fourth case shows a falling off at the high frequencies, but this loss in hearing proceeds gradually as the pitch increases rather than abruptly as in the first case. As indicated in the figure the per cent. of hearing is 58 per cent.

The fifth case is one of extreme deafness and is typical of such cases. The per cent. of hearing is only 12 per cent. The last case shows not only the minimum audibility curve, but the quality of the sensation perceived. As indicated on the chart, in certain regions noises are heard when the stimulus is a pure tone. When computing the per cent. of hearing in such cases, it seems reasonable to take only the area where sensation of good quality is perceived. In some cases, this poor quality extends through practically the whole area and although the person hears sounds, he is unable to properly interpret them. Consequently, from a practical point of view, his per cent. of hearing is very low. For such cases, deaf sets or other aids to the hearing do not give any satisfactory help.

Masking of One Pure Tone by Another

We are now in a position to discuss another set of facts concerning the perception of tones, namely, the ability of the ear to perceive certain sounds in the presence of other sounds. Such data for pure tones have been obtained in our laboratories and will soon be published in some detail. The apparatus used consisted simply of two vacuum tube oscillators generating the two tones used and two attenuators which made it possible to introduce the tones into a single receiver with any desired intensities. In other words, it consists of two audiometers with a common receiver for generating the two tones. The curves shown in Fig. 8 give the general character of the results of this work.

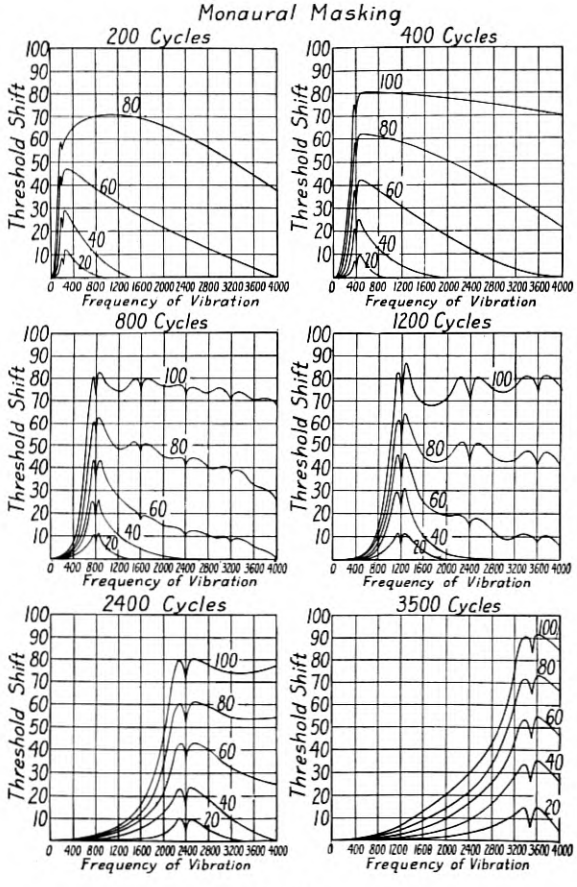


Fig. 8

The ordinates show the amounts in loudness units that the threshold value of a tone of any frequency called the "masked tone" is shifted due to the presence of another tone called the "masking tone." The frequency of the masking tone is given at the top of each set of curves.

The experimental procedure was as follows: The threshold values for the two tones were first determined. The intensity of the masking tone (the frequency of which is given above each graph) was then increased beyond its threshold value by the number of units indicated just above the curve. The masked tone was then increased in intensity until its presence was just perceived. The amount of this latter increase, measured on the loudness scale is called the *threshold shift* and is plotted as ordinate in Figs. 8, 9 and 10. The frequencies of the masked tones are given by the abscissæ.

For example, in the fourth chart, the masking effects of the tone having a frequency of 1,200 cycles are shown. It is seen that the greatest masking effect is near 1,200 cycles, which is the frequency of the masking tone. A tone of 1,250 cycles must be raised to 46 units above the threshold to be perceived in the presence of a 1,200-cycle tone which is 60 units above its threshold, or it must be raised to within 14 units of the masking tone before it is perceived. This corresponds to an intensity ratio between the tones of only 25. A tone of 3,000 cycles, however, can be perceived in the presence of a 1,200-cycle tone which is 60 units loud when it is only 8 units above its threshold. This means that the intensity ratio between these two tones, under such circumstances, corresponds to 52 units or to a ratio of approximately 160,000 in intensity. However, as the loudness of the masking tone is increased, all of the high tones must be increased to fairly large values before they can be heard. For example, the high frequencies must be raised 75 units above the threshold to be heard in the presence of a 1,200-cycle tone having a loudness of 100 units. But even for such large intensities for the masking tone, those frequencies below 300 are perceived by raising their loudness only slightly above the threshold value. It should be noticed that in all cases, those tones having frequencies near the masking frequency, whether they are higher or lower, are easily masked.

It is thus seen that Mayer's conclusion, that a low pitch sound completely obliterates higher pitched tones of considerable intensity and that higher pitched frequencies will never obliterate lower pitched tones, is true only under certain circumstances. A low tone will not obliterate to any degree a high tone far removed in frequency, except when the former is raised to very high intensities. Also a tone of higher frequency can easily obliterate a tone of lower fre-

quency if the frequencies of the two tones are near together. When the two tones are very close together in pitch the presence of the masked tone is perceived by the beats it produces. This accounts for the sharp drop in the curves at these frequencies. A similar thing happens for those regions corresponding to harmonics of the masking frequency. In the charts for the 200- and 400-cycle masking tones these drops are not shown inasmuch as they were small, but in an accurate picture they should be shown.

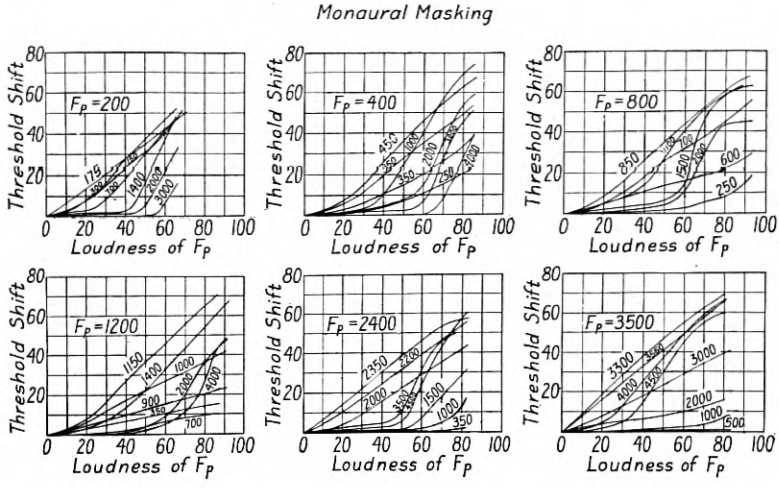


Fig. 9

In Fig. 9, these results are shown plotted in a different way. The abscissæ represent the loudness of the primary tones whose frequency is indicated at the top of each of the charts. The amounts that the threshold is shifted are plotted as ordinates as in the previous figure. For example, in Chart 1, the results are shown for a masking tone of 200 cycles. The curve marked 3,000 indicates the masking effect of a 200-cycle upon a 3,000-cycle tone. It is seen that the loudness of the low pitched tone can be raised to 55 units before it has any interfering effect upon the high pitched tone. For louder values than this it has a very marked effect. It will be noticed that in nearly all of the charts the curves for different frequencies intersect. This leads to some rather interesting conclusions, regarding the perception of a complex tone. For example, consider the curves for a masking tone having a frequency of 400 cycles. Assume we have a complex tone having three frequencies of 400, 300 and 200 cycles with relative loudness values of 50, 10 and 10, respectively. The ear will hear only

the 400-cycle tone and the 200-cycle tone as is evident from the curves. It would be necessary to raise the 300-cycle tone above 16 units for it to be heard in the presence of 400 cycles of loudness 50. However, if the sound is magnified without distortion 30 loudness units, so that these three frequencies have loudness values of 80, 40

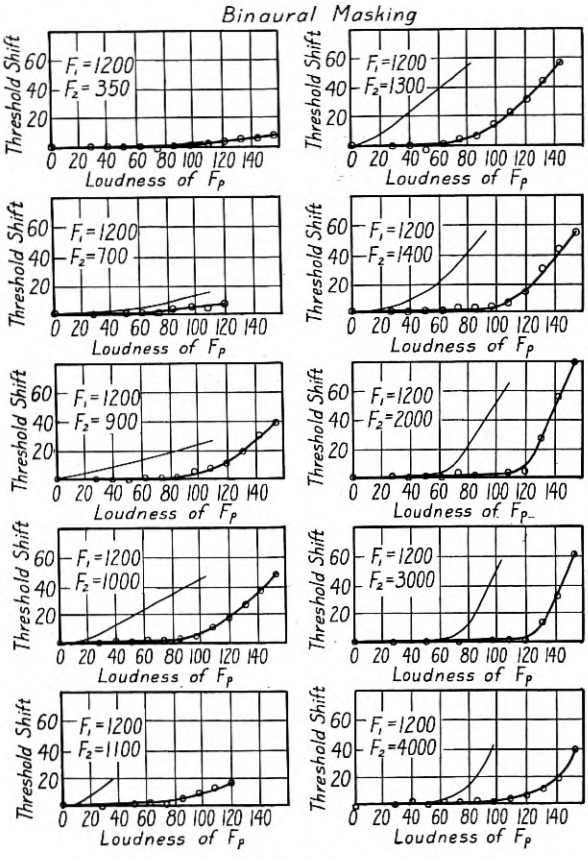


Fig. 10

and 40, respectively, then the 400-cycle tone and 300-cycle tone only will be heard. Under such conditions, the 300-cycle tone could be attenuated approximately 15 units before it would disappear. This means that the sensation produced by a complex sound is different in character as well as intensity when the sound is increased or decreased in intensity without distortion. In general, as the tone becomes more intense the low tones become more prominent because the high

tones are masked. It is a common experience of one working with complex sounds to have the low frequencies always gain in prominence as the sound is amplified.

The question naturally arises, Does the same interfering effect exist when the two tones are introduced into opposite ears instead of both being introduced into the same ear ? The answer is No. Curves showing the results in such tests are shown in Fig. 10. For comparison the results for the case when in tones are both in the same ear are given by the light lines. Take the case of 1,200 and 1,300 cycles. It is rather remarkable that a tone in one ear can be raised to 60 units, that is, increased in intensity one million times, before the threshold value for the tone in the other ear is noticeably affected. If the 1,300-cycle tone were introduced into the same ear as the 1,200-cycle tone, its loudness would need to be shifted 40 units, corresponding to a 10,000-fold magnification in intensity above its threshold intensity in the free ear before it can be heard. It is seen that if one set of curves is shifted about 50 units it will coincide with the second set. This strongly suggests that the interference in this case is due to the loud tone being transmitted by bone conduction through the head with sufficient energy to cause masking. The vibration is probably picked up by the base of the incus and transmitted from there to the cochlea in the usual way. There is other evidence * which I shall not have space here to discuss, which indicates that the effective attenuation from one ear to the other is approximately 50 units.

Theories of Audition

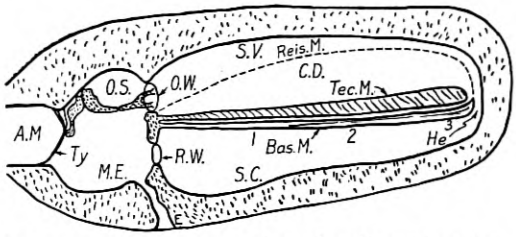
With these facts in mind, we are now ready to discuss the theory of hearing which will best account for them. I will refer briefly to just a few of the principal theories of hearing which have been proposed. The sketch shown in Fig. 11 gives a diagrammatic picture of the internal ear. In the Helmholtz theory, as first formulated, it is stated that the organ of Corti located between the basilar membrane and the tectorial membrane act like a set of resonators which are sharply tuned. Each tone stimulates a single organ depending upon its pitch. Later this theory was somewhat modified as it was thought that the resonant property might reside in one of the membranes in the cochlea.

* See paper by Wegel and Lane soon to be published in the *Physical Review* entitled "The Auditory Masking of One Pure Tone by Another and its Relation to the Dynamics of the Inner Ear."

In the "telephone" theory, as expounded by Voltalini, Rutherford, Waller and others, it is assumed that the basilar membrane vibrates as a whole like the diaphragm of a telephone receiver, and consequently responds to all frequencies with varying degress of amplitude. The discrimination of pitch takes place in the brain.

Meyers in his theory states that various lengths of the basilar membrane are set in motion depending upon the intensity of the stimulating tone. As in the previous theory, the pitch discrimination is accomplished in some way in the brain.

In the "non-resonant" theory of Emile ter Kuile it is assumed that the sound disturbance penetrates different distances into the



Diagrammatic representation of auditory function

A.M.	Auditory meatus	O. W.	Oval window
Bas. M.	Bas. mem. including organ of Corti	Reis. M.	Reissner's mem.
C. D.	Cochlear duct	R. W.	Rd. window
E.	Eustachian tube	S. C.	Scala cochlea
He.	Helicotrema	S. V.	Scala vestibuli
M. E.	Middle ear	Tec. M.	Tectorial membrane
O. S.	Ossicles (malleus, incus, stapes)	Ty.	Tympanic membrane

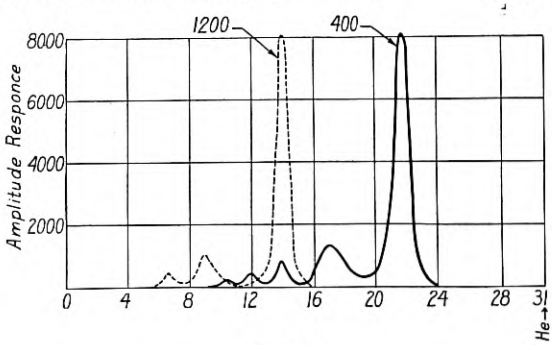


Fig. 11

cochlea depending upon the frequency of the stimulating tone. The further along the membrane the disturbance reaches, the lower will be the pitch sensation. A low pitch tone then stimulates all of the nerve fibres that would be stimulated by tones of higher pitch plus some additional nerve fibres.

The theory of maximum amplitudes was first put into definite form by Gray in 1899.[69] It assumes that the position of maximum amplitude of the basilar membrane varies with the pitch of the stimulating tone. Although a considerable portion of the membrane vibrates when stimulated by a pure tone, the ear judges the pitch by the position of maximum response of the basilar membrane. Roaf has shown that some action of this sort must take place due to the dynamical constants involved.[61] It is an amplification of this theory that I desire to propose as the one which most satisfactorily accounts for the facts.

When a sound wave impinges upon an ear-drum, its vibrational motion is communicated through the middle ear (Fig. 11) by means of the chain of small ossicles (malleus, incus and stapes) to the oval window. Here the vibration is communicated to the fluid contained in the cochlea. If the pitch of the tone is low, say below 20 vibrations, the fluid is moved bodily back and forth around the basilar membrane through the helicotrema, the motion of the membrane at the round window and the oval window being just opposite in phase, the former moving inward while the latter moves outward. For very high frequencies, the mass reactions of the ossicles and the fluid are so great that very little energy can be transmitted to the cochlea. For example, when the elastic forces are negligible it requires a force 10,000 times as large as produce the same amplitude of vibration at 10,000 cycles as that required at 100 cycles. For intermediate frequencies the mass reactions, the elastic restoring forces and the frictional resistances which are brought into play are such that the wave is transmitted through the basilar membrane causing the nerves to be excited.

It is thus seen that the upper and lower limits of audibility are easily explained. When the forces upon the drum of the ear or walls of the ear canal are large enough to excite the sensation of feeling and the pitch of the tone is either too low or too high to cause any perceptible vibration of the basilar membrane, we are beyond the lower or upper limit of audibility respectively. At frequencies between these limits, the vibrational energy is first communicated to the fluid in the scala vestibuli and then transmitted through the two membranes into the fluid of the scala cochlea. As the basilar membrane transmits the sound wave it takes up a vibration amplitude which stimulates the nerve fibres located in it. The entire membrane vibrates for every incident tone, but for each frequency there is a corresponding spot on the membrane where the amplitude of

the vibration is greater than anywhere else. Our postulate is that only those nerves are stimulated which are at the particular parts of the membrane vibrating with more than a certain critical amplitude; and that we judge the pitch from the part of the membrane where the nerves are stimulated. According to this conception, the variation with frequency of the minimum audible intensity is due principally to the variation with frequency of the transmission efficiency of the mechanical system between the auditory meatus and the basilar membrane. Pure tones of equal loudness correspond either to equal amplitudes or to equal velocities of vibration of the basilar membrane or to some function of the two. Whatever is assumed, the dependence of the minimum audible intensity upon frequency for the ear can be explained entirely by the vibrational characteristics of the ear mechanism. For the sake of clearness it will be assumed that equal amplitudes of vibration of the basilar membrane correspond to equal sensations. For loud pure tones, there are several regions of maximum amplitude on the membrane, corresponding to the tone and to the harmonic introduced by the nonlinear response of the middle ear, the latter maxima increasing very rapidly as the stimulation increases.

It is a strange thing that the phenomenon of the masking of tones which, as stated in the beginning, has been considered by some to be so fatal to any resonator theory, is the very thing that has furnished experimental data which makes it possible to calculate the vibration characteristics of the inner ear. Such a calculation must be based upon assumptions which will be uncertain, but will seem reasonable. It is not my purpose to discuss those here, but I shall give only the final result of such a calculation made by Mr. Wegel and Mr. Lane of our laboratories. At the bottom of Fig. 11, the two curves show the amplitude of vibration of different portions of the basilar membrane for the two frequencies 400 and 1,200 cycles. For purposes here these curves may be considered to be simply illustrative. This membrane has a length of 31 mm. and a width of .2 mm. at the base and .36 mm. at the helicotrema end. The x-axis in this figure gives the distance in millimeters from the oval window and the y-axis gives the amplitudes of vibration in terms of the amplitude corresponding to the threshold of audibility. The loudness of the stimulating tones in both cases is 80 units. It will be seen that the maximum response for the high frequencies is near the base of the cochlea, while that for the low frequencies is near the helicoptrema. It will be noticed that the amplitude of the membrane has several maxima corresponding to the subjective harmonics.

With this picture in mind, it is clear why the perception of one tone is interfered with by the presence of a second tone when their frequencies are close together, since the nerves necessary to perceive the first tone are already stimulated by the second tone. Also when their frequencies are widely separated, entirely different sets of nerves carry the impulses to the brain, and consequently there is no interference between the tones except that which occurs in the brain. Although this brain interference may not be entirely negligible, especially for very loud sounds, it is certainly very much smaller than that existing in the ear for tones close together in pitch.

It is also seen that the reason why the low tones mask the high tones very much more easily that the reverse is due to the harmonics introduced by the transmission mechanism of the ear. Inasmuch as these harmonics are due to the second order modulations, they are proportional to the square of the amplitude and, therefore, become much more prominent for the large amplitudes. When two tones are introduced, summation and difference tones as well as the harmonics will necessarily be present (see Appendix B). With the proper apparatus for generating continuously sounding tones, these subjective tones are easily heard. Their frequency can be quite accurately located by introducing from an external source a frequency which can be varied until it produces beats with the subjective tone.

Messrs. Wegel and Lane who are working in this field have observed modulation frequencies created in the ear as high as the fourth order. They will soon publish* an account of this work on the vibrational characteristics of the basilar membrane. It is seen that the quality as well as the intensity of the sensation produced by a pure tone should change as the intensity of stimulus is increased due to the increasing prominence of the harmonics. This is in accordance with one's experience while listening to pure tones of varying intensity. The non-linear character of the hearing mechanism is also sufficient to account for the falling off in the ability of one to interpret speech when it becomes louder than about 75 units. The introduction of the summation and difference tones and the harmonics makes the interpretation by the brain more difficult. Its action in this respect is very similar to the carbon transmitter used in commercial telephone work or to an overloaded vacuum tub. This characteristic of the ear also explains why we should expect departures from non-linearity when making loudness balances for complex tones. It also suggests that a similar thing might be ex-

* Wegel and Lane, see paper already cited.

pected when comparing the loudness of pure tones if the balances are made at very high intensities. No such balances have yet been made.

What happens to the ear when one becomes deaf? This question, of course, is one for the medical profession to answer, but let us take one or two simple cases and see if they fit into this theory. First assume that the nerve endings are diseased for a short distance away from the base of the cochlea so that they send no impulses to the brain. Under certain assumptions the kind of an audiogram one should obtain can be calculated from the vibrational characteristics determined as mentioned above. Such a calculation shows that an audiogram similar to that shown in Fig. 4, which has a rapid falling off in sensitiveness, can be accounted for, both quantitatively as well as qualitatively. On a pure resonant theory corresponding to that first proposed by Helmholtz, a tone island would exist corresponding to the affected region for such a case. Although we have tested a large number of cases, no such islands have ever been found. When the intensity of the tone is raised sufficiently to bring the amplitude of the area containing the healthy nerve cells which are adjacent to the diseased portion to a value above that corresponding to the threshold, the tone will then be perceived.

Again assume that due to some pathological condition, the tissue around the oval window where the stapes join the cochlea has become hardened. Its elasticity will then be greatly increased so that vibrational energy at low frequencies will be greatly discriminated against. For such a case, an audiogram similar to that shown in Fig. 7-B would be obtained.

A number of things can cause a general lowering of the ear sensitivity, such as wax in the ear canal, affections of the ear-drum, fixation of any of the ossicles, thickening of the basilar membrane, affections of the nerve endings or loss in nervous energy being supplied to the membrane, etc. However, one would expect that each type of trouble would discriminate, at least to some extent, against certain frequency regions so as to produce some characteristic in the audiogram. Ear specialists are beginning to realize the possibility of obtaining considerable aid in the diagnosis of abnormal hearing from such accurate audiograms.

There are a large number of facts obtained from medical research which necessarily have a bearing upon the theory of hearing, but as far as I know none of them is contrary to the theory of hearing given above. It was seen that there are approximately 300,000 tone units in the auditory-sensation area. According to the anatomists, there are

only 4,000 nerve cells in the basilar membrane with four or five fibre hairs for each cell. Assuming that each hair fibre acts as a unit there are still insufficient units for each perceivable tone and according to the theory given above, a large number of these units must act at one time. Consequently the ear must be able to interpret differences in the intensity of excitation of each nerve cell as well as determine the position of each nerve cell excited.

Most modern neurologists believe in the "none or all" excitation theory of nerve impulses.[59-60] They also claim that nerve impulses can never be much more rapid than about 50 per second and cannot therefore follow frequencies as high as those found in sound waves. The second statement only emphasizes the necessity of assuming that the intensity position as well as place position is necessary to

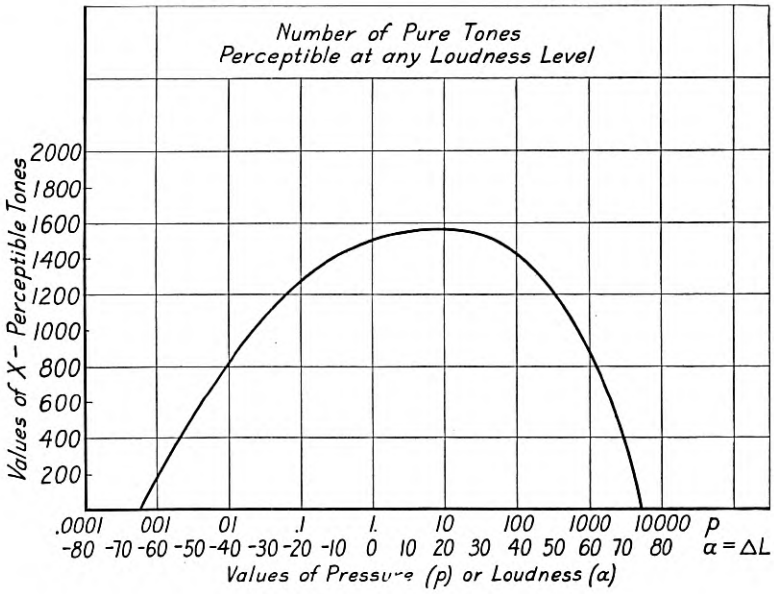


Fig. 12

account for the differentiation of pure tones. The first statement is not necessarily in conflict with such an idea since anatomists are not agreed upon the number of nerve fibres radiating from each nerve cell. Since each nerve fibre can serve to give a unit nerve impulse, the intensity of stimulation sent from a single nerve cell can increase with stimulation depending upon the number of nerve fibres brought into action. The intensity of the sensation produced

is then directly related to the total number of nerve fibres giving off impulses. It seems to me that the spacial and intensity configurations which are possible, according to this theory, are sufficient for an educated brain to interpret all the complex sounds which are common to our experience.

In conclusion then, it is seen that the pitch of pure tones is determined by the position of maximum response of the basilar membrane, the high tones stimulating regions near the base and the low tones regions near the apex of the cochlea.

A person can sense two mixed tones as being distinctly two tones while he cannot sense two mixed colors, since in the ear mechanism there is a spacial frequency selectivity while in the eye mechanism there is no such selectivity.

The limiting frequencies which can be perceived are due entirely to the dynamical constants of the inner ear as is also the dependence of minimum audible intensity on frequency.

The so-called subjective harmonics, summation and difference tones are probably due to the non-linear transmission characteristics of the middle and inner ear.

These subjective harmonics account for the greater masking effect of low tones on high tones than high tones on low tones. Due to this non-linear characteristic, the quality as well as the intensity of the sensation produced, especially by complex tones, change as the intensity of the stimulus increases.

The facts obtained from audiograms of abnormal hearing are consistent with the theory of hearing which has been outlined.

Although this theory of hearing involves the principle of resonance, it is very different from the Helmholtz theory as usually understood. In the latter it as assumed that there are four or five thousand small resonators in the ear, each responding only to a single tone; while in the former it is assumed that a single vibrating membrane which vibrates for every impressed sound is sufficient to differentiate the various recognizable sounds by its various configurations of vibration form.

A loudness scale has been chosen such that the loudness change is equal to ten times the common logarithm of the intensity ratio. A pitch scale has been chosen such that the pitch change is equal to 100 times the logarithm to the base two of the frequency ratio. The loudness of complex or simple tones is measured in terms of the number of loudness units a tone of 700 cycles must be raised above its average threshold value before it sounds equally loud to the sound measured.

The degree of deafness is measured by the fractional part of the normal area of audition in which the sensation is either lacking or false.

Appendix A

The calculations of the number of pure tones perceivable as being different in pitch at a given intensity or being different in loudness at a given pitch involves a line integral. The calculation of the number of pure tones perceivable as being different either in loudness or pitch involves a surface integral.

Let the coordinates used in Fig. 1 corresponding to ΔL and ΔP be designated α and β, respectively. Then the relations shown in Figs. 2 and 3 can be expressed by the equations

$$\frac{\Delta E}{E} = f(\alpha - \alpha_0) \text{ and} \tag{1}$$

$$\frac{\Delta N}{N} = \varphi(\beta) \tag{2}$$

where α_0 is the value of α along the normal minimum audibility curve shown in Fig. 1. Knudsen's data indicated that the curve shown in Fig. 3 held only for values of $\alpha - \alpha_0$ corresponding to the flat part of the curve in Fig. 2. For lower intensities the pitch discrimination fell off in about the same way as that shown for the intensity discrimination. To represent this mathematically, $\varphi(\beta)$ can be multiplied by a factor which is unity for the loud tones and which increases similarly to $f(\alpha - \alpha_0)$ for the weaker tones. Such a factor is $10\, f(\alpha - \alpha_0)$ since $f(\alpha - \alpha_0)$ is approximately $\frac{1}{10}$ for the louder tones. So the corrected formula for $\frac{\Delta N}{N}$ is

$$\frac{\Delta N}{N} = 10\varphi(\beta) \cdot f(\alpha - \alpha_0). \tag{3}$$

Let dx be the number of perceivable tones of constant intensity corresponding to α in the pitch region between β and $\beta + d\beta$ and let dy be the number of perceivable tones of constant pitch corresponding to β in the region between α and $\alpha + d\alpha$. Then

$$dx = \frac{dN}{\Delta N} \tag{4}$$

$$dy = \frac{dE}{\Delta E}. \tag{5}$$

But the values of β and α are given by

$$\beta = 100 \log_2 N \tag{6}$$

$$\alpha = 10(\log_{10} E - \log_{10} E_1) \tag{7}$$

where E_1 is the value of intensity corresponding to a pressure amplitude of 1 dyne.

Substituting values of dN and dE in terms of α and β we have

$$dx = \frac{1}{100}\frac{N}{\Delta N}\log_e 2\, d\beta = \frac{\log_e 2}{1000}\frac{d\beta}{\varphi(\beta)\cdot f(\alpha-\alpha_0)} \tag{4'}$$

$$dy = \frac{1}{10}\frac{E}{\Delta E}\log_e 10\, d\alpha = \frac{\log_e 10}{10}\frac{d\alpha}{f(\alpha-\alpha_0)}. \tag{5'}$$

The number of tones of constant intensity which are perceivable as different in pitch is then

$$x = \frac{\log_e 2}{1000}\int_{\beta_1}^{\beta_2}\frac{d\beta}{\varphi(\beta)\cdot f(\alpha-\alpha_0)}$$

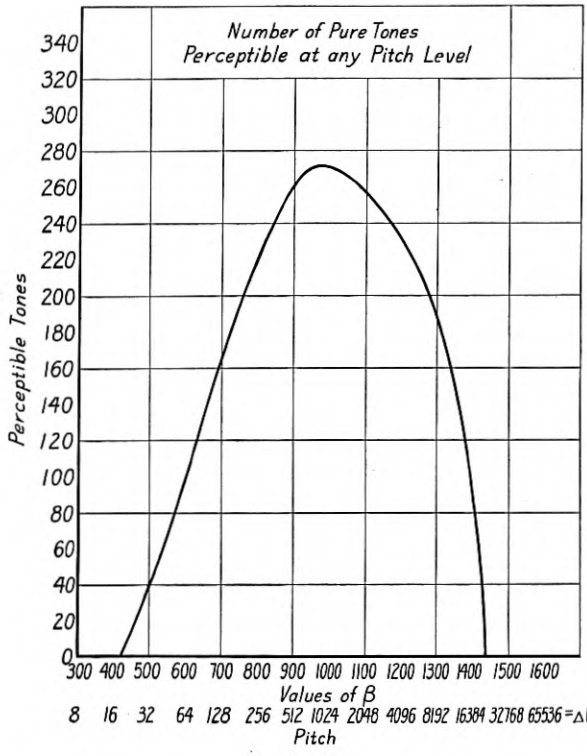


Fig. 13

where β_1 and β_2 are the points where the particular intensity line cuts the boundary lines of the auditory-sensation area. For example, the limits for the line corresponding to 1-dyne pressure ampli-

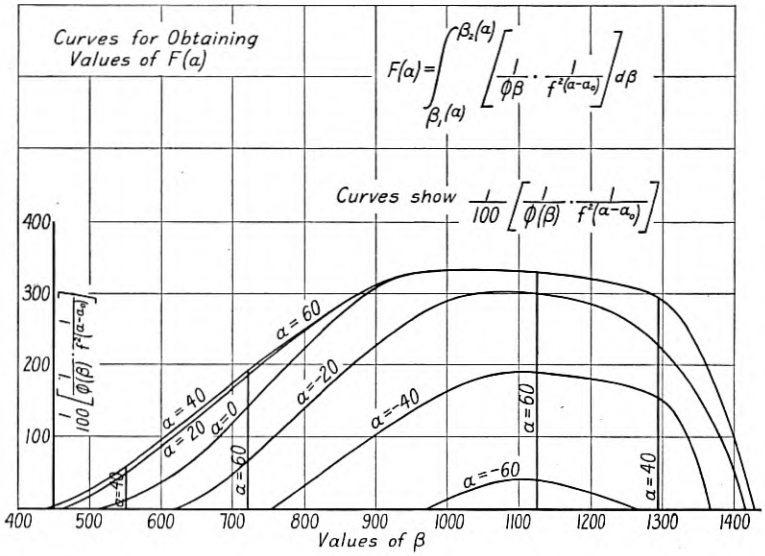


Fig. 14

tude are 500 and 1420. Similarly the number of tones of constant pitch which are perceivable as being different in loudness is given by

$$y = \frac{\log_2 10}{10} \int_{\alpha_1}^{\alpha_2} \frac{d\alpha}{f(\alpha - \alpha_1)}$$

where α_1 and α_2 are determined by the intersection of the particular pitch line with the boundary lines of the auditory-sensation area.

The values of these integrals were computed graphically. Figs. 12 and 13 show the results of these calculations. It is seen that the maximum number of tones perceivable as different in loudness is in the frequency range 700 to 1,500 which is also the important speech range. The number in this range is approximately 270.

In the pressure range from 1 to 100 there are approxmiately 1,500 tones which can be perceived as being different in pitch.

The number of tones ΔT in a small area $d\beta\ d\alpha$ situated with one corner at the point (α, β) is given by $dx\ dy$ or

$$\Delta T = dx\ dy = \frac{\log_e 2 \log_e 10}{10{,}000} \frac{d\alpha\ d\beta}{\varphi(\beta) f^2(\alpha - \alpha_0)},$$

$$T = \frac{\log_e 2 \log_e 10}{10{,}000} \int\int \frac{d\beta\ d\alpha}{\varphi(\beta) f^2(\alpha - \alpha_0)}.$$

The function $\frac{1}{\varphi(\beta)\cdot f^2(\alpha-\alpha_0)}$ must be integrated throughout the auditory-sensation area. This was done by graphical methods as shown in Figs. 14 and 15 with the result that $T=324{,}000$.

Appendix B

Let the pressure variation of the air in front of the drum of the ear be designated by δp. Since the pressure of the air in the middle

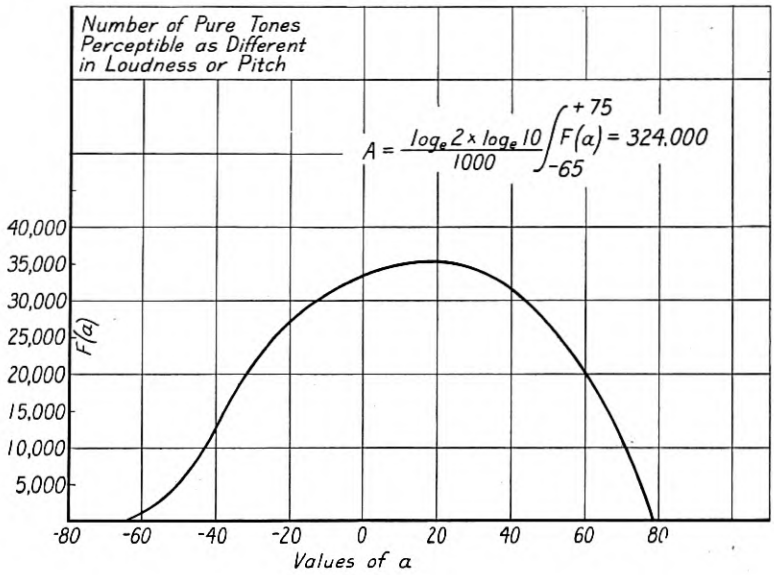


Fig. 15

ear balances the undisturbed outside air pressure this change in pressure multiplied by the effective area of the ear-drum is the only effective force that produces displacements. Let the displacement of the fluid of the cochlea near the oval window be designated by X. If Hookes law held for all the elastic members taking part in the transmission of sound to the inner ear then

$$X=k\delta p \tag{1}$$

where k is a constant.

It would be expected from the anatomy of the ear that Hookes law would start to break down even for small displacements. So in general the relation between the force δp and the displacement X can be represented by

$$X=f(\delta p)=a_0+a_1\delta p+a_2(\delta p)^2+a_3(\delta p)^3+\ldots. \tag{2}$$

where the coefficients α_0, α_1, α_2 . . . belong to the expansion of the function into a power series. Now if δp is a sinusoidal variation then

$$\delta p = p_0 \cos \omega t \tag{3}$$

where $\frac{\omega}{2\pi}$ is the frequency of vibration. Substituting this value in (2), terms containing the cosine raised to integral powers are obtained. These can be expanded into multiple angle functions. For example, for the first four powers

$$\cos^2 \omega t = \tfrac{1}{2} \cos 2\,\omega t + \tfrac{1}{2}, \tag{4}$$

$$\cos^3 \omega t = \tfrac{1}{4} \cos 3\,\omega t + \tfrac{3}{4} \cos \omega t, \tag{5}$$

$$\cos^4 \omega t = \tfrac{1}{8} \cos 4\,\omega t + \tfrac{1}{2} \cos 2\,\omega t + \tfrac{3}{8}. \tag{6}$$

It is evident then that the displacement X will be represented by a formula

$$X = b_0 + b_1 \cos \omega t + b_2 \cos 2\,\omega t + b_3 \cos 3\,\omega t + \ldots .$$

In other words when a periodic force of only one frequency is impressed upon the ear-drum this same frequency and in addition all its harmonic frequencies are impressed upon the fluid of the inner ear.

If two pure tones are impressed upon the ear then δp is given by

$$\delta p = p_1 \cos \omega_1 t + p_2 \cos \omega_2 t.$$

If this value is substituted in equation (2), terms of the form $\cos^n \omega_1 t$ and $\cos^m \omega_2 t$ and $\cos^n \omega_1 t \cos^m \omega_2 t$ are obtained. The first two forms give rise to all the harmonics and the third form gives rise to the summation and difference tones. For example, the first four terms are

$$a_0 = a_0$$

$$a_1 \delta p = a_1 (p_1 \cos \omega_1 t + p_2 \cos \omega_2 t)$$

$$a_2 (\delta p)^2 = a_2 [\tfrac{1}{2} p_1^2 \cos 2\,\omega_1 t + \tfrac{1}{2} p_2^2 \cos 2\,\omega_2 t + p_1 p_2 \{\cos (\omega_1 - \omega_2) t + \cos (\omega_1 + \omega_2) t\} + \tfrac{1}{2}(p_1^2 + p_2^2)]$$

$$a_3 (\delta p)^3 = a_3 \Big[\Big(\frac{3}{4} p_1^3 + \frac{3}{2} p_1 p_2^2\Big) \cos \omega_1 t + \frac{1}{4} p_1^3 \cos 3\,\omega_1 t + \Big(\frac{3}{4} p_2^3 + \frac{3}{2} p_1^2 p_2\Big) \cos \omega_2 t + \frac{1}{4} p_2^3 \cos 3\,\omega_2 t + \frac{3}{4} p_1^2 p_2 \cos (\omega_2 t + 2\,\omega_1 t) + \frac{3}{4} p_1^2 p_2 \cos (\omega_2 t - 2\,\omega_1 t) + \frac{3}{4} p_1 p_2^2 \cos (\omega_1 t + 2\,\omega_2 t) + \frac{3}{4} p_1 p_2^2 \cos (\omega_1 t - 2\omega_2 t) \Big].$$

Therefore unless there is a linear relation between a force acting on the ear-drum and the displacement at the oval window, that is unless all the coefficients in equation (2) are zero except a_1, all the harmonics and the summation and difference tones will be impressed upon the fluid in the cochlea of the inner ear.

BIBLIOGRAPHY

Pitch Discrimination

[1] Preyer, "Grenzen d. Tonwahr.," Jena, 1876.
[2] Luft, *Phil. Studien.*, 4, p. 511, 1888.
[3] Meyer, *Ztschr. f. Psychol. u. Physiol.*, 16, p. 352, 1898.
[4] Schaefer and Guttman, *Ztschr. f. Psychol. u. Physiol.*, 32, p. 87, 1903.
[5] Stucher, *Ztschr. f. Psychol. u. Physiol.*, 42, part 2, p. 392, 1907.
[6] Seashore, *Psychol. Monogr.*, 13, p. 21, 1910.
[7] Vance, *Psychol. Monogr.*, 16, No. 3, p. 115, 1914.
[8] Smith, *Psychol. Monogr.*, 16, No. 3, p. 65, 1914.
[9] Knudsen, *Phys. Rev.*, 21, No. 1, p. 84, Jan., 1923.

Intensity Discrimination

[10] Merkel, *Phil. Studien.*, 4, pp. 117–251, 1887.
[11] *Wien. Ann. d. Phys.*, 36, p. 834, 1889.
[12] Zwaardemaker, *Proc. Konink. Akad. Wetensch, Amsterdam*, 8, p. 421, 1905.
[13] Pillsbury, *Psychol. Monogr.*, 13, No. 1, p. 5, 1910.
[14] Knudsen, *Phys. Rev.*, 21, No. 1, p. 84, Jan., 1923.

Absolute Sensitivity

[15] Toepler and Boltzman, *Ann. d. Phys.*, 141, p. 321, 1870.
[16] Wead, *Am. Jour. Sci.*, (3), 26, p. 177, 1883.
[17] Rayleigh, *Proc. Roy. Soc.*, 26, p. 248, 1887; *Phil. Mag.*, [5], 38, pp. 295, 365, 1894; *Phil. Mag.*, [6], 14, p. 596, 1907.
[18] *Wien. Archiv. fur die gesamte Physiologie*, 97, pp. 1–57, 1903.
[19] Zwaardemaker and Quix, *Arch. f. Anat. u. Physiol. Abt.*, p. 321, 1903.
[20] Webster, *Festschrif. f. L. Boltzmann*, Leipzig, p. 866, 1904.
[21] Shaw, *Proc. Roy. Soc.*, Ser. A, 76, 360, 1905.
[22] Abraham, *Compt. Rend.*, 144, p. 1099, 1907.
[23] Koehler, *Ztschr. f. Psychol. u. Physiol.*, 54, (1), p. 241, 1910.
[24] Bernbaum, *Ann. d. Phys.*, 49, pp. 201–228, 1916.
[25] Kranz, *Phys. Rev.*, 17, No. 3, p. 384, 1921.
[26] Minton, *Phys. Rev.*, 19, No. 2, p. 80, 1922.
[27] Hewlett, *Phys. Rev.*, p. 52, Jan., 1922.
[28] Fletcher and Wegel, *Phys. Rev.*, 19, No. 6, p. 553, 1922.
[29] Lane, *Phys. Rev.*, 19, No. 5, May, 1922.
[30] Wegel, *Proc. Nat'l Acad.*, July 15, 1922.
[31] MacKenzie, *Phys. Rev.*, 20, No. 4, Oct., 1922.

Upper Limit of Audibility

[32] Savart, *Compt. Rend.*, 20, pp. 12–14, 1845.
[33] Rayleigh, *Proc. Roy. Inst. of Gr. Br.*, 15, p. 417, 1897.
[34] Koenig, *Handb. d. Physiol.*, 3, p. 112, 1880; *Ann. d. Phys.*, 69, pp. 626–721, 1899.

[35] Scripture and Smith, *Stud. from Yale Psychol. Lab.*, 2, p. 105, 1894.
[36] Stumpf and Meyer, *Ann. d. Phys.*, 61, p. 773, 1897.
[37] Schwendt, *Arch. f. d. ges. Physiol.*, 75, p. 346, 1899.
[38] Edelmann, *Ann. d. Phys.*, 2, p. 469, 1900.
[39] Meyer, *Jour. Physiol.*, 28, p. 417, 1902.
[40] *Wien. Arch. f. ges. Physiol.*, 97, p. 1, 1903.
[41] Bezold, *Funk. Pruef. d. Mensch. Gehoer.*, 2, p. 162, 1903.
[42] Schulze, *Ann. d. Phys.*, 24, p. 785, 1907.
[43] Stueker, *Sitz. Ber. d. Akad. d. Wiss. Wien.*, 116, 2a, p. 367, 1907.
[44] Wegel, *Proc. Nat'l Acad.*, July 15, 1922.

Lower Limit of Audibility

[45] Savart, *Ann. de Phys. et de Chem.*, 47.
[46] Helmholtz, "Sens. of Tone," Eng. Trans., p. 175, 1885.
[47] Preyer, "Ueber d. Grenze d. Tonwahr," Jena, p. 8, 1876.
[48] Benzold, *Ztschr. f. Psychol. u. Physiol.*, 13, p. 161, 1897.
[49] Schaefer, *Ztschr. f. Psychol. u. Physiol.*, 21, p. 161, 1899.
[50] Imai and Vance, *Psychol. Monogr.*, 16, p. 104, 1914.
[51] Wegel, *Proc. Nat'l Acad.*, July 15, 1922.

Miscellaneous References

[52] Mayer, A. M., *Phil. Mag.*, 11, p. 500, 1875, "Researches in Acoustics."
[53] Peterson, *Jos. Psychol. Rev.*, 23, No. 5, p. 333, 1916, "The Place of Stimulation in the Cochlea *vs.* Frequency as a Direct Determiner of Pitch."
[54] Arnold and Crandall, *Phys. Rev.*, 10, No. 1, July, 1917, "The Thermophone as a Precision Source of Sound."
[55] Wente, *Phys. Rev.*, 19, No. 5, p. 498, 1922, "The Sensitivity and Precision of the Electrostatic Transmitter for Measuring Sound Intensities."
[56] Fletcher and Wegel, *Phys. Rev.*, 19, No. 6, June, 1922, "The Frequency Sensitivity of Normal Ears."
[57] Fowler and Wegel, *Annals of the Amer. Rhin., Larg. and Ot. Soc.*, June, 1923, "Audiometric Methods and their Applications."
[58] Fletcher, "Nature of Speech and its Interpretation," *Jour. Frank. Inst.*, June, 1922.
[59] Lillie, R. S., "The Relation of Stimulation and Conduction in Irritable Tissues to Changes in Permeability of the Limiting Membranes," *Amer. J. Physiol.*, 28, 197–223, 1911. Also other papers.
[60] Lucas, K., "The Conduction of the Nervous Impulse," London, 1919.

Discussions of the Theory of Hearing

[61] Roaf, *Phil. Mag.*, 43, pp. 349–354, Feb., 1922.
[62] Wrightson, "Analytical Mechanism of Internal Ear," 1918.
[63] Morton, *Phys. Soc. Lond.*, 31, p. 101, Apr., 1919.
[64] Peterson, *Psych. Rev.*, 20, p. 312, 1913.
[65] Boring and Titchener, *Physiolog. Abs.*, 6, p. 27, April, 1921.
[66] Meyer, *Amer. Jour. of Psych.*, 18, pp. 170–6, 1907.
[67] Hartridge, *Brit. Jour. Psychol.*, 12, pp. 248–52, 1921.
[68] Hartridge, *Brit. Jour. Psychol.*, 12, pp. 277–88, 1921.
[69] Gray, *Jour. Anat. Phys.*, 34, p. 324, 1900.
[70] Hartridge, *Brit. Jour. Psychol.*, 12, pp. 142–6, 1921.
[71] Hartridge, *Brit. Jour. Psychol.*, 12, pp. 248–52, 1921.
[72] Hartridge, *Nature*, 107, pp. 394–5, May 26, 1921; 107, p. 204, April 14, 1921.
[73] Hartridge, *Brit. Jour. Psychol.*, 12, pp. 362–82, Aug. 16, 1921.

[74] Hartridge, *Nature*, 107, p. 811, Aug. 25, 1921.
[75] Hartridge, *Nature*, 109, p. 649, May 20, 1920.
[76] Ackerman, *Nature*, 109, p. 649, May 20, 1920.
[77] Hartridge, *Nature*, 110, pp. 9–10, July 1, 1922.
[78] Wilkinson, *Proc. of Roy. Soc. Med.*, 15, Sect. of Otol., pp. 51–3, 1922.
[79] Bayliss, *Nature*, 110, p. 632, Nov. 11, 1922.
[80] Broemser, *Physiol. Abs.*, 6, p. 28, April, 1921. (Abs. from *Sitzung. d. Ges. f. Morphol. u. Physiol. in Munchen*, p. 67, 1920.)
[81] Weiss, *Psychol. Rev.*, 25, p. 50, 1918.
[82] Abraham, *Ann. d. Phys.*, 60, No. 17, p. 55, 1919.
[83] Buck, *N. Y. Med. Jour.*, June, 1874.
[84] Bryant, Repr. Trans. of the *Amer. Otol. Soc. Transact.*, 1909.
[85] Marage, *Comp. Rend.*, 175, p. 724, Oct. 23, 1922.
[86] Rayleigh, *Sci. Abs.*, Sec. A., 22, p. 124, Mar. 31, 1919. (Abs. from *Nature*, 102, p. 304, Dec. 19, 1918.)
[87] Marage, *Comp. Rend.*, 172, p. 178, Jan. 17, 1921.
[88] Dahns, *Monasschr. f. Ohrenheilk. u. Laryng.-Rhinol.*, 56, p. 23, 1922.
[89] Barton, *Nature*, 110, pp. 316–9, Sept. 2, 1922.
[90] Scripture, *Nature*, 109, p. 518, Apr. 22, 1922.
[91] Ogden, *Psychol. Bull.*, 15, p. 76, 1918.
[92] Ogden, *Psychol. Bull.*, 16, p. 142, 1919.
[93] Ogden, *Psychol. Bull.*, 17, p. 228, 1920.

The Contributors to this Issue

GEORGE A. CAMPBELL, B.S., Massachusetts Institute of Technology, 1891; A.B., Harvard, 1892; Ph.D., 1901; Göttingen, Vienna and Paris, 1893–96. Mechanical Department, American Bell Telephone Company, 1897; Engineering Department, American Telephone and Telegraph Company, 1903–1919; Department of Development and Research, 1919—; Research Engineer, 1908—. Dr. Campbell has published papers on loading and the theory of electric circuits and is also well-known to telephone engineers for his contributions to repeater and substation circuits. The electric filter which is one of his inventions plays a fundamental röle in telephone repeater, carrier current and radio systems.

ROBERT W. KING, A.B., Cornell University, 1912; Ph.D., 1915; assistant and instructor in physics, Cornell, 1913–17; Engineering Department of the Western Electric Company, 1917–20; Department of Development and Research, American Telephone and Telegraph Company, 1920–21; Information Department, 1921—. While with the Western Electric Company, Mr. King's work related to the design and construction of vacuum tubes and allied high vacuum apparatus.

KARL K. DARROW, S.B., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917—. At the Western Electric, Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

H. D. ARNOLD, Ph.B., Wesleyan, 1906; M.S., 1907; Ph.D., Chicago, 1911; assistant in physics, Wesleyan, 1906–07; Chicago, 1908; professor, Mt. Allison, 1909–10; Engineering Department of the Western Electric Company, Research Engineer, 1911—; Director of Research, 1923—. Dr. Arnold has been in direct charge of the development of the vacuum tube for telephone repeaters and radio purposes, and also other items of telephone equipment.

LLOYD ESPENCHIED, Pratt Institute, 1909; United Wireless Telegraph Company as radio operator, summers, 1907–08; Telefunken Wireless Telegraph Company of America, assistant engineer, 1909–10;

American Telephone and Telegraph Company, Engineering Department and Department of Development and Research, 1910—. Took part in long distance radio telephone experiments from Washington to Hawaii and Paris, 1915; since then his work has been connected with the development of radio and carrier systems.

HARVEY FLETCHER, B.S., Brigham Young, 1907; Ph.D., Chicago, 1911; instructor of physics, Brigham Young, 1907–08; Chicago, 1909–10; Professor, Brigham Young, 1911–16; Engineering Department, Western Electric Company, 1916—. The present paper by Dr. Fletcher gives some of the results of an investigation which is being made of the relation between the frequency characteristics of telephone circuits and the intelligibility of transmitted speech. Dr. Fletcher has also published on Brownian movements, ionization and electronics.