The Bell System Technical Journal

Vol. XVI *October, 1937* *No. 4*

Resistance Compensated Band-Pass Crystal Filters for Use in Unbalanced Circuits

By W. P. MASON

In this paper are discussed several types of crystal band-pass filters which can be used in unbalanced circuits. These types of filters are all resistance compensated, i.e., the resistances associated with the filter elements are in such a position in the filter that they can be effectively brought to the ends of the filter and combined with the terminal resistances with the result that the dissipation produces an additive loss for the filter characteristic and does not affect the sharpness of cut-off attainable. It is shown that all these types of networks can be reduced to three lattice types of crystal filters, and the formulae for these three networks are given. A comparison is given between the characteristics obtainable with resistance compensated crystal and electrical filters and a conclusion regarding their comparison given by V. D. Landon [4] is shown to be incomplete.

I. Introduction

IN a recent paper [1] a description is given of a number of wave filters employing quartz crystals as elements. Most of these filters were of the lattice type and hence were inherently balanced. For some purposes, however, such as connecting together unbalanced tubes, it is desirable to obtain a filter in an unbalanced form and it is the purpose of this paper to show several forms for constructing resistance compensated band-pass crystal filters which will give results similar to those described previously. Another purpose is to give a numerical comparison between the characteristics obtainable with resistance compensated crystal and electrical filters.

II. A Comparison of the Performance Characteristics of Crystal vs. Coil and Condenser Filters

In order to show the properties of resistance compensated crystal filters it is instructive to give a comparison between the types of characteristics which can be obtained by using crystal and coil and

[1] "Electrical Wave Filters Employing Quartz Crystals as Elements," W. P. Mason, *B. S. T. J.*, July, 1934, p. 405.

condenser filters. The quartz crystal filter considered here is shown on Fig. 1.

By using the balancing resistance R_x of Fig. 1 the crystal filter can be made entirely compensated for coil resistance; i.e. the resistance associated with the coils of the network is in such a place in the network that it can be effectively brought to the ends of the filter and combined with the terminal impedances with the result that the effect of the dissipation in the coils is only to produce an additive loss for the filter characteristic and does not affect the sharpness of cut-off attainable. In fact if the filter works into a vacuum tube the dissipation in the coil can be used to terminate the filter completely, and introduces no loss.

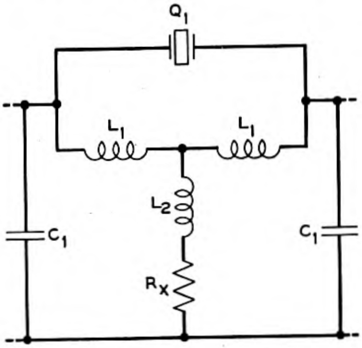


Fig. 1—A bridge T quartz crystal filter.

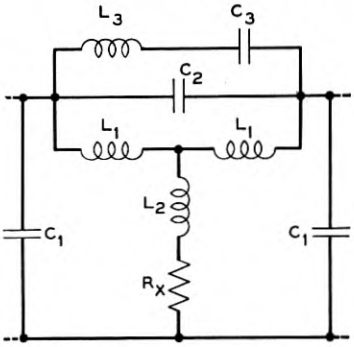


Fig. 2—Electrical network equivalent to crystal filter of Fig. 1.

For the electrical filter, however, the dissipation introduced by the electrical elements which replace the crystal is not compensated and causes a considerable distortion of the pass band which becomes more prominent as the band width is narrowed. To show this let us consider the network of Fig. 2. In analyzing such networks it is usually more convenient to reduce them to their equivalent lattice form and apply network equivalences holding for lattice type networks. This can be done by applying Bartlett's Theorem [2] which states that any network which can be divided into two mirror image halves can be reduced to an equivalent lattice network by placing in the series arms of the lattice a two-terminal impedance formed by connecting the two input terminals of one half of the network in this arm and short-circuiting all of the cut wires of the network, and in the lattice arm placing the same network with all its cut wires open-circuited. Applying this process to Fig. 1, a lattice network equivalent to the network of Fig. 1 is that shown on Fig. 3. In this network the capacitances can be considered as substantially dissipationless and if the network representing the crystal can also be considered dissipationless, the resistance introduced by the coils can be effectively brought outside the lattice and incorporated with the terminal resistances. This follows from the fact that an inductance with an associated series resistance can just as well be represented over the narrow-frequency range of the filter by an inductance paralleled by a much higher resistance. The impedance of an inductance and resistance in series and the impedance of an inductance and resistance in parallel are given by the expressions

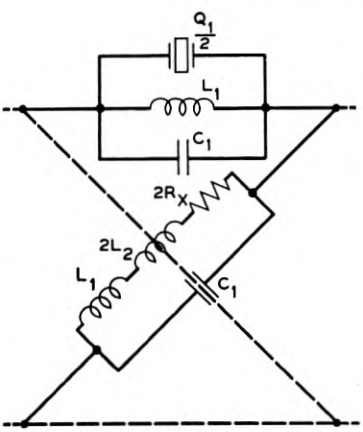


Fig. 3—Lattice equivalent of crystal filter of Fig. 1.

$$R_1 + j\omega L_1 = \frac{R_2(j\omega L_2)}{R_2 + j\omega L_2} = \frac{R_2\omega^2 L_2^2 + j\omega L_2 R_2^2}{R_2^2 + \omega^2 L_2^2}. \qquad (1)$$

[2] "Extension of a Property of Artificial Lines," A. C. Bartlett. *Phil. Mag.*, **4**, pp. 902–907, Nov. 1927.

Defining Q, the ratio of reactance to resistance, as $Q = \omega L_1/R_1$, we have

$$R_2 = R_1(1 + Q^2); \quad L_2 = L_1(1 + 1/Q^2). \tag{2}$$

This relation holds strictly only for a single frequency, but over a narrow-band filter the relation holds quite accurately.

Employing this conception, the lattice network can be reduced to that of Fig. 4 in which a resistance R parallels each arm of the lattice.

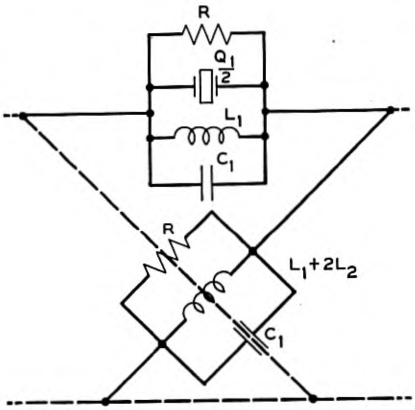


Fig. 4—Filter showing paralleling resistance.

This is made possible by the adjustable resistance R_x which is fixed at such a value that the parallel resistance associated with the inductance $L_1 + 2L_2$ is equal to that associated with L_1. Then by employing the two lattice equivalents shown on Fig. 5, first proved by the writer,[3] it is

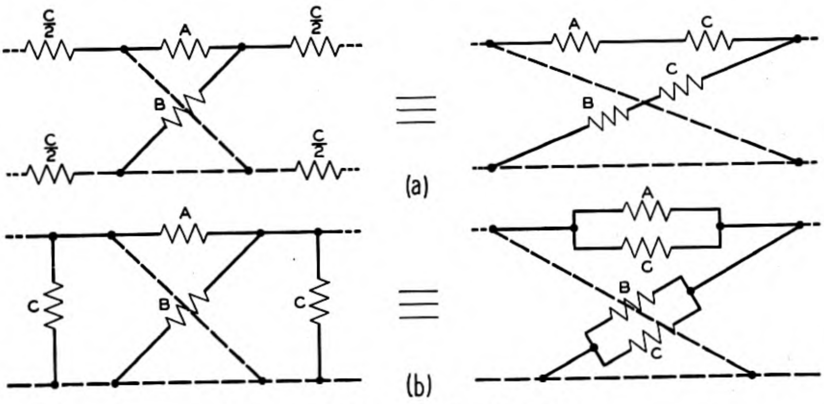


Fig. 5—Lattice network equivalences.

[3] Reference 1, page 418.

possible to take these resistances outside the lattice and combine them with the terminating impedance, leaving all the elements inside the lattice dissipationless. The two remaining arms of the lattice have the impedance characteristic shown on Fig. 6*A*. A lattice filter has a pass band when the two impedance arms have opposite signs and an attenuation band when they have the same sign. When the impedance of two arms cross, an infinite attenuation exists. Hence the characteristic obtainable with this network is that shown on Fig. 6*B*.

Next let us consider an electrical filter in which coils and condensers take the place of the essentially dissipationless crystal. In this case the dissipation due to L_1 and L_2 can be balanced as before and the only question to consider is the effect of the dissipation associated with L_3 and C_3. In a similar manner to that employed for the coil we can show that a series tuned circuit with a series resistance R_4 is equivalent to a second series tuned circuit having the same resonant frequency as the first shunted by a resistance R_4' where

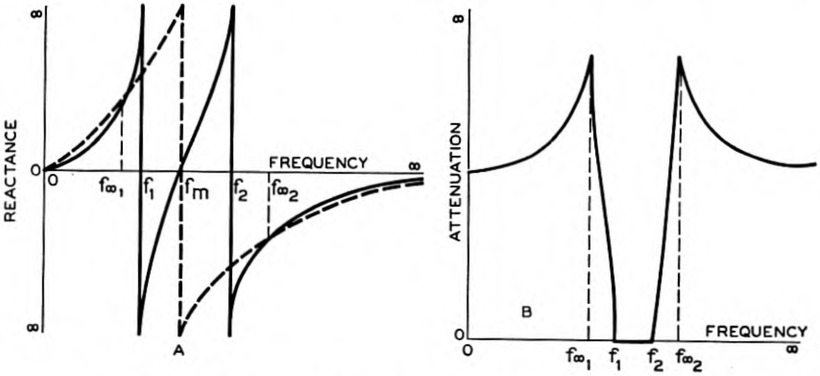


Fig. 6—Characteristics obtainable with the crystal filter of Fig. 1.

$$R_4' = R_4(1 + Q^2); \quad C_4' = C_4(1 + 1/Q^2)$$

$$\text{where } Q = \left| \frac{\frac{1}{\omega C_4}\left(1 - \frac{\omega^2}{\omega_R{}^2}\right)}{R_4} \right|. \qquad (3)$$

At two frequencies for which the absolute values of the reactances are the same and therefore the value of Q equal, it is possible to replace the series resistance by a shunt resistance and hence compensate it by varying the resistance R_x. Since, however, the reactance of the tuned circuit varies from a negative value through zero to a positive value over the pass band of the filter, the value of this shunt resistance is not even approximately constant and hence the filter cannot be resistance compensated throughout the band of the filter. It can, however, be compensated at the frequencies of infinite attenuation and high losses can be obtained at these frequencies.

The effect of the lack of resistance compensation throughout the band can best be shown by a numerical computation of the loss of an electrical filter as compared to that for a crystal filter. A practical example has been taken of a filter whose band width is 12 kilocycles wide with the mean frequency at 465 kilocycles. In order to obtain the best Q's with reasonably sized coils an arrangement suggested by R. A. Sykes is used. The coils L_1 are obtained by using the two equal windings of a coupled coil series aiding so that L_1 equals the primary inductance plus the mutual inductance. Since in an air core coil all of the dissipation is associated with the primary inductance and none with the mutual this gives a high Q for L_1. The inductance L_2 neutralizes the negative mutual inductance $-M$ and supplies in addition a small positive inductance. The Q of this combination is poor but it makes unnecessary the use of a high resistance R_x for balancing purposes. By this method a much higher effective Q is obtained than can be obtained by a single coupled coil or by three separate coils.

The calculated curve for the electrical filter assuming Q's of 150 for all the coils is shown on Fig. 7 by the dotted lines. A similar curve for a crystal filter is shown on Fig. 7 by the full lines. As is evident the effect of the coil dissipation is to round off the edges of the pass band and to limit the effective discrimination between the passed and attenuated bands.

This result does not agree with that given by Landon,[4] who in a recent paper makes a comparison between the results obtained with crystal and electrical filters which appears to be somewhat misleading. It is stated in this paper that the electrical filter circuits given are completely resistance compensated and "in crystal filters in which the crystal is confined to the rejector meshes of the network, the limitation is about the same as for electrical filters." By referring to the curves of Fig. 7 it is readily seen that high losses can be obtained outside the pass band with resistance compensated electrical filters,[5] but that the

[4] "'*M* Derived' Band-Pass Filters with Resistance Cancellation," Vernon D. Landon, *R. C. A. Review*, Oct. 1936, Vol. 1, No. 2, Page 93.

[5] The use of resistance for compensating and balancing the attenuation in electrical filters has been worked out by H. W. Bode and S. Darlington (see U. S. patents 2,002,216, 1,955,788, 2,029,014, 2,035,258). The first work was done for low- and high-pass filters but it was later extended also to band-pass filters. Some of these results are analogous to those of Landon, while others give a better compensation within the transmitted band. The use of the resistance in the crystal filter of Fig. 1 was suggested by Mr. Darlington.

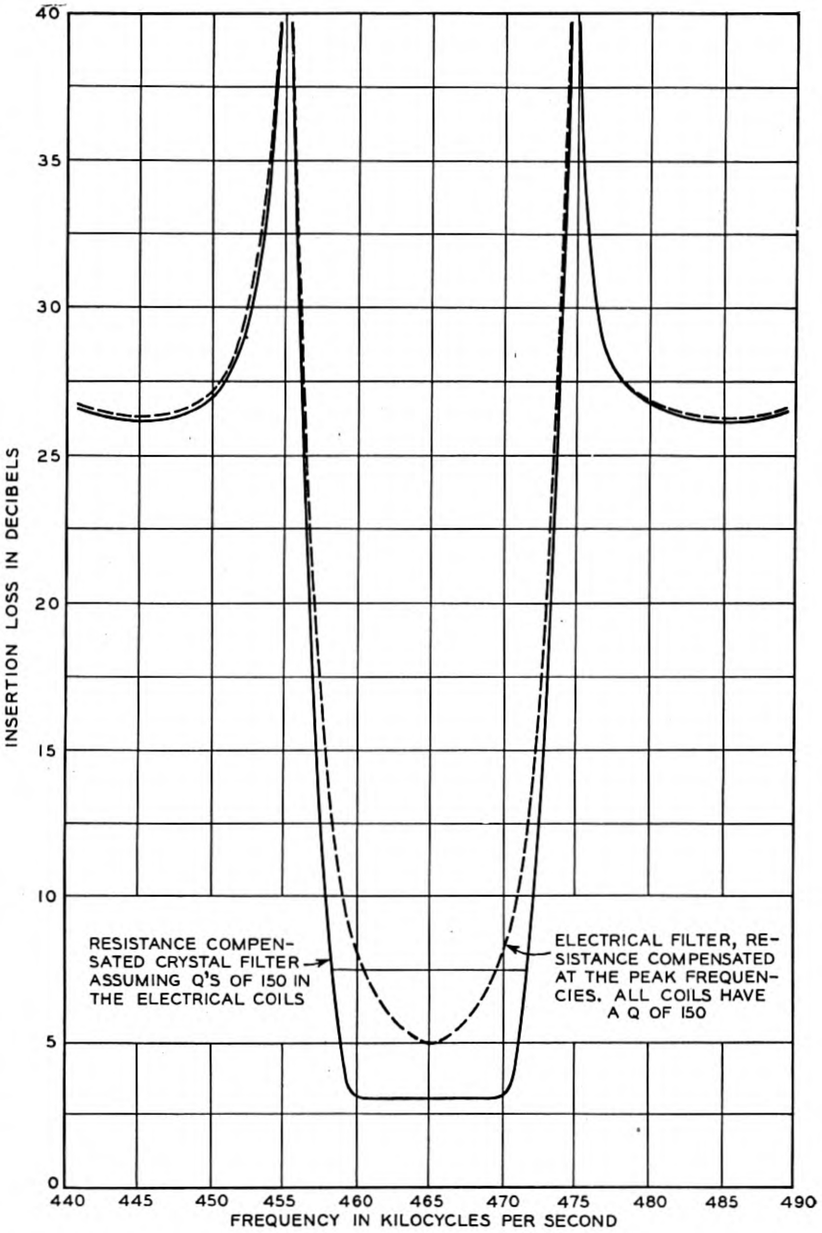


Fig. 7—Numerical comparison between the loss characteristics of a crystal filter and a coil and condenser filter.

pass band of the filter is seriously distorted unless elements, such as crystals, are used which have negligible dissipation.

III. Band-Pass Resistance Compensated Crystal Filters

All of the wide-band resistance compensated crystal filters proposed so far can be shown to be equivalent to the two general types of lattice crystal filters shown on Fig. 8. For example the crystal filter of Fig. 1 was shown to be equivalent to the lattice type filter of Fig. 8 (b) in which the crystals in the lattice arms are left out.

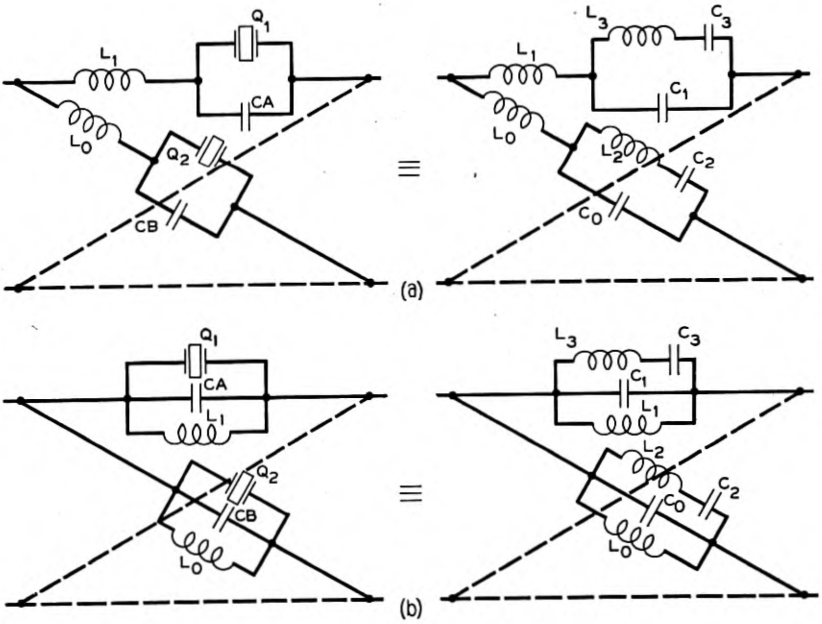


Fig. 8—Wide-band lattice crystal filters.

In the lattice filters of Fig. 8 the number of crystals employed can be cut in half by employing in two similar arms a crystal with two pair of equal plates. It can be shown that such a crystal used in similar arms is equivalent to two identical crystals of twice the impedance of the crystal used as a single plate and having the same resonance frequency. Hence the lattice filters of Fig. 8 are as economical of elements—except for two condensers—as an unbalanced type filter. For some purposes, however, such as connecting together unbalanced tubes, it is desirable to obtain a filter in an unbalanced form. Also, at high frequencies the crystals become quite small and hence it becomes difficult to divide the plating on such crystals. It is the purpose of this section to list a number of filters of the unbalanced type which are equivalent to the lattice filters of Fig. 8. They do not have as general filter character-

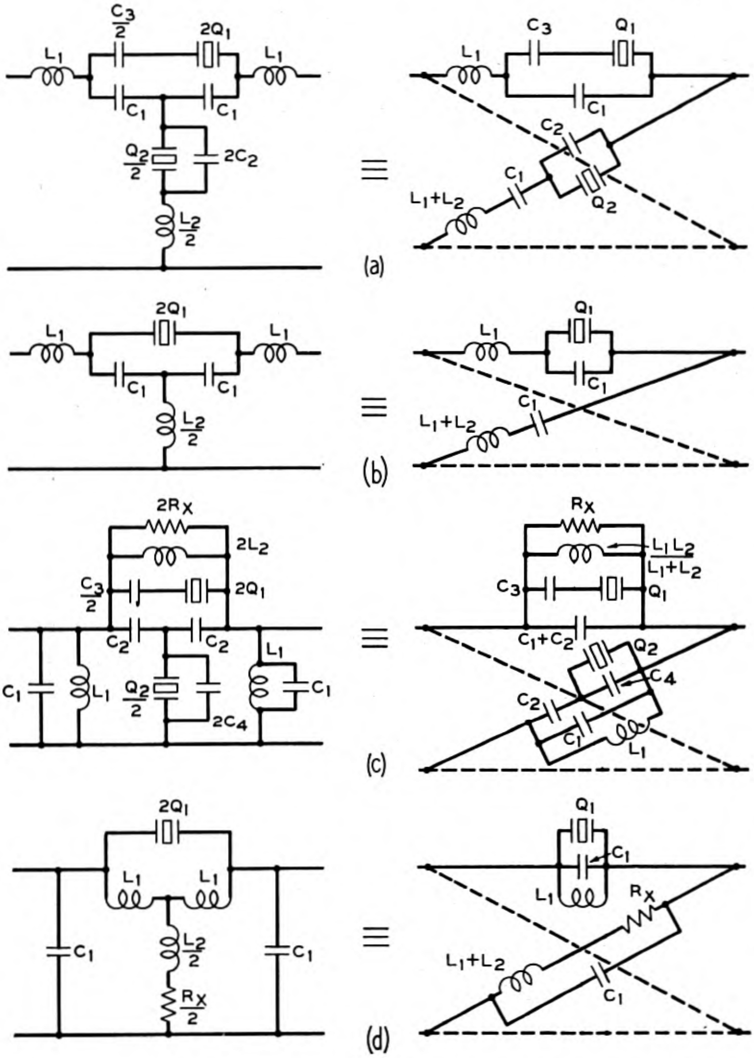


Fig. 9—Wide-band bridge T crystal filters.

istics as the equivalent lattice networks but for a number of purposes are satisfactory.

Fig. 9 shows four bridge *T* crystal filters which are equivalent respectively to the lattice crystal filters of Fig. 8. The equivalent lattice configurations are shown on Fig. 9. The first two filters have series coils which inherently give low-impedance filters. The second of these is equivalent to the filter of Fig. 8 (a) with one pair of the crystals eliminated. If the inductance L_2 were eliminated from Fig. 9 (a) or (b) the filters will be resistance compensated, for all of the resistance will be on the ends of the filter. Furthermore if a small amount of coupling is allowed between the two end coils, the effect of this will be to introduce the small coil $L_2/2$ in the desired place as can be seen from the T network equivalent of a coupled coil as shown on Fig. 10. Furthermore if the coils are air core, no dissipation is associated with the mutual inductance and hence if coupled coils are used the networks still have a resistance balance. Similarly the filters shown on Figs. 9 (c) and (d) are equivalent to the high-impedance type filter shown on Fig. 8 (b) with all crystals present or with crystals missing from the lattice arms. By employing coils with a small amount of mutual inductance these types can also be made with a resistance balance. They can also be made to balance for physical coils by employing the resistances shown. It is obvious from the equivalent lattice structures that these networks have limitations on band widths and allowable attenuation which are not present for the original lattice structures of Fig. 8. However, for filters whose pass bands are less than the maximum pass bands, useful results can be obtained.

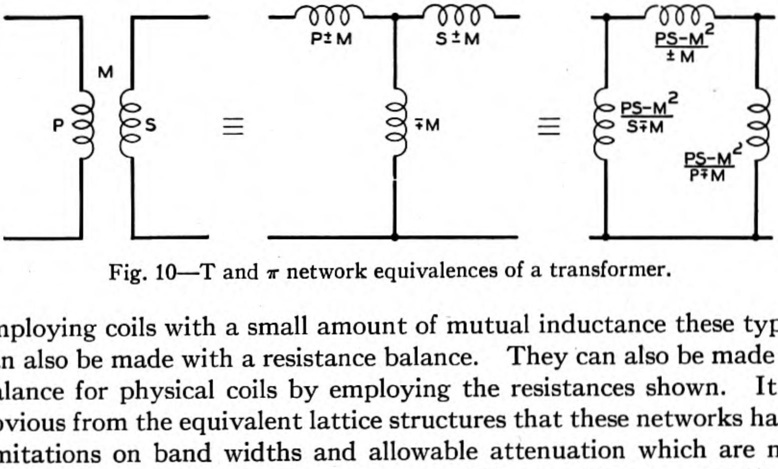


Fig. 10—T and π network equivalences of a transformer.

Another method for obtaining results similar to that obtainable in a lattice network is to use a hybrid coil with series aiding secondaries which are connected to a crystal and a condenser as shown on Fig. 11. This circuit, which has been used extensively to provide a narrow band crystal filter in telegraph work, was invented first by W. A. Marrison [6] of the Bell Telephone Laboratories. Under certain circumstances this configuration can be shown to give results similar to the narrow-band

[6] Patent 1,994,658 filed June 7, 1927, granted March 19, 1935.

lattice filter of Fig. 12. A hybrid coil with series aiding windings connected to two impedances $2Z_1$ and $2Z_2$ as shown by Fig. 13*A* can be shown to be equivalent to the circuit of Fig. 13*B* in which a lattice network with the branches Z_1 and Z_2 is placed in series with the transforming network and the series terminating inductances. Hence if the hybrid coil has nearly a unity coupling between its secondary coils and the remainder of the transformer is designed to work into the impedance of the filter, the network of Fig. 11 is equivalent to the narrow-band lattice filter of Fig. 12 with crystals removed from the lattice arms, plus

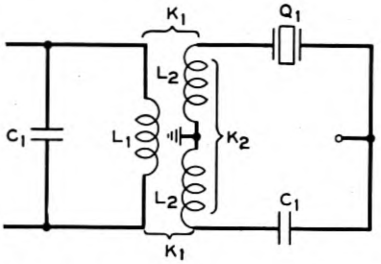


Fig. 11—A three-winding transformer crystal filter.

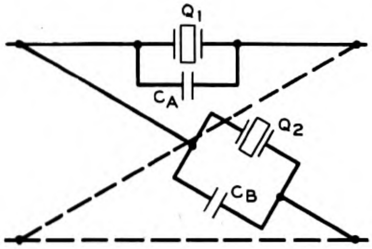


Fig. 12—A narrow-band lattice crystal filter.

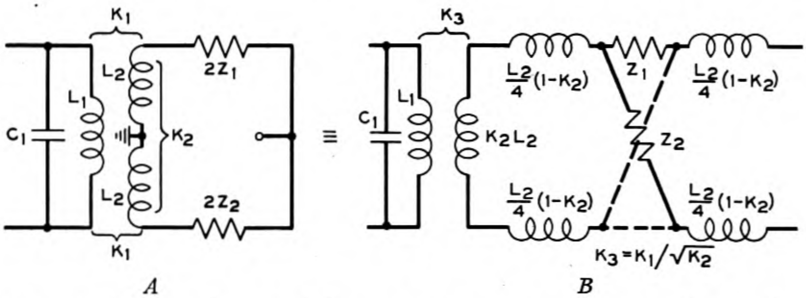


Fig. 13—An equivalent circuit for a three-winding transformer and network.

a transformer. As usually used, however, the impedance of the transformer is much lower than that of the filter and as a consequence the band-pass characteristic of the filter is lost. As a result the network passes only a single frequency and gives results similar to those obtainable with a very sharply tuned circuit. By placing a crystal in the other arm of the network as shown by Fig. 14,[7] this configuration can be made equivalent to the filter shown in Fig. 12.

It is obvious from the equivalence of Fig. 13 that the configurations of Fig. 11 and Fig. 14 can also be used to give a wide-band filter. This follows since the series inductances can be taken inside the lattice and the low-impedance crystal filter of Fig. 8 (a) results. The Q of the coils included in the filter will ordinarily not be high since the inductance is obtained by a difference of primary and mutual inductances, and a better result will be obtained by making the secondary coupling high and including physical coils in series with the crystals.

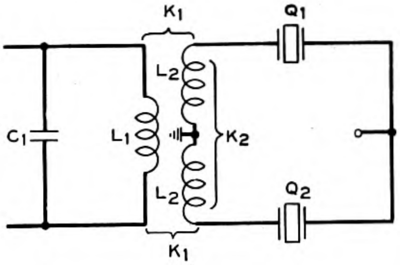


Fig. 14—A three-winding transformer crystal filter with two crystals.

We see then that all of the resistance compensated wide-band filters are equivalent to the lattice filters of Figs. 8 and 12, and all their design equations are known when the design equations of the equivalent lattices are calculated. This requires two steps, first the calculation of the spacing of the resonant frequencies of the network to give the required attenuation and secondly the calculation of the element values from the known resonances by means of Foster's theorem. Such calculations are familiar in filter theory and hence only the results are given here. The results are given in Tables I, II, and III for the network of Figs. 8 (a), 8 (b) and 12 respectively. These values are given in terms of the characteristic impedance Z_0 of the filter at the mean frequency, the lower and upper cut-off frequencies f_1 and f_2 respectively and the b's of the network. These last are parameters which specify

[7] This configuration is covered by patent 2,001,387 issued to C. A. Hansell.

TABLE I

Element	Formula
L_0	$\dfrac{Z_0 f_2(f_2^2 + f_1^2 B)}{2\pi f_1(f_2 - f_1)(f_2^2 A + f_1^2 C)}$
L_1	$\dfrac{Z_0 f_1(f_2^2 A + f_1^2 C)}{2\pi f_2(f_2 - f_1)(f_2^2 + f_1^2 B)}$
L_2	$\dfrac{Z_0[f_2^4(A(1 + B) - C) + 2f_1^2 f_2^2 C + f_1^4 BC]^2}{2\pi f_1 f_2(f_2 - f_1)^3(f_2 + f_1)^2[f_2^2 A + f_1^2 C]C(AB - C)}$
L_3	$\dfrac{Z_0[f_2^4 A + 2f_1^2 f_2^2 C + f_1^4(B(A + C) - C)]^2}{2\pi f_1 f_2(f_2 - f_1)^3(f_2 + f_1)^2(f_2^2 + f_1^2 B)(AB - C)}$
C_0	$\dfrac{(f_2 - f_1)(f_2^2 A + f_1^2 C)^2}{2\pi Z_0 f_1 f_2[f_2^4(A(1 + B) - C) + 2f_1^2 f_2^2 C + f_1^4 BC]}$
C_1	$\dfrac{(f_2 - f_1)(f_2^2 + f_1^2 B)^2}{2\pi Z_0 f_1 f_2[f_2^4 A + 2f_1^2 f_2^2 C + f_1^4(B(A + C) - C)]}$
C_2	$\dfrac{(AB - C)C(f_2 - f_1)^3(f_2 + f_1)^2}{2\pi Z_0 f_1 f_2[f_2^4(A(1 + B) - C) + 2f_1^2 f_2^2 C + f_1^4 BC](1 + B)}$
C_3	$\dfrac{(AB - C)(f_2 - f_1)^3(f_2 + f_1)^2}{2\pi Z_0 f_1 f_2[f_2^4 A + 2f_2^2 f_1^2 C + f_1^4(B(A + C) - C)](A + C)}$

where $A = b_1 + b_2 + b_3;$ $B = b_1 b_2 + b_1 b_3 + b_2 b_3;$ $C = b_1 b_2 b_3;$

$$b_n = \sqrt{\frac{1 - f\infty_n^2/f_1^2}{1 - f\infty_n^2/f_2^2}}; \qquad n = 1, 2, 3$$

TABLE II

Element	Formula	Element	Formula
L_0	$\dfrac{Z_0(f_2 - f_1)(A + C)}{2\pi f_1 f_2(1 + B)}$	C_0	$\dfrac{(f_2^2 + f_1^2 B)f_2}{2\pi Z_0 f_1(f_2 - f_1)(f_2^2 A + f_1^2 C)}$
L_1	$\dfrac{Z_0(f_2 - f_1)(1 + B)}{2\pi f_1 f_2(A + C)}$	C_1	$\dfrac{(f_2^2 A + f_1^2 C)f_1}{2\pi Z_0 f_2(f_2 - f_1)(f_2^2 + f_1^2 B)}$
L_2	$\dfrac{Z_0(A + C)(f_2^2 A + f_1^2 C)^2}{2\pi f_1 f_2(f_2 - f_1)(f_2 + f_1)^2 C(AB - C)}$	C_2	$\dfrac{(f_2 - f_1)(f_2 + f_1)^2 C(AB - C)}{2\pi Z_0 f_1 f_2(f_2^2 A + f_1^2 C)(A + C)^2}$
L_3	$\dfrac{Z_0(1 + B)(f_2^2 + f_1^2 B)^2}{2\pi f_1 f_2(f_2 - f_1)(f_2 + f_1)^2(AB - C)}$	C_3	$\dfrac{(f_2 - f_1)(f_2 + f_1)^2(AB - C)}{2\pi Z_0 f_1 f_2(1 + B)^2(f_2^2 + f_1^2 B)}$

where $A = b_1 + b_2 + b_3;$ $B = b_1 b_2 + b_1 b_3 + b_2 b_3;$ $C = b_1 b_2 b_3;$

$$b_n = \sqrt{\frac{1 - f\infty_n^2/f_1^2}{1 - f\infty_n^2/f_2^2}}; \qquad n = 1, 2, 3$$

TABLE III

Element	Formula	Element	Formula
C_0	$\frac{f_1(b_1 + b_2)}{2\pi Z_0(f_2^2 + f_1^2 b_1 b_2)}$	C_2	$\frac{b_1 b_2(f_2^2 - f_1^2)}{2\pi Z_0 f_1 f_2^2(b_1 + b_2)}$
C_0'	$\frac{f_2^2 + f_1^2 b_1 b_2}{2\pi Z_0 f_1 f_2^2(b_1 + b_2)}$	L_1	$\frac{Z_0(f_2^2 + f_1^2 b_1 b_2)^2}{2\pi f_1 f_2^2(b_1 + b_2)(f_2^2 - f_1^2)}$
C_1	$\frac{(b_1 + b_2)(f_2^2 - f_1^2)}{2\pi Z_0 f_1(1 + b_1 b_2)(f_2^2 + f_1^2 b_1 b_2)}$	L_2	$\frac{Z_0 f_2^2(b_1 + b_2)}{2\pi f_1 b_1 b_2(f_2^2 - f_1^2)}$

$$b_1 = \sqrt{\frac{1 - f\infty_1^2/f_1^2}{1 - f\infty_1^2/f_2^2}}; \qquad b_2 = \sqrt{\frac{1 - f\infty_2^2/f_1^2}{1 - f\infty_2^2/f_2^2}}$$

the location of the attenuation peaks of the network with relation to the cut-off frequencies and are given by the expression:

$$b_n = \sqrt{\frac{1 - f\infty_n^2/f_1^2}{1 - f\infty_n^2/f_2^2}}; \quad n = 1, 2, 3,$$

where $f\infty_n$ is the frequency of infinite attenuation.

These tables give the design formulae for the networks of Figs. 8 and 12. To obtain the equations for a network having crystals in the series arms alone, it is only necessary to let $b_3 = 0$. If one of the peaks of the filter of Fig. 8 (a) is placed at infinity—which results when $b_2 = f_2/f_1$—the two coils will have equal values and by the theorem illustrated by Fig. 5 can be brought out to the ends of the filter, simplifying the construction. In a similar manner if one of the peaks of the filter of Fig. 8 (b) is placed at zero frequency, i.e. $b_2 = 1$, the two shunt inductances are equal and can be brought out to the ends of the filter. The design equation of the narrow band filter of Fig. 12 with the lattice crystals replaced by condensers can be obtained from Table III by letting $b_2 = 0$.

Magnetic Generation of a Group of Harmonics*

By E. PETERSON, J. M. MANLEY and L. R. WRATHALL

A harmonic generator circuit is described which produces a number of harmonics simultaneously at substantially uniform amplitudes by means of a non-linear coil. Generators of this type have been used for the supply of carrier currents to multi-channel carrier telephone systems, for the synchronization ‘ carrier frequencies in radio transmitters, and for frequency comparison and standardization.

A simple physical picture of the action of the circu t has been derived from an approximate mathematical analysis. The principal roles of the non-linear coil may be regarded as fixing the amount of charge, and timing the charge and discharge of a condenser in series with the resistance load. By suitably proportioning the capacity, the load resistance, and the saturation inductance of the non-linear coil, the amplitudes of the harmonics may be made to approximate uniformity over a wide frequency range. The sharply peaked current pulse developed by condenser discharge passes through the non-linear coil in its saturated state and so contributes nothing to the eddy current loss in the core. In this way the efficiency of frequency transformation is maintained at a comparatively high value for the harmonics in a wide frequency band, even with small core structures. The theory has also been adequate in establishing a basis for design, and in evaluating the effects of extraneous input components.

I. Outline of Development

THE use of non-linear ferromagnetic core coils to generate harmonics started with a simple type of circuit due to Epstein[1] which appeared in 1902. Application of the idea was not made to any great extent until it was elaborated by Joly[2] and by Vallauri[3] in 1911. The frequency multipliers thus developed were limited to doublers and to triplers, polarization being required for the doubler. In these, as well as in subsequent developments, single and polyphase circuits were used, and various arrangements were adopted for the structure of the magnetic core and for the circuit, by which unwanted components were balanced out of the harmonic output path. Later developments had to do with improvements in detail, and with the generation of higher harmonics in a single stage and in a series of stages. The applications

* Presented at the Pacific Coast Convention of A. I. E. E., Spokane, Washington, September 2, 1937. Published in *Elec. Engg.* August 1937.

of perhaps greatest importance were to high power, long-wave radiotelegraph transmitters, where the fundamental input was obtained from an alternator. Other applications of the idea of harmonic production by magnetic means have been made in the power and communication fields.[4]

It appears that these circuits were all developed primarily to generate a single harmonic. Comparatively good efficiencies were obtained, values from 60 to 90 per cent being reported for the lower harmonics. The theory of frequency multiplication was investigated by a number of workers, among whom may be mentioned Zenneck [5] and Guillemin.[6] The latter, after analysis which determined the optimum conditions for the generation of any single harmonic, found experimentally that the efficiency of harmonic production decreased as the order of the harmonic increased. He obtained efficiencies of 10 per cent for the 9th harmonic, and 3 per cent for the 13th harmonic of 60 cycles.

Where the circuits are properly tuned and the losses low, free oscillations may be developed. The frequencies of these free oscillations may be harmonic, or subharmonic as in the circuit described by Fallou; [7] they may be rational fractional multiples of the fundamental, or incommensurable with the fundamental, as in Heegner's circuit.[8] The amplitudes of these free oscillations are usually critical functions of the circuit parameters and input amplitudes, and where the developed frequencies are not harmonic, they are characterized by the fact that the generated potentials are zero on open circuit. The theory of the effect has been worked out by Hartley.[9] It is presumably this effect which is involved in the generation of even harmonics by means of an initially unpolarized ferromagnetic core, an observation which has been attributed to Osnos.[10]

II. Circuit Description

The harmonic producer circuit which forms the subject of the present paper differs from those mentioned in that it is designed to generate simultaneously a number of harmonics at approximately the same amplitude.

Harmonics developed in circuits of this type have been used for the supply of carrier currents to various multi-channel carrier telephone systems, for synchronizing carriers used in radio transmitters, and for frequency comparison and standardization. Only odd harmonics are generated by the harmonic producer when the core of the non-linear coil is unpolarized, as is the case here. To generate the required even harmonics, rectification is employed. This is accomplished by means of a well balanced copper oxide bridge, which provides the even harmonics in a path conjugate to the path followed by the odd harmonics.

A typical circuit used for the simultaneous generation of a number of odd and even harmonics at approximately equal amplitudes is shown schematically in Fig. 1. Starting with the fundamental frequency input, a sharply selective circuit (F) is used to remove interfering components, and an amplifier (A) provides the input to the harmonic generator. The shunt resonant circuit (L_0C_0) tuned to the fundamental serves primarily to remove the second harmonic generated in the amplifier. The elements C_1L_1 are inserted to maintain a sinusoidal current into the harmonic producer proper, as well as to tune out the circuit reactance.

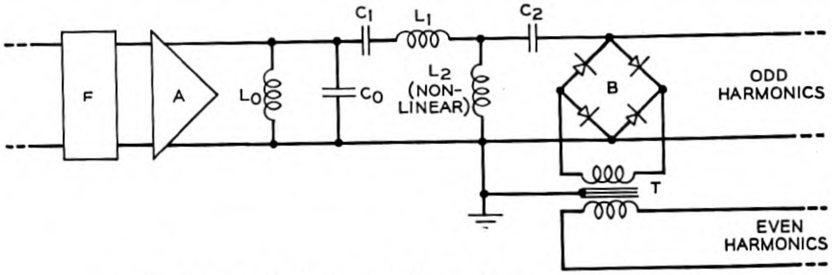


Fig. 1—Circuit diagram of channel harmonic generator.

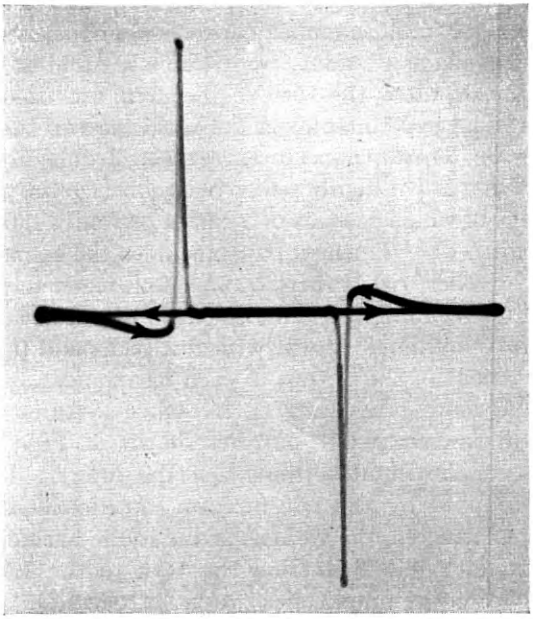


Fig. 2—Cathode ray oscillogram of output current wave form with fundamental input current as abscissa.

L_2 is a small permalloy core coil which is operated at high magnetizing forces well into the saturated region. The circuit including L_2, C_2, and the load impedance, which is practically resistive to the desired harmonics, is so proportioned that highly peaked current pulses rich in harmonics flow through it. Two such pulses, oppositely directed, are produced during each cycle of the fundamental wave, the duration of each being a small fraction of the fundamental period. The typical output wave shown in Fig. 2 was obtained by means of a cathode ray oscillograph, the ordinate representing the current in the load resistance, and the abscissa representing the fundamental current into the coil. The desired odd harmonics are selected by filters connected across the input terminals of the copper oxide bridge. The even harmonics are obtained by full-wave rectification in the copper-oxide bridge. They appear at the conjugate points of the bridge, and are connected through an isolating transformer to the appropriate filters. Thus the harmonics are produced in two groups, with the even harmonics separated from the odds to a degree depending largely upon the balance of the copper-oxide bridge, as well as upon the amount of second harmonic passed on from the amplifier. In this way the required discrimination properties of any filter against adjacent harmonics are reduced to the extent of the balance.

A particular application of the circuit described above to the generation of carriers for multi-channel carrier telephone systems uses a fundamental frequency of 4 kc., from which a number of harmonics are developed. Of these the 16th to the 27th are used as carriers. A photograph of an experimental model of this carrier supply system * is shown in Fig. 3. The top panel includes an electromagnetically driven tuning fork serving as the highly selective circuit (F), the amplifier (A), the output stage of which consists of a pair of pentodes in push-pull, and the tuned circuit L_0C_0. The next panel includes the elements L_1C_1, L_2, C_2, B, and T, together with a thermocouple and meter terminating in a cord and plug for test and maintenance purposes. The last two panels include the twelve harmonic filters, with test jacks and potentiometers for close adjustment of the output of each harmonic.

A few of the more interesting performance features are given in Fig. 4. The harmonic power outputs shown in Fig. 4*a* represent measurements at the input terminals of the filters. The variation observed is produced by the non-uniform impedances of the filters. When these are corrected, the variations due to the harmonic generator proper are less than ± 0.2 db from the 16th to the 27th harmonic. Outside this region the amplitudes gradually decrease to the extent

* Developed by J. M. West.

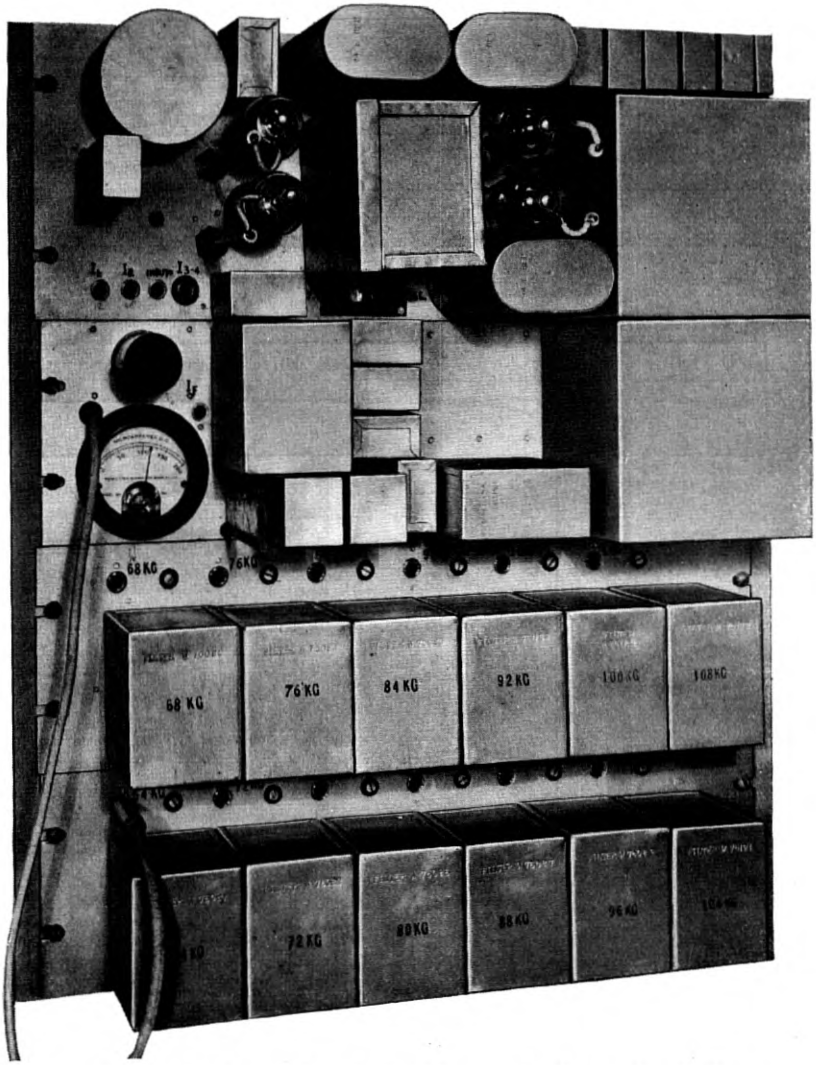


Fig. 3—Carrier supply unit, furnishing twelve harmonics of 4 kc. (experimental model).

of 4 db at the 3d and 35th harmonics, and 11 db at the fundamental and the 61st harmonic. The variation of harmonic output with change of amplifier plate potential is given for the two harmonics indicated in Fig. 4*b*. Figure 4*c* shows the 104 kc. output as a function of the 4 kc. input. Arrows are used to indicate normal operating points. The input amplifier is operated in an overloaded state so that beyond a critical input, the fundamental output of the amplifier and

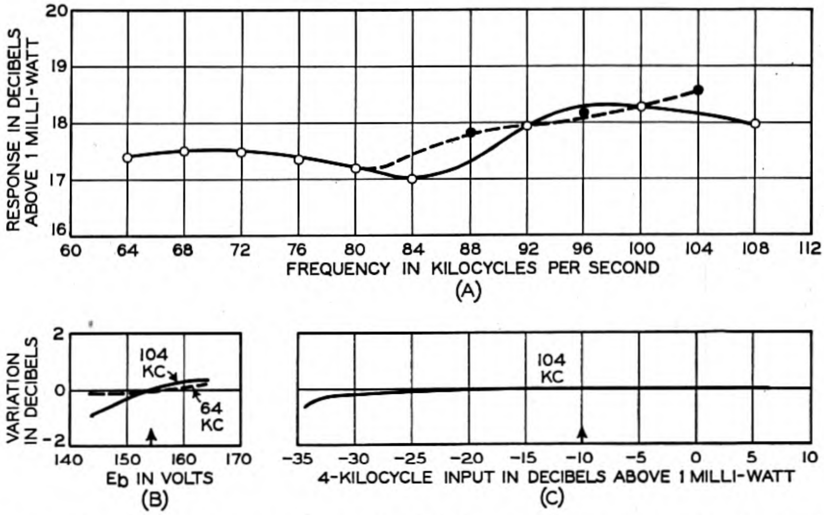


Fig. 4—Performance curves of channel harmonic generator. (*A*) Harmonic outputs; (*B*) Variation of 16th and 26th harmonics with amplifier plate potential; (*C*) Variation of 26th harmonic with fundamental input.

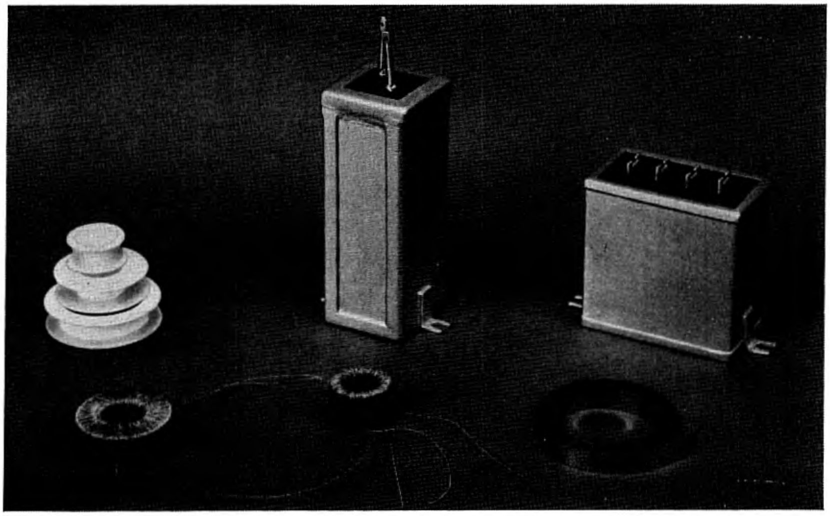


Fig. 5—Construction of experimental non-linear coils used for harmonic generation, showing core forms, magnetic tape, wound coils, and assembled units.

the harmonic output corresponding are but little affected by change of input amplitude. With a linear amplifier the harmonic output current would vary roughly as the four-tenths power of the input current.

Another application involving higher frequencies has been made to the generation of the so-called "group" carriers used in conjunction with a coaxial conductor.[11] There odd harmonics of 24 kc. from the 9th to the 45th are used. The circuit differs from Fig. 1 in that the copper oxide bridge is omitted, and the non-linear coil is provided with two windings to facilitate impedance matching. The performance of an experimental model is similar to that of the generator described above. A notion of the physical size and construction of the non-linear coils used may be had from the photographs of Fig. 5.

In both applications the required harmonics are generated at amplitudes high enough to avoid the necessity for amplification.

III. Theory of Operation

The analysis of operation of the harmonic generating circuit described above meets with difficulties, since a high degree of non-linearity is involved in working the coil well into its saturated region.

To avoid these difficulties, an expedient is adopted by which the hysteresis loop is replaced by a single-valued characteristic made up of connected linear segments [6] as shown in Fig. 6*b*. It is then possible to formulate a set of linear differential equations with constant coefficients, one for each linear segment. The solutions are readily arrived at and may be pieced together by imposing appropriate conditions at the junctions, so that a solution for the whole characteristic is thereby obtained. From this solution the wave form of current or voltage associated with any circuit element may be calculated. Resolution of the wave form into components may then be accomplished by an independent Fourier analysis.

The assumed *B-H* characteristic of Fig. 6*b* is made up of but three segments. While it is manifestly a naive representation of a hysteresis loop, it will be shown by comparison with experiment that the main performance features of harmonic generators may be reproduced by this crude model.

It will be noted on Fig. 6*b* that the differential permeability of the assumed non-linear core, a quantity proportional to dB/dH, takes on one of two values, determined by the absolute value of the magnetizing force. These are designated by μ in the permeable region and μ_s in the saturated region. The corresponding inductances are L_{20} and L_{2s}, L_{20} being many times greater than L_{2s}. The values of current through the coil at which the differential inductance changes are designated $\pm I_0$,

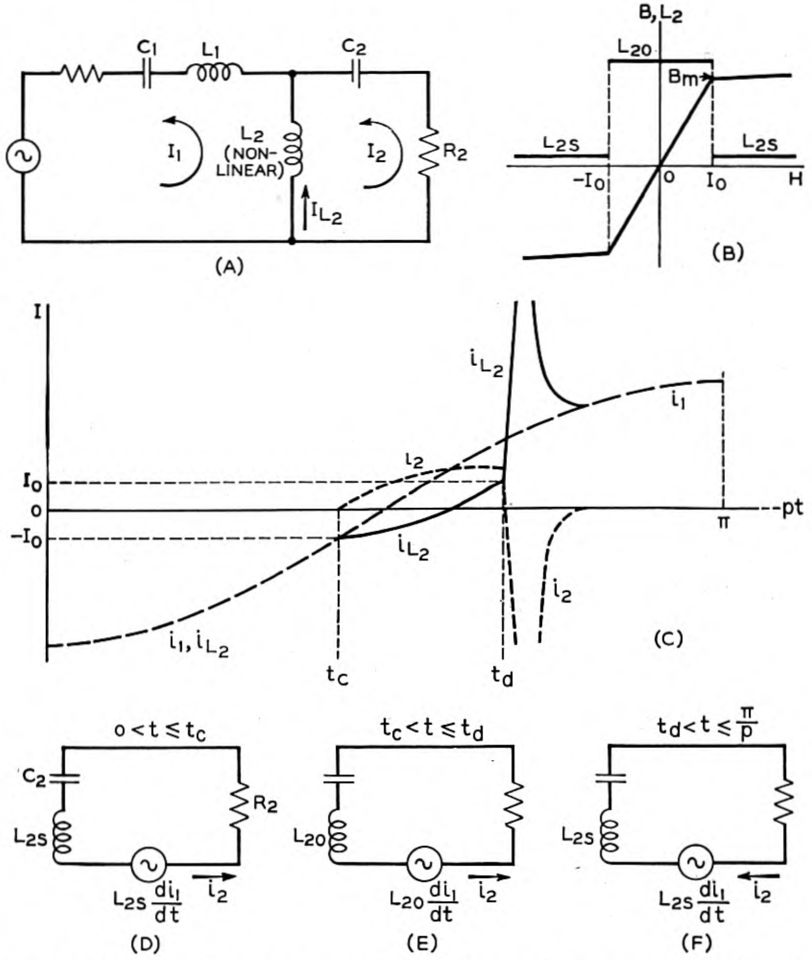


Fig. 6—Diagrams illustrating operation of the harmonic generator. (*A*) Harmonic generator circuit; (*B*) Differential inductance and flux density of assumed non-linear coil as functions of magnetizing force; (*C*) Variation with time of currents in primary and secondary meshes, and in non-linear coil; (*D*) (*E*) (*F*) Equivalent circuits of the harmonic generator for the three time intervals indicated.

corresponding to the magnetizing forces $\pm H_0$. With this simple representation of the non-linear inductance, the operation of the circuit shown in Fig. 6*a* will be described over a complete cycle of the fundamental input wave.

The current flowing in the input mesh is made practically sinusoidal by tuning L_1, C_1. If now we start at the negative peak of the sinusoidal input current of amplitude I_1 and frequency $p/2\pi$, the non-linear coil is worked in the saturated state where its inductance L_{2s} is low. Since the resistance of the winding is small, the potential drop across the coil is correspondingly small. The current i_2 which charges the condenser C_2, assuming the latter to have zero charge at the start, is therefore negligible as indicated in Fig. 6*c*. This state of affairs is maintained until the current through L_2 reaches the value $-I_0$, at time t_c. At this point the inductance of the coil increases suddenly to L_{20} and the voltage across the coil tends to increase. Hence the current i_2 increases and C_2 is charged much more rapidly than in the preceding interval. Charging continues until the current through the coil increases through I_0 at time t_d. At that time, the coil inductance returns to the low saturation value L_{2s}, and the potential across the coil decreases. The condenser potential is no longer opposed by the potential drop across the coil and the condenser discharges through R_2 and L_{2s}; i_2 reverses its direction, maintaining the coil in the saturated region. The form and duration of the sharply peaked discharge pulse characteristic of this type of harmonic generator are determined by the values of the elements just mentioned. The resistance, capacity, and saturation inductance effectively in circuit are adjusted to permit the current to rise to a high maximum, to damp the pulse, and to shorten the pulse duration to the point at which the highest harmonic required reaches the desired amplitude. Under the working conditions which will be assumed in the following, this insures that the pulse dies away before the end of the half-cycle as shown in Fig. 6*c*. At that time the currents and potentials are the same, except for reversals of sign, as those at the start, so that the current wave consists of an alternating succession of these pulses. Equivalent circuits for the three respective time intervals of a half-cycle are shown in Figs. 6*d*, 6*e*, 6*f*. The similarity of the load current wave form derived above to that experimentally observed and shown in Fig. 2, is to be noted.

The course of events described above parallels closely conclusions drawn from the mathematical analysis. This picture attributes to the coil L_2 a sort of switching property which permits the condenser C_2 in the load circuit to be charged and discharged alternately. The charge starts when the large inductance L_{20} is switched across the primary and secondary meshes, thus permitting energy to flow from the primary circuit into the condenser C_2. This corresponds to that part of the wave described above during which the load current slowly rises as the charge accumulates on C_2. Discharge starts when the large inductance L_{20} is switched out and the much smaller inductance L_{2s} is switched in. This sharply reduces the voltage across L_2, and the condenser is discharged through the load resistance and the saturation inductance.

During this interval the secondary circuit is practically isolated from the primary. The switching process is sustained by the alternations of the sinusoidal primary current and is periodic, as we have seen, since similar conditions exist at the start of each pulse. The times at which switching occurs are those at which the current through the coil passes through the critical values ($\pm I_0$) where the inductance changes.

Since the narrow discharge pulse provides the principal contribution to the higher harmonics in which we are interested, and since this discharge takes place in the secondary independently of the primary, the elements of the secondary mesh during discharge determine the form of the output spectrum. From this viewpoint we may regard the condenser as the source of energy for these harmonics and hence as a possible location for equivalent harmonic generator e.m.f.'s. In this light, the discharge circuit becomes a half-section of low-pass filter terminated in resistance R_2, with L_{2s} as the series element and C_2 as the shunt element.

IV. Quantitative Results of Analysis

To connect the three solutions which hold for the three linear regions of the B-H characteristic, conditions at the junctions are introduced which lead to transcendental equations. These may be solved graphically when definite values are assigned to the circuit parameters. From these may be obtained the maximum value Q_m of charge on C_2 which is reached at the end of the charging stage.

By plotting a representative group of these final charges over a range of parameters ordinarily encountered, an empirical equation has been deduced for Q_m as follows:

$$Q_m = \sqrt{2}\,\frac{I_1}{p}\left(\frac{pL_{20}}{R_2}\right)^{0.75}(pC_2R_2)^{0.65}\left(\frac{I_0}{I_1}\right)^{0.6}. \tag{1}$$

For the usual operating conditions the narrow peaked discharge part of the current pulse is most important in the determination of the higher harmonics (say beyond the 9th) with which we are concerned here. The charging interval then may be neglected in calculating the higher harmonics. The form of the discharge pulses is determined by the parameters pC_2R_2 and k, where

$$k = L_{2s}/R_2^2C_2.$$

The familiar criterion for oscillation in a series circuit containing inductance, capacity and resistance may be expressed in terms of k. If $k > \frac{1}{4}$, the discharge is an exponentially decaying oscillation; if

$k \leq \frac{1}{4}$, the discharge is an exponentially decaying pulse. This last condition is the one assumed in the description of operation given above.

If the discharge is oscillatory, and if further the second peak is large enough, the current through the coil may become less than I_0 during the discharge interval. Thus L_2 will return to its larger value, and recharging of the condenser will result. This process may lead to large and undesired variations in the amplitudes of the harmonics. To maintain the frequency distribution as uniform as possible over the frequency range of interest, the circuit parameters are usually adjusted so that recharging does not occur.

Harmonic analysis shows that the nth harmonic amplitude under the above assumptions is given by

$$I(n) = \frac{(2/\pi)pQ_m}{\sqrt{1 + (1 - 2k)(npC_2R_2)^2 + k^2(npC_2R_2)^4}}, \tag{2}$$

where n is odd. This expression neglects the contributions due to the charging stage, which are usually small for harmonics higher than the ninth.

The corresponding harmonic power output is

$$W_n = \frac{I(n)^2R_2}{2} = \frac{W_0}{1 + (1 - 2k)(npC_2R_2)^2 + k^2(npC_2R_2)^4}, \tag{3}$$

where W_0 is a convenient parameter which does not vary with n and hence serves as an indication of the power of the output spectrum. It is related to W, the total power delivered to the load resistance, by the equation,

$$W_0 = \frac{4}{\pi} pC_2R_2W.$$

For purposes of calculation, W_0 may be found from (1) and (2) to be

$$W_0 = \frac{10^{-7}}{\pi^2} pB_mAdH_1 \left(\frac{H_0}{H_1}\right)^{0.2} (pC_2R_2)^{1.3} \left(\frac{pL_{20}}{R_2}\right)^{0.5} \text{watts}, \tag{4}$$

where

$$L_{20} = \frac{4N^2A\mu 10^{-9}}{d}, \quad H_1 = 0.4NI_1/d,$$

and N is the number of turns wound on the toroidal core of diameter d cm. and cross-sectional area A cm.^{2}

In Fig. 7 the power spectrum is shown by plotting W_n in db above or below W_0 as a function of npC_2R_2 for several values of k. These curves illustrate the degree of uniformity obtainable in harmonic amplitudes under different conditions. It may be shown from (3) that W_n has a maximum with respect to n when k is greater than $\frac{1}{2}$, if

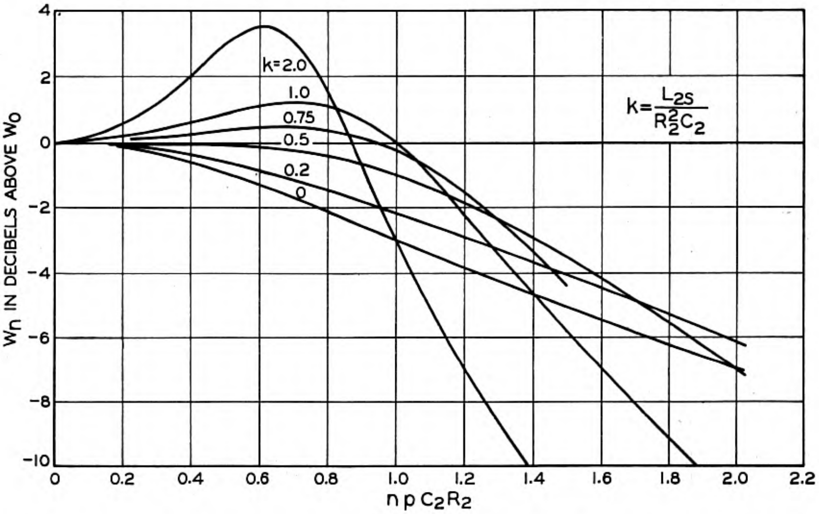


Fig. 7—Harmonic power spectrum plotted from eq. (2) as function of npC_2R_2 with k as parameter.

$$npC_2R_2 = \frac{1}{k}\sqrt{k - \tfrac{1}{2}},$$

and that its value at this point is

$$(W_n)_{\max} = W_0k^2/(k - \tfrac{1}{4}).$$

A number of relations may be derived from these equations which are useful for design purposes. Thus the form of harmonic distribution is fixed by k and pC_2R_2. The output power for a given magnetic material worked at a given fundamental magnetizing force then depends solely upon the volume of core material. Finally, the impedance is fixed by the number of turns per unit length of core. If the impedances desired for primary and secondary circuits differ, separate windings may be used for each circuit.

V. Calculated and Observed Performance

In order to make practical use of the results given above, we need some basis for deriving the assumed parameters of the non-linear coil from the physical properties of the magnetic materials used in harmonic producers.

The fact that the actual magnetization curve is a loop instead of a single-valued curve as assumed requires increased power input to the circuit to provide for the hysteresis and eddy losses in the core. Other than this, the principal remaining effect of the existence of a loop is a lag in the time at which the pulses occur, an effect which is of no great moment in determining the form or magnitude of the resulting pulses.

The next point requiring consideration is the effect introduced by the assumed abrupt change of slope contrasted to the smooth approach to saturation actually observed. While no rigorous comparisons can be drawn, the effect of the more gradual approach to saturation was approximated analytically by introducing an additional linear segment between the permeable region and each saturated region of the *B-H* characteristic, at a slope intermediate between the two, so as to form a *B-H* characteristic of five segments in place of the original three. The solutions for these two characteristics were found to yield negligibly small differences in the amplitudes of the higher harmonics. It was inferred from this result that no substantial change would be introduced by a smooth approach to saturation.

Finally, the actual *B-H* characteristic has a slight curvature in the saturated region, while the analysis considered a small linear variation. A rough approximation for the effect of this curvature, which leads to fair agreement with experiment, consists in taking for L_{2s} the average of the actual slope, from its minimum value reached during the discharge peak down to the point at which the slope is one-tenth maximum. To this is added the linear inductance contributed by the dielectric included within the winding.

To summarize then, the harmonic outputs obtained from the analysis with the assumed *B-H* characteristic may be brought into line with experimental observations by the introduction of quantities obtained from actual *B-H* loops at appropriate frequencies and magnetizing forces. In these the maximum slope found on the loop is taken for L_{20}, the average slope over the saturated region is taken for L_{2s}, and the energy corresponding to the area of the real *B-H* loop must be added to that originally supplied the harmonic generator input.

A comparison between measured and calculated harmonic distributions obtained with a 4-kc. fundamental input is shown in Fig. 8. In this case the harmonic distributions were measured for four different values of the secondary condenser C_2 as shown by the plotted points. The power output of each harmonic is plotted in terms of the quantity npC_2R_2. Calculated values are indicated by dashed lines. It is observed that while the agreement between calculation and experiment

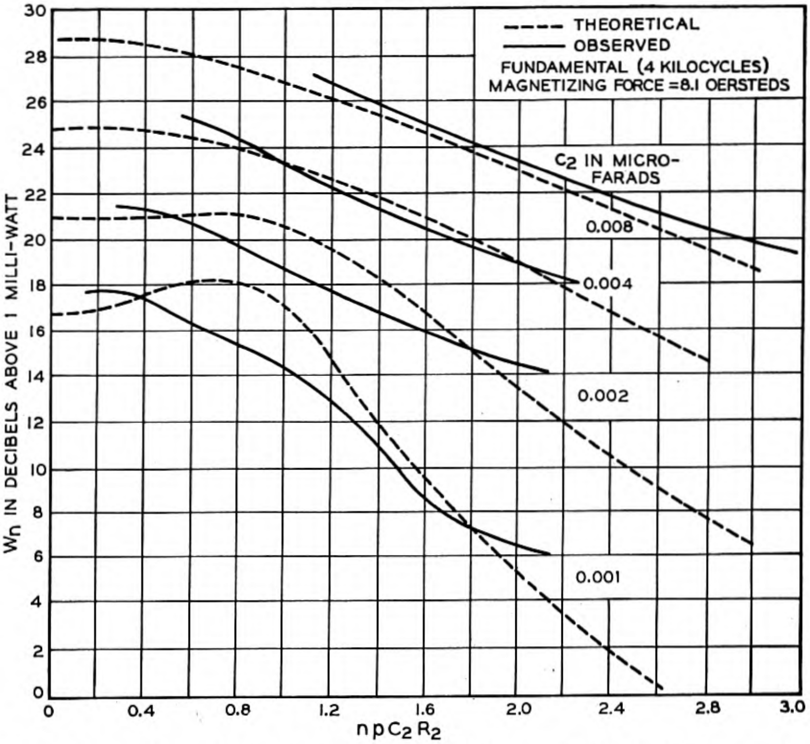


Fig. 8—Comparisons of calculated and measured harmonic distributions, plotted as functions of npC_2R_2, with C_2 as parameter.

is perhaps as good as could be expected for the two highest curves, a substantial divergence is noticed in the two lowest sets; the forms of the two sets are significantly different, and it seems that the divergence might become even greater at larger values of npC_2R_2 than those shown. Upon examination of the equations, however, it turns out that the conditions existing for the lowest pair of curves are just those for which recharging occurs, so that the conditions for which the equations were framed hold no longer. The calculated distributions might be expected to be too low for the higher harmonics, since we have taken an average value for the saturation inductance. This means that the peak of the discharge pulse will be sharper than that calculated, with a corresponding effect upon the higher harmonics.

Another comparison between calculated and observed values is shown in Fig. 9 for a fundamental input of 120 kc. with two values of resistance load. Fair agreement is observed over the greater part of the frequency range which extended to 5 MC. The distribution curve for the smaller resistance load undulates as the load resistance is reduced, since multiple oscillations and recharging are then promoted, in consequence of which the output power tends to become concentrated in definite bands of harmonics. In general, agreement within a few db is found over a wide range of circuit parameters when working into a resistance load, provided that recharging does not occur.

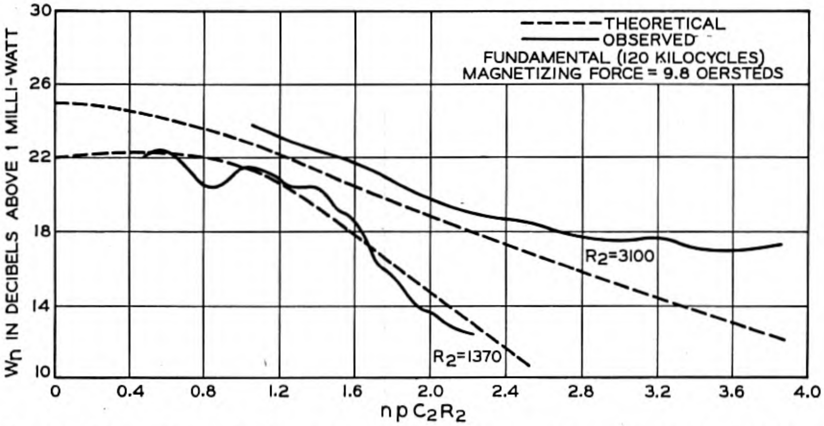


Fig. 9—Comparisons of calculated and measured harmonic distributions plotted as functions of npC_2R_2, with R_2 as parameter.

When the resistance termination is replaced by a bank of filters as it is in practice, the resistance termination is approximated over the frequency band covered by the filters. Where the band is wide the results obtained do not differ greatly from those with the pure resistance load, but when only a few harmonics are taken off by filters and the impedances to the other harmonics of large amplitude vary widely over the frequency range, then the wave form of the current pulse is substantially altered, with corresponding effect upon the frequency distribution, and the calculations for a pure resistance termination do not apply.

A difficulty sometimes arises in getting a desired value of fundamental current into the coil. Under certain circuit conditions the current amplitude is found to change rapidly as the input voltage is smoothly varied. This phenomenon has been described by various terms such as Kippeffekt, ferro-resonance, and current-hysteresis.[13] If the operating point is located close to one of these discontinuities, the fundamental input and harmonic output may vary widely with small changes in supply potentials and circuit parameters. This troublesome source of variation may be avoided in a number of different ways, of which the simplest is to increase the resistance of the resonant mesh. In the present case this is effectively accomplished without sacrificing efficiency by using pentodes, which have high internal resistances, in the amplifier stage connected to the resonant mesh.

The efficiency of power conversion from fundamental to harmonics may be found from the fundamental power input to the circuit, as derived from measurements on a cathode ray oscillograph, and from the total harmonic output measured by means of a thermocouple. The maximum efficiency obtainable with the low-power circuits described in the second section is in the neighborhood of 75 per cent, and decreases with increasing fundamental frequency because of the increased dissipation due to eddy currents. It should be noted that this figure does not include losses in the primary inductance L_1. When only a few harmonics are used, the efficiency of obtaining this useful power naturally drops to a much lower value, which for the particular cases mentioned in the second section, is between 15 and 25 per cent.

VI. Effect of Extraneous Components

In any practical case the fundamental input to the harmonic producer is accompanied by extraneous components introduced by crosstalk, by modulation, or by an impure source. Thus if the fundamental is derived as a harmonic of a base frequency, small amounts of adjacent harmonics will be present. Or if the amplifiers are a.-c. operated, side-frequencies are produced differing from the fundamental by 60 cycles and its multiples. Extraneous components of this sort in the input modulate the fundamental and produce side-frequencies about the harmonics in the output. When the harmonics are used as carriers, the accompanying products must be reduced to a definite level below the fundamental if the quality of the transmitted signal is to be unimpaired. The requirements imposed by this condition can be calculated by simple analysis, the results of which agree rather well with experimental values.

The method of analysis used is to consider the extraneous component at any instant as introducing a bias [14] to the non-linear coil. The primary effect of a small bias (b) is to shift the phase of the discharge pulse by $\mp b/H_1$ radians, H_1 being the amplitude of the fundamental magnetizing force. The sign of the shift alternates so that intervals between pulses are alternately narrowed and widened.

The effect of this shift on the harmonics produced may be found by straightforward means in which the amplitude of any harmonic is expressed in terms of the bias. Hence when the extraneous component or components vary with time, the sidebands produced may be evaluated when the bias is expressed by the appropriate time function.

If the bias is held constant, the wave is found to include both odd and even harmonics, the amplitudes of which are given by

$$\left.\begin{aligned} I_n &= I(n)\,|\cos nb/H_1|, \quad (n \text{ odd}), \\ &= I(n)\,|\sin nb/H_1|, \quad (n \text{ even}), \end{aligned}\right\} \tag{5}$$

$I(n)$ being the harmonic distribution in the absence of bias as given by eq. (2).

If the extraneous input component is sinusoidal, we have

$$b = Q \sin (qt + \varphi). \tag{6}$$

Substituting this expression for b in the equation for the harmonic components yields odd harmonics of the fundamental, and modulation products with the angular frequencies $mp \pm lq$, which may be grouped as side-frequencies about the odd harmonics. The amplitude of the nth (odd) harmonic is

$$I_n = I(n)\left|J_0\left(\frac{nQ}{H_1}\right)\right|, \tag{7}$$

and the amplitude of the modulation product $mp \pm lq$ is

$$I_{m.\,\pm l} = I(m)\left|J_l\left(\frac{mQ}{H_1}\right)\right|, (m + l \text{ odd}), \tag{8}$$

where $J_l(x)$ is the Bessel function of order l.

Considering the side-frequencies about the nth harmonic, the largest and nearest of these are $(n + 1)p - q$ and $(n - 1)p + q$, n being odd. The ratio of the amplitudes of either side-frequency to the nth harmonic is

$$\frac{I_{n\pm1,\,\mp1}}{I_n} = \left|\frac{J_1[(n \pm 1)Q/H_1]}{J_0(nQ/H_1)}\right|, \tag{9}$$

on the assumption that the harmonic distribution in the neighborhood of n is uniform so that $I(n \pm 1) \doteq I(n)$. If the arguments of the Bessel functions are less than four-tenths, a good approximation to the right member of eq. (9) is $(n \pm 1)Q/2H_1$. Hence with sufficiently small values of interference, the sidebands produced are proportional to the amplitude of the interference, and increase linearly with the order of the harmonic. These relations apply to harmonic generators which produce sharply peaked waves in general, and are not peculiar to the magnetic type.

Neighboring modulation products involving the interfering component q more than once have much smaller amplitudes in normal circumstances than the product considered above. Because of the tuning in the input mesh, interfering components far removed in frequency from the fundamental are greatly reduced and the most troublesome interference is likely to be close in frequency to the fundamental.

It may be noted that where the interference is produced by amplitude modulation of the fundamental, so that two interfering components enter the input, the distortion produced may be approximated by doubling the amplitudes of the side-frequencies produced by one of the interfering components. If the disturbance is the second harmonic of the fundamental, the effect is nearly the same as that for constant bias, and the relations (5) may be used if b is taken as the amplitude of the second harmonic magnetizing force.

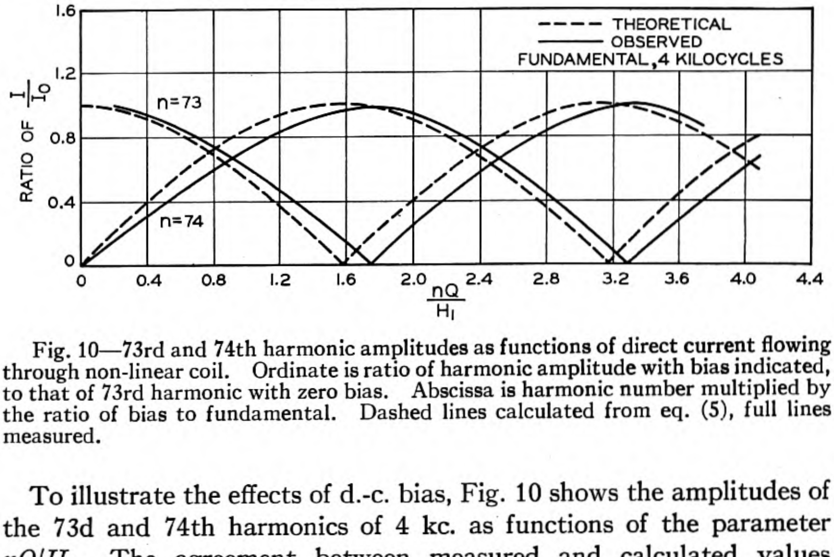


Fig. 10—73rd and 74th harmonic amplitudes as functions of direct current flowing through non-linear coil. Ordinate is ratio of harmonic amplitude with bias indicated, to that of 73rd harmonic with zero bias. Abscissa is harmonic number multiplied by the ratio of bias to fundamental. Dashed lines calculated from eq. (5), full lines measured.

To illustrate the effects of d.-c. bias, Fig. 10 shows the amplitudes of the 73d and 74th harmonics of 4 kc. as functions of the parameter nQ/H_1. The agreement between measured and calculated values indicates that the most important effects of bias have been included in the simple analysis.

References

1. *E. T. Z.*, v. 25, p. 1100, 1904.
2. *C. R.*, v. 152, p. 856, 1911.
3. *E. T. Z.*, v. 32, p. 988, 1911.
4. Cantwell, *Elec. Engg.*, v. 55, p. 784, 1936.
5. *Proc. I. R. E.*, v. 8, p. 468, 1920.
6. *Arch. f. El.*, v. 17, p. 17, 1926, and *Proc. I. R. E.*, v. 17, p. 629, 1929.
7. *Rev. Gen. d'El.*, v. 19, p. 987, 1926.
8. *Zeit. f. Fernmeldetechnik*, v. 5, p. 115, 1924.
9. Peterson, *Bell Labs. Record*, v. 7, p. 231, 1929.
10. Kasarnowski, *Zeit. f. Phys.*, v. 30, p. 225, 1924.
11. Espenschied and Strieby, *Elec. Engg.*, v. 53, p. 1371, 1934; *Bell Sys. Tech. Jour.*, v. 53, p. 654, 1934.
12. Elmen, *Elec. Engg.*, v. 54, p. 1292, 1935; *Bell Sys. Tech. Jour.*, v.15, p. 113, 1936.
13. Casper, Hubmann and Zenneck, *Jahrbuch*, v. 23, p. 63, 1924; Rouelle, *C. R.*, v. 188, p. 1392, 1929.
14. Peterson and Llewellyn, *Proc. I. R. E.*, v. 18, p. 38, 1930.

The Vodas *

By S. B. WRIGHT

Since the first transatlantic radio telephone circuit was opened for service over ten years ago, an increasing number of voice-operated switching devices has been added to the international telephone network. All of these have the common purpose of preventing echo and singing effects due to arranging the facilities to give the best possible transmission, even under difficult radio conditions. Differences in the design and performance of the several types of devices suggest that the advantages and disadvantages of each be made available.

The characteristics of two types of "vodas" used on circuits connecting with the United States are described in this paper. For reference purposes, a complete list of Bell System papers relating to these devices is included.

Introduction

THE interconnection of ordinary telephone systems by means of long radio-telephone links presents some unique and interesting technical problems. Since radio noise is often severe as compared with that in wire lines, radio transmitter power capacity is relatively large and expensive, and it is in general economical to control the speech volumes so that the radio transmitter will be fully loaded and thus the effect of noise minimized for a given transmitter power rating. This volume control, to be fully effective, calls for voice-operated switching devices to suppress echoes and singing.

This paper describes the measures which have been developed for use at radio-wire junctions in the United States. They are based upon an arrangement called a "vodas." This word, devised to fill a need for verbal economy, is formed from the initial letters of the words "*v*oice-*o*perated *d*evice *a*nti-*s*inging"; and thus implies not only a suppressor of feedback or singing, but also automatic operation by voice waves.

The general principles and applications of the vodas have been discussed from time to time in various papers listed at the end of this text. The present paper goes somewhat more into detail regarding the transmission performance of the vodas, including a description of an improved form of circuit which discriminates between line noise and the syllabic characteristics of speech.

* Presented at the Pacific Coast Convention of A.I.E.E., Spokane, Washington, September 2, 1937. Published in *Elec. Engg.*, August, 1937.

Historical Background

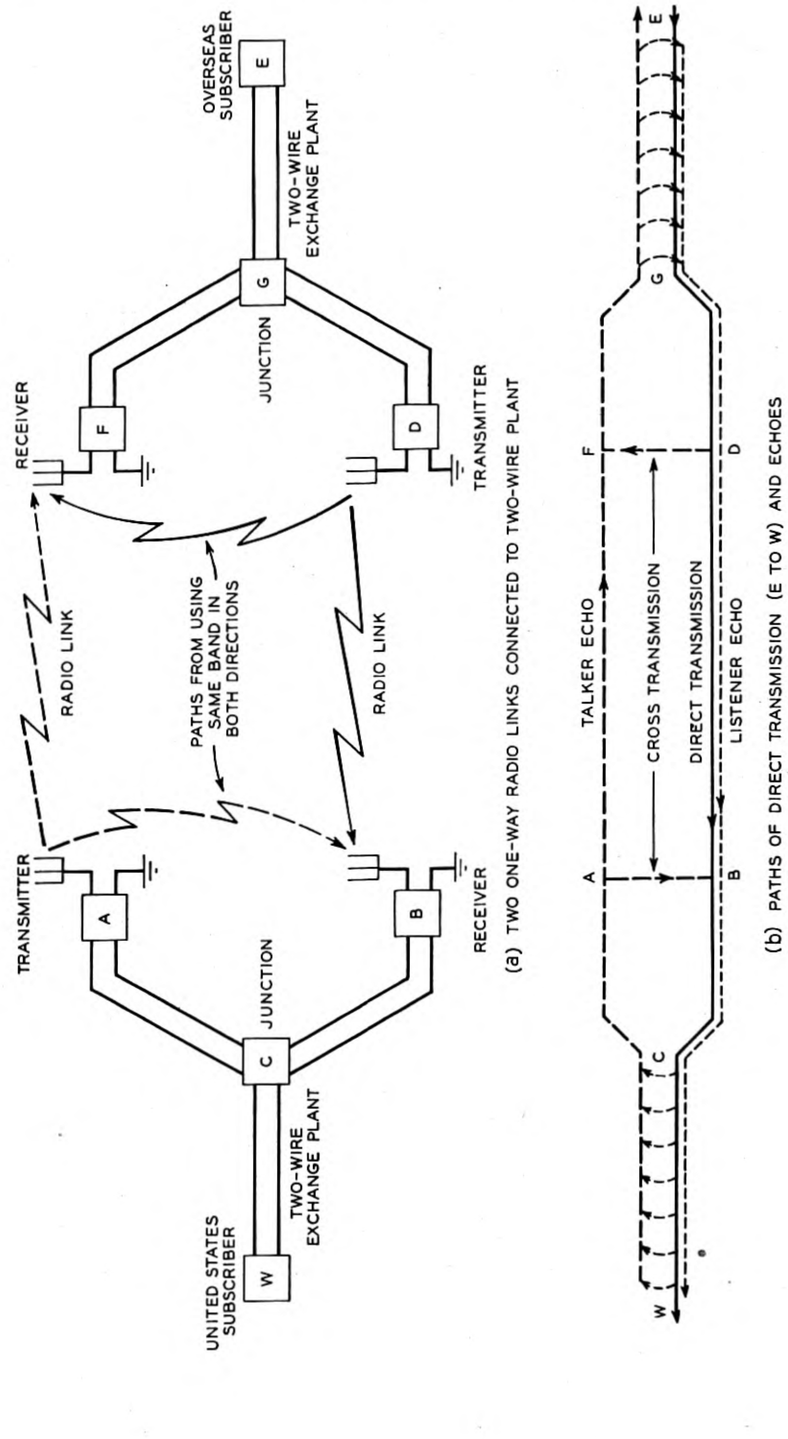
The two-way problem in telephony began with the invention of the telephone itself, and was the subject of considerable pioneering activity during the latter part of the nineteenth century. The invention of the amplifier brought about new problems when applied in a repeater for two-way operation. Even before a practical repeater had been devised, inventors visualized controlling the direction of transmission through amplifiers in a line by relays controlled from switches associated with the subscribers' instruments, an idea which is in use today on airplanes and small boats and in special circuits where this type of two-way operation is practicable. It is also used by amateur radio telephone operators. But for public telephone service more rapid and automatic control of two-way conversation is preferable.

To control the direction of transmission in a manner that would meet public convenience, invention progressed through the early part of the twentieth century toward devices for switching the speech paths automatically by voice waves. During this period, long distance radio telephony was first demonstrated to be practical on a one-way basis.

From that time until the first transatlantic radio telephone circuit was placed in service on January 7, 1927, anti-singing voice-operated devices underwent a process of development aimed at meeting the requirements of two-way radio telephone service. The vodas was one result. Since 1927, improvements have been made in cheapening and simplifying the equipment and in making a vodas that will operate better on speech and not so frequently on noise. It has also been possible to arrange a vodas so as to permit using the same privacy apparatus for both directions of transmission, thereby saving the cost of duplicate apparatus.

The Radio Telephone Problem

The conditions encountered when joining two-wire two-way circuits by radio links are illustrated in Fig. 1 in which (*a*) shows a connection between two subscribers, *W* and *E*, while (*b*) shows the paths of direct transmission and echo when *E* talks. In addition to the talker and listener echoes which arise in such a connection, singing can occur around the closed circuit *CAFGDBC* if the amplification is great enough. Also, when the same frequency band is used to transmit in both directions, two cross-transmission paths *AB* and *DF* are set up, and echoes and singing can take place around the end paths *ABC* and *DFG*. Any echoes or singing are of course primarily due to reflections of energy at points of impedance irregularities in the two-wire plant, including the subscribers' telephones themselves.



(a) TWO ONE-WAY RADIO LINKS CONNECTED TO TWO-WIRE PLANT

(b) PATHS OF DIRECT TRANSMISSION (E TO W) AND ECHOES

Fig. 1—Echoes in a radio telephone connection.

In wire circuits, simple hybrid coils and echo suppressors [2] are usually adequate to prevent such effects because the gains are not increased to provide for loading the circuit with energy when speech is weak, and also because the cross-transmission paths are absent. In long radio circuits, however, singing may result from the adjustments of amplification made to load the radio transmitter in case of weak speech and thus override noise, even though separate frequency bands are used in the two directions. Moreover, it is desired that the users of the service have as good transmission over the entire connection, including these radio links, as that to which they are accustomed in their own wire telephone systems, and even better transmission may be desired owing to differences in the language habits of the subscribers. Consequently, the overall transmission efficiencies of intercontinental radio circuits are sometimes better than those of the best land lines in the areas to be interconnected.

Fundamentals of Vodas Operation

A voice-operated device to suppress singing effects can be designed to have three possible arrangements:

1. The terminal can normally be blocked in one direction and connected through in the other.

2. Both directions of transmission can normally be blocked and activated in either direction but not both directions by the voice waves.

3. The circuit can remain activated in the last direction of speech and blocked in the other direction.

Where there is no noise on the transmission system under consideration any of these three arrangements will give satisfactory operation as there is then nothing to prevent making the voice-operated devices as sensitive as may be necessary to obtain full operation on weak as well as on strong voice waves. If there is any noise on the system which tends to operate the device it is necessary to make it less sensitive to avoid false operation. A point may be reached where the sensitivity is so low that the weakest parts of speech will not cause operation, and the weak consonants will be lost. The reduction in articulation has been found to be proportional to the time occupied by these lost or "clipped" sounds.[9]

If the device is located at a point in the circuit where the signal-to-noise ratio coming from one direction is poorer than that coming from the opposite direction it is obvious that a considerable advantage will be gained by using arrangement 1, since the device may be pointed in

[2] See references at end of text.

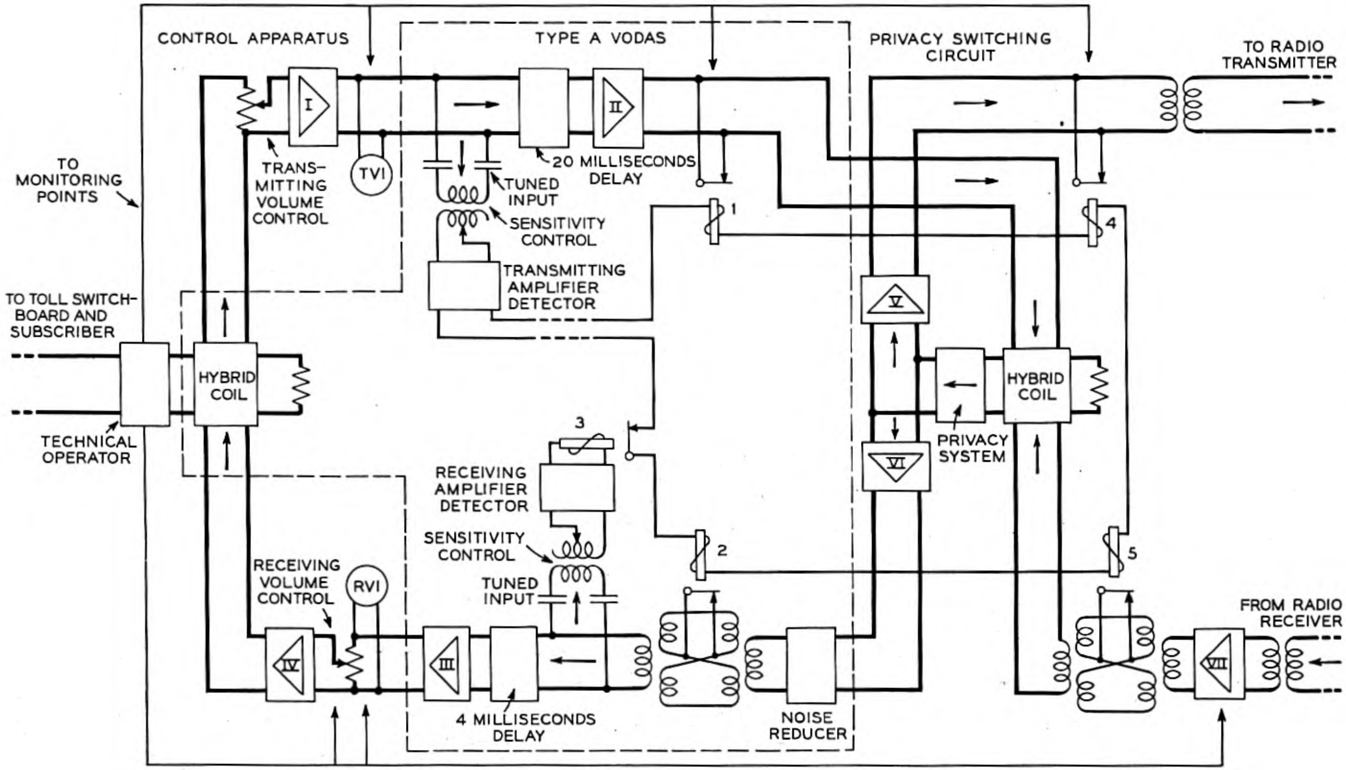


Fig. 2—Schematic of type A control terminal.

the direction in which the normally blocked path is exposed to the better signal-to-noise ratio and the normally activated path is exposed to the poorer signal-to-noise ratio. The vodas is, of course, arranged so that the normally blocked (transmitting) side is exposed to the land lines, which are usually quieter than the radio links. In the receiving side, the device can be less sensitive because there is no need for having it completely operated under control of the voice waves. All that is necessary is to have this side sensitive enough to operate in response to comparatively large voice or noise waves which might otherwise, after reflection and passage into the outbound path, result in false operation of the more sensitive side associated with this path.

In the vodas the principle of balance is used to keep the reflected currents small and thus allow the sensitivity of the normally activated device to be further reduced if necessary. Where a high degree of balance is not obtained and when noise from the radio limits the sensitivity of the receiving device it is sometimes necessary, particularly for weak outgoing volumes, to reduce the incoming volume so as to prevent echoes from operating the normally blocked transmitting side.

This echo limitation is primarily due to noise in the radio link, reflections from the two-wire plant and weak volumes from the subscribers. It is difficult to produce any large improvement in talker volumes and balance; so it would appear that the solution of the difficulty would probably come from the direction of improving radio transmission. Some benefit has also been obtained by reducing the effect of radio noise on the vodas with special devices of which the "Compandor"[17, 18] and the "Codan"[19, 20] are examples. More recently, use has been made of a new voice-controlled device called a "Noise Reducer"[21, 22] which reduces the received noise between speech sounds.

Vodas Design—Type A Control Terminal

Figure 2 shows a schematic diagram of a vodas * arranged to use the same privacy device for both transmitting and receiving. This is the type used on transatlantic and other long routes. Since the operation of this arrangement has been described before,[13] it will not be repeated here.

The diagram of the relay circuit in Fig. 3 shows how various time actions are obtained. Relays 1, 2, 4 and 5 are operated from battery B_1 when the ground contact of relay TM is opened. Thus the travel time of any relay armature is not a factor in securing fast initial

* The vodas apparatus, together with the volume control devices and technical operator's circuits, go to make up what is called a *Type A Control Terminal.*

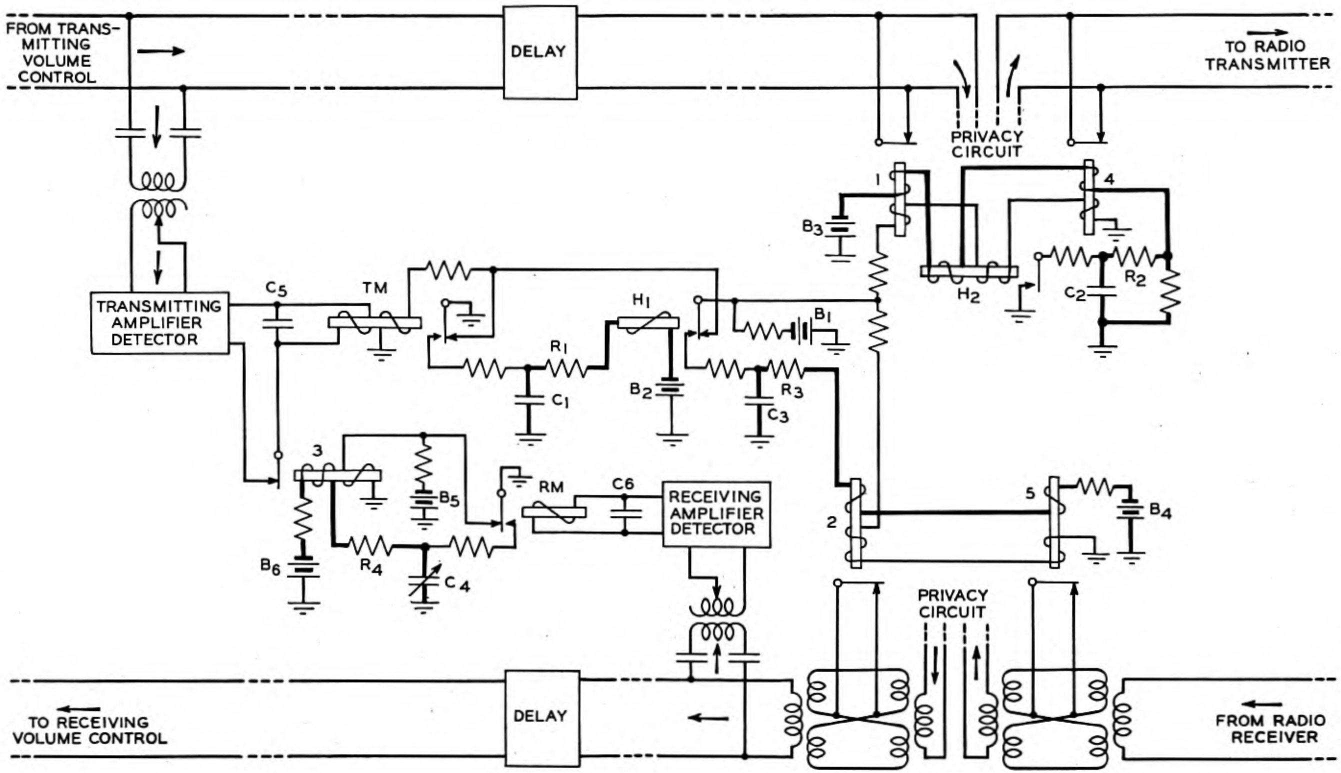


Fig. 3—Vodas relay circuits.

operation. When the armature of relay TM reaches its left-hand contact, relay H_1 operates and delays release of the relay train even if TM is at once restored to normal. H_1 is delayed in releasing by the time required to charge condenser C_1. The final release of relays 1 and 4 is then controlled by the time constant of an auxiliary circuit involving relay H_2 and condenser C_2, while that of relays 2 and 5, which is made later so as to suppress delayed echoes, is controlled by the circuit charging C_3. On the receiving side, condenser C_4 is adjustable so as to permit the technical operator to select the shortest release time for suppressing the delayed echoes in a given land line extension.

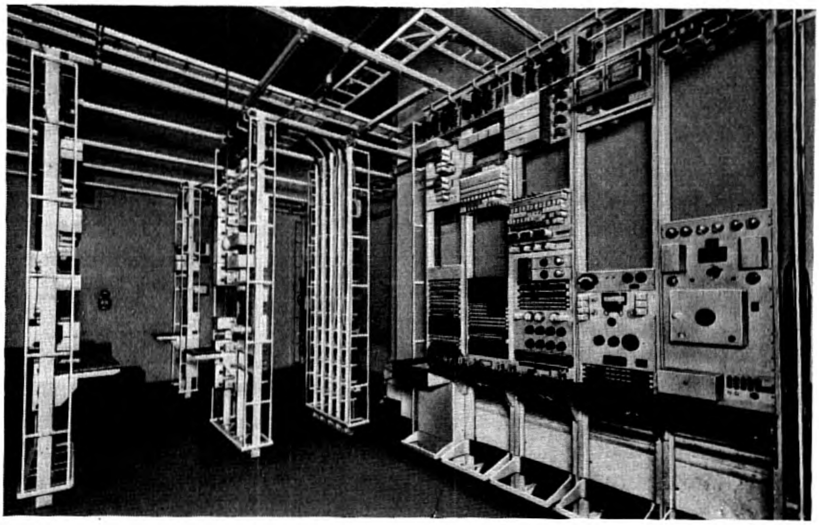


Fig. 4—Type A control terminal at San Francisco.

The vodas control terminal of the A type [8] used at New York consists of a line of technical operating positions with cross-connections to other lines of equipment containing the delay units, repeaters, vodas amplifier-detectors and privacy apparatus. Figure 4 shows an arrangement of a single terminal at San Francisco. The control bay is placed between two line testing bays on the left and two transmission testing bays on the right of the operating lineup. The distributing frame is in the center of the picture; and repeaters, ringers and privacy apparatus are shown at its left. At the extreme left is the vodas bay.

Syllabic Vodas—Type B Control Terminal

The desire for a cheaper control terminal than the Type A led to the development of a second type, known as *Type B*, in which the vodas employs the same fundamental principles. In this vodas added protection against false operation from line noise is secured by the use of a new principle in voice-operated devices, called "syllabic" operation.

It is observed that in many types of noise a large component of the long-time average power is steady. Speech, however, comes as a series of wave combinations of relatively short duration. These facts suggested a device which distinguishes between the rates of variation of the envelopes of the impressed waves. This is accomplished by a filter in the detector circuit which passes the intermodulated components of speech in the syllabic range, but suppresses those of line noise which are above or below this range.

Figure 5 shows a schematic diagram of the application of this device to a Type B control terminal. The privacy switching circuits are omitted from this drawing, as are also the circuits for delaying the release of the relays. In comparing this drawing with Fig. 2, it will be seen that relays 1, 2 and 3 perform the same functions, but the transmitting branch of the vodas consists of two portions, one a sensitive detector with a syllabic frequency filter, which on operation increases the sensitivity of the second portion.

Considering the action of Fig. 5 on transmitted speech, the output of the sensitive detector of the syllabic device is a complex function of the applied wave having intermodulated components in the range passed by the tuned input circuit, together with a d-c. component and various low frequency components set up by the syllabic nature of the speech. There are also various components of any noise waves which may be present including a d-c. component. The first step in getting rid of the noise is to pass the detector output through a repeating coil which blocks the d-c. component of both the speech and noise, but passes frequencies above about ½ cycle per second. The resulting waves enter the low-pass filter, the output of which contains frequencies between ½ and 25 cycles per second, which "syllabic range" is between the d-c. component of zero frequency and the fundamental frequency of the line noise. These syllabic frequency currents cause momentary operations of relays (*I*) and (*F*). Relay (*I*) operates when a speech wave is commencing and relay (*F*), which is poled oppositely, operates while the impulse is dying out, thus sending current out of the filter in the opposite direction. Operation of either (*I*) or (*F*) effectively inserts gain ahead of the upper detector, thereby

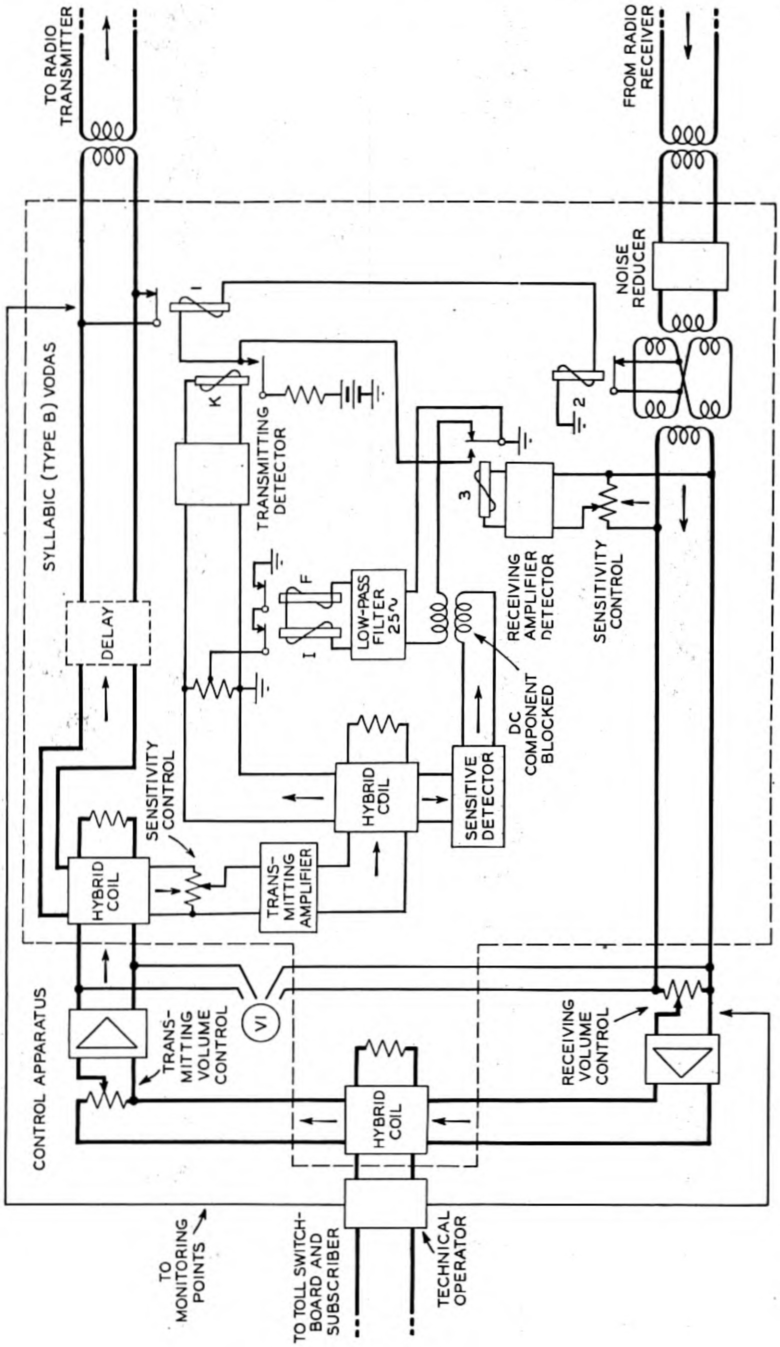


Fig. 5—Schematic of type B control terminal with syllabic vodas.

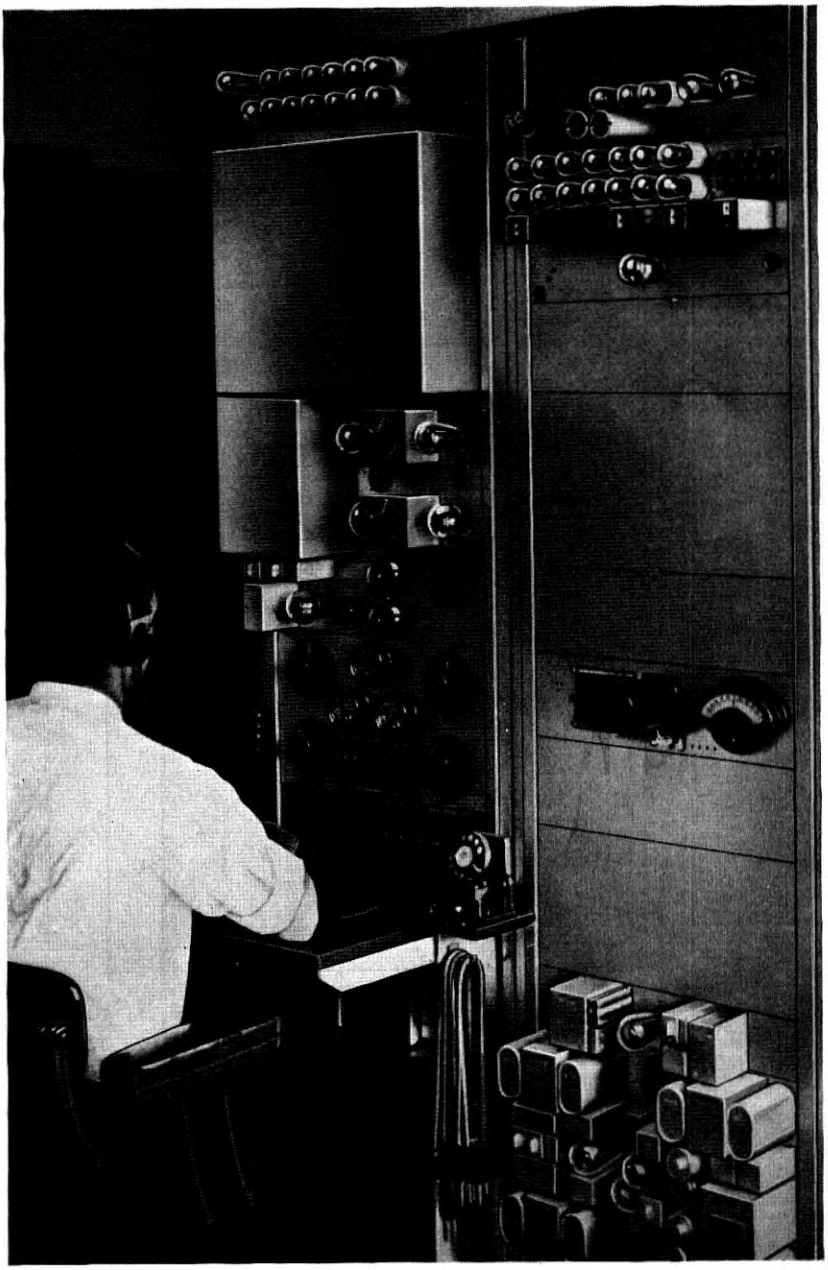


Fig. 6—Technical operator at Forked River, N. J., using a type B control terminal to establish a circuit between a steamship and a shore telephone operator.

increasing the sensitivity of relay (K), when speech is present. Even if the noise is strong enough to operate relay (K) over the upper branch when the gain is inserted, the release of relay (F) at the end of a speech sound will remove the gain and permit (K) to fall back. Thus, it is possible to work relay (K) more sensitively on weak speech than would be possible without the syllabic device.

Figure 6 shows a photograph of a B-type terminal in ship-to-shore service at Forked River, New Jersey. The vodas and volume control apparatus are in the left-hand cabinet. The right-hand cabinet contains privacy apparatus, a signaling oscillator and a vodas relay test panel.

Performance

In any system employing voice-operated devices it is necessary for the time actions to provide for to-and-fro conversation with a minimum of difficulty when the subscribers desire to reverse the direction. The electromagnetic relays used in the vodas have advantages over other types of switching arrangements which have been proposed in that they (1) operate and release at definite current values, (2) have fast operating and constant releasing times, (3) have their windings and their contacts electrically separated, thus simplifying the circuits, and (4) operate in circuits having low impedances.

The operating times of the two types of vodas are shown in Fig. 7 as a function of the strength of suddenly-applied single-frequency sine waves in the voice range. These measurements were made with a capacitance bridge.[5] The sensitivities of the two types were adjusted so that observers noted an equivalent amount of clipping. The Type A vodas was provided with a 20-millisecond delay circuit; the Type B had no delay. For the Type A vodas, the operating time is quite small and constant just above the threshold of operation.

For weak inputs the operating time of the syllabic device is determined by relay (I) and the filter, as shown in Fig. 7. As the suddenly-applied input is increased, a point is reached where the less sensitive detector operates relay (K), reducing the operating time from around 20 milliseconds to values comparable to those of the Type A.

The operation was also tested on waves formed by applying simultaneously two sine waves of equal amplitude but slightly different frequencies. These waves were recorded on an oscillograph, together with a d-c. indication of the operation of each of the vodas relays, with the sensitivities adjusted the same as for Fig. 7. The time from the beginning of a beat wave (null point) to the time of operation was measured from these oscillograms and plotted against various values of total applied voltages. Figure 8 shows the results for a 5-cycle-per-second difference between the two frequencies. Negative values of time indicate that the path was cleared before the beginning of the wave, and these occur only with the Type A vodas due to the delay circuit. The curves for frequency differences of less than 5 cycles per second show more clipping and greater differences between the devices, while those for greater frequency differences show less time clipped and less difference between the two types of vodas. In the case of weak waves it is evident that the syllabic will give less clipping because the energy of the wave does not rise to the value required to operate the Type A device until after the syllabic device has operated; and for very weak waves the Type A does not operate at all. In the case of strong waves, the Type A vodas is better due to its delay circuit. However, since the clipped time is greater on weak sounds than on strong ones, the two types give performances on speech which are judged to be equivalent.

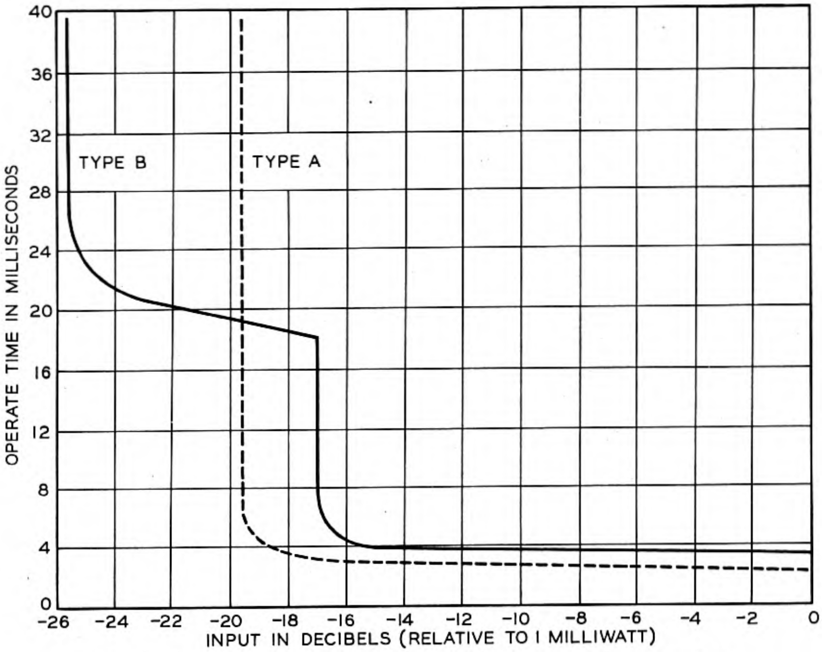


Fig. 7—Vodas operating times with sine waves suddenly applied.

A comparison of operation of the two types of vodas on a speech wave is shown in Fig. 9. Reading from left to right, the middle trace of this oscillogram shows the wave recorded by saying the word "six" over a telephone circuit transmitting a band of frequencies from about 800 to 2000 cycles per second, which is the range normally effective in operating the vodas. The upper trace shows the point at

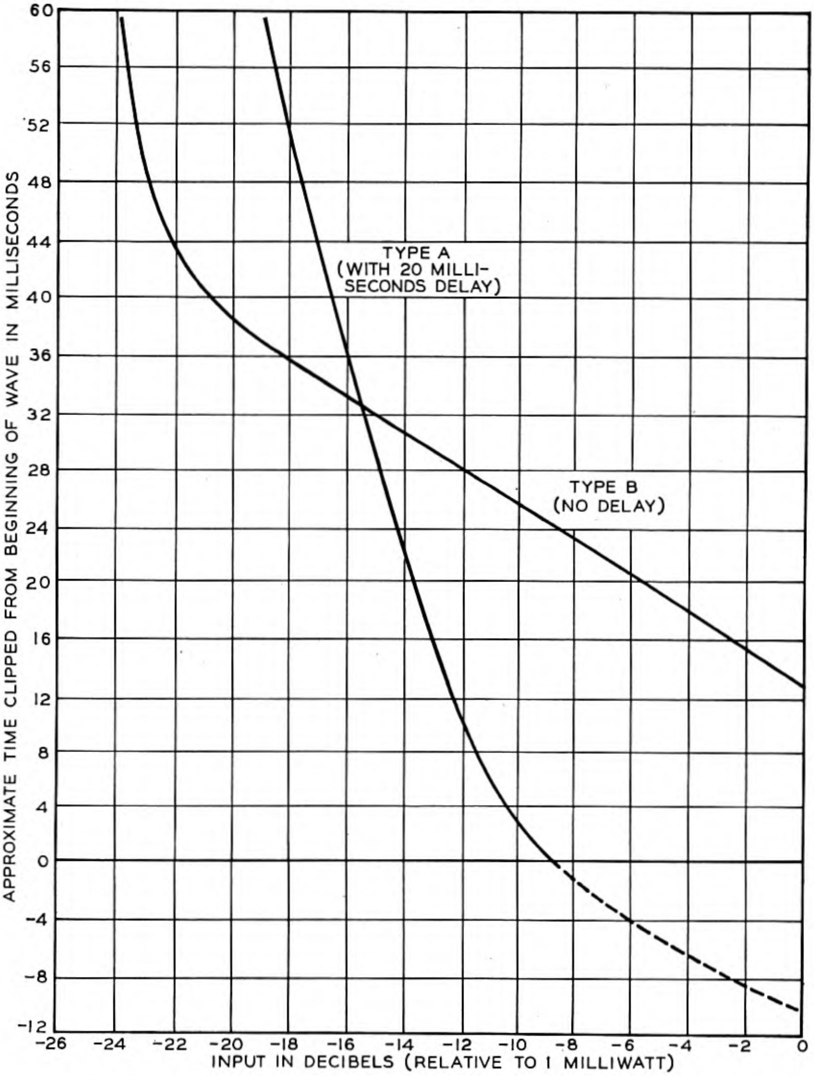


Fig. 8—Operation on a 5-C.P.S. sine wave.

which the syllabic Type B device operated and the lower trace shows the point at which the Type A device operated. Since the speech wave shown was used to operate both devices, the reduction of clipping by the delay circuit in the Type A vodas was not recorded. However, the effect of a transmission delay of 20 milliseconds is shown by subtracting 20 milliseconds from the point at which operation occurred. This is indicated on the oscillogram for both devices. It is concluded that on this wave the syllabic device without a delay circuit would give about the same clipping as the Type A vodas with its dèlay circuit. Figure 8 indicates that the Type A would be better for stronger speech and the Type B would be better for weaker speech. The advantage of a delay circuit in either case is evident.

It is evident from this analysis that the reason for using delay circuits is not primarily because the relays are slow in operating. When the sensitivity is limited by noise, clipping of initial consonants can occur with infinitesimal operating times. One way of reducing the clipping is to use long releasing times so that the relays remain operated between syllables. This has the disadvantage of making it harder for the opposite talker to break in. To avoid this difficulty, the relays in the vodas are given releasing times that permit the distant speech to break in about one sixth of a second after a United States talker ceases to speak.

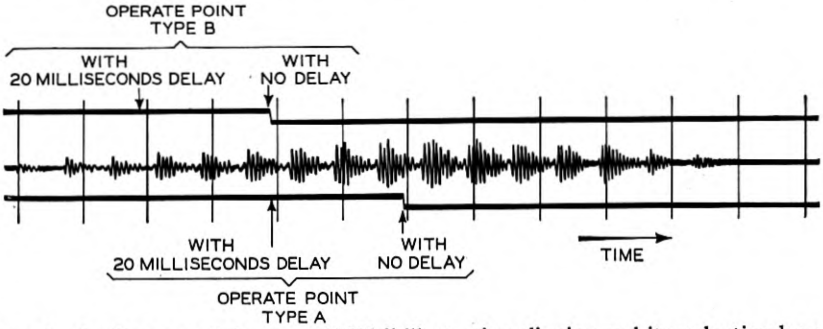


Fig. 9—Oscillogram of the word "SIX," illustrating clipping and its reduction by a delay circuit in the transmission path.

One advantage of delay circuits is to reduce the clipping of initial consonants and thus permit using short releasing times, thereby making it possible to reverse the circuit more readily. In addition, delay circuits permit using a lower relay sensitivity which has two advantages. First, more noise can be tolerated without causing false operation. Second, more received volume can be delivered without the echoes causing false operation of the normally blocked transmitting side.

The advantage of artificial delay of various amounts has been determined by using different types of normally blocked arrangements to find the relation between the delay and the sensitivity required to produce given amounts of clipping of initial sounds. The results are shown for a Type A vodas in Fig. 10. The curves for the syllabic device are similar. The set-up was arranged so that various delays could be inserted in either the transmission circuit (Delay *X*) or the relay circuit (Delay *Y*). The left ends of the curves indicate that when delay *Y* is used, that is, when the net operating time of the relay is great, a point will be reached where no reasonable increase in sensitivity is sufficient to prevent intolerable clipping. The value of 20 milliseconds of delay *X* as compared to zero is equivalent to an increase of about 5 db in sensitivity for a given amount of noticeable clipping.

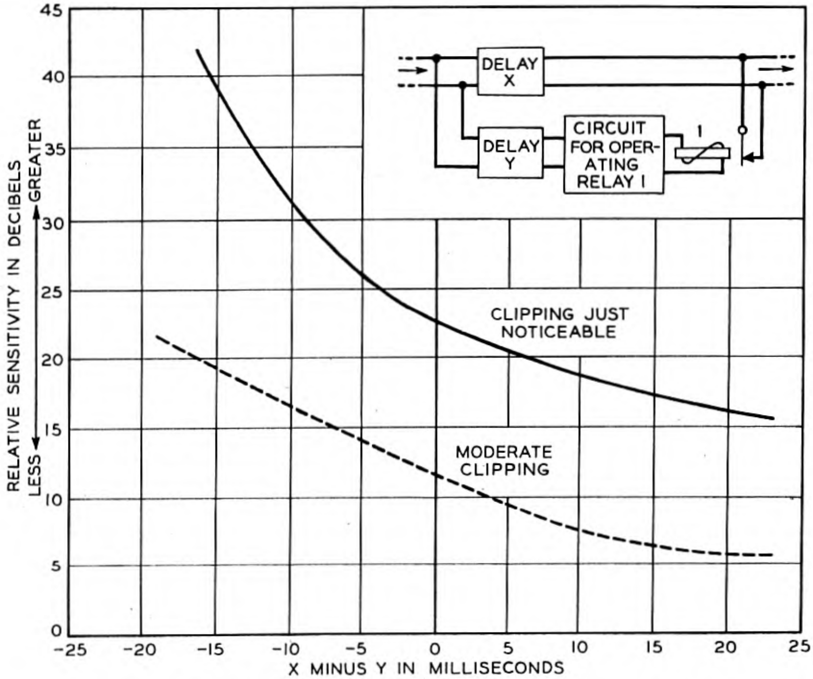


Fig. 10—Typical delay vs. sensitivity for certain clipping effects.

A reasonable release time is of value in preventing clipping, as it causes the relays to remain operated not only for trailing weak endings of sounds, but also when the energy is temporarily reduced by intermediate consonants which may be comparable with noise. Delayed release is also important when it is required to maintain the blocked condition while delayed echoes are being dissipated. For these echoes, the hangover or release times should be constant for various applied voltages. In the vodas, the change in release time over a wide range of inputs is less than 1 per cent. Adjustments are made by varying the condensers and resistances of the auxiliary circuits shown in Fig. 3. Typical values obtained by this method are indicated in Fig. 11.

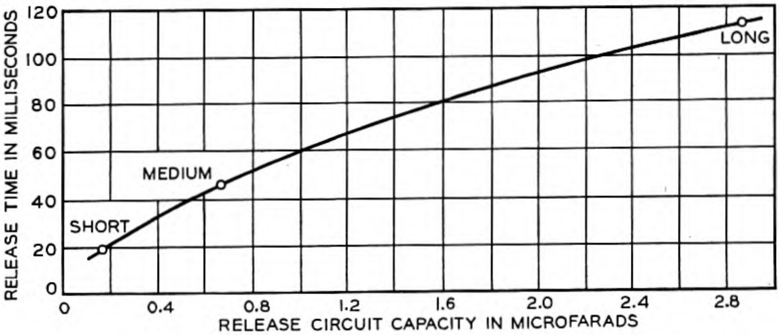


Fig. 11—Release time vs. capacitance.

The vodas amplifier-detectors have broadly tuned input circuits to exclude by frequency discrimination many of the frequencies induced by power sources and those which are unnecessary for speech operation. The sensitivity-frequency characteristic is shown on Fig. 12.

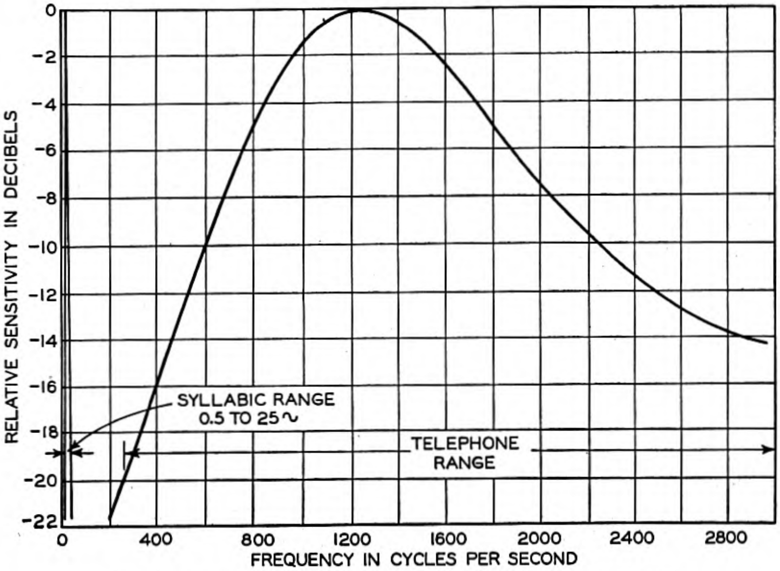


Fig. 12—Sensitivity-frequency characteristics of the vodas.

This figure also shows the relatively narrow frequency range passed by the repeating coil and syllabic frequency filter of the Type B vodas.

Operating Attendance

To insure proper operation of a vodas a technical operator[3] is in attendance. He is provided with circuits which enable him to talk and monitor on the circuit as indicated in Figs. 2 and 5. His duties include adjusting the sensitivity of the receiving relays for the particular value of radio noise existing and adjusting the transmitting and receiving speech volumes by the aid of potentiometers and volume indicators. He selects the proper hangover time and coordinates the operation of the circuit as a whole with the distant end. At times, he may be required to increase the sensitivity of the transmitting side of the vodas in the case of talkers with poor ability to operate relays or to decrease the sensitivity when weak volumes are supplied from land lines with more than the usual amount of noise.

Summary

The vodas is used in radio telephony to switch the voice paths rapidly to and fro, and thus prevent echoes and singing that would otherwise occur at unpredictable times. It is also used to save privacy apparatus by permitting the use of the same apparatus for both directions of transmission. The performance characteristics of the electromagnetic relays used in the vodas are very suitable in that they have small operating and constant releasing times.

Improved performance of the voice-operated relays in the presence of line noise can be secured by the use of a syllabic type of vodas which discriminates between the characteristic voltage-time envelopes of the noise and speech waves. Laboratory and field tests indicate that this device, even without delay circuits, gives slightly better performance on most conditions than the original vodas with delay. When provided with a transmitting delay circuit, the syllabic device is decidedly better than the older vodas.

References

The International Bibliography on the Coordination of Radio Telephony and Wire Telephony is given in the C.C.I.F. Green Book Volume I of the Proceedings of the Xth Plenary Meeting, held at Budapest, September 1934. Below is a chronological list of Bell System papers relating to the vodas.

1. "The Limitation of the Gain of Two-Way Telephone Repeaters by Impedance Irregularities," George Crisson, *Bell Sys. Tech. Jour.*, Vol. 4, No. 1, January, 1925, pp. 15–25.

2. "Echo Suppressors for Long Telephone Circuits," A. B. Clark and R. C. Mathes, *A.I.E.E., Jour.*, Vol. 44, No. 6, June, 1925, pp. 618–626; *Elec. Commun.*, Vol. 4, No. 1, July, 1925, pp. 40–50; *A.I.E.E., Trans.*, Vol. 44, 1925, pp. 481–490.
3. "The New York-London Telephone Circuit," S. B. Wright and H. C. Silent, *Bell Sys. Tech. Jour.*, Vol. 6, No. 4, October, 1927, pp. 736–749.
4. "Echo Elimination in Transatlantic Service," G. C. Crawford, *Bell Lab. Record*, Vol. 5, No. 3, November, 1927, pp. 80–84.
5. "Bridge for Measuring Small Time Intervals," J. Herman, *Bell Sys. Tech. Jour.*, Vol. 7, No. 2, April, 1928, pp. 343–349.
6. "Problems in Power-Line Carrier Telephony," W. B. Wolfe and J. D. Sarros, *A.I.E.E., Jour.*, Vol. 47, No. 10, October, 1928, pp. 727–731; *A.I.E.E., Trans.*, Vol. 48, 1929, pp. 107–113.
7. "A Carrier Telephone System for Power Lines," C. F. Boeck, *Bell Lab. Record*, Vol. 7, No. 11, July, 1929, pp. 451–458.
8. "Voice-Frequency Equipment for the Transatlantic Radio Telephone," J. A. Coy, *Bell Lab. Record*, Vol. 8, No. 1, September, 1929, pp. 15–20.
9. "Effects of Phase Distortion on Telephone Quality," J. C. Steinberg, *Bell Sys. Tech. Jour.*, Vol. 9, No. 3, July, 1930, pp. 550–556.
10. "Electrical Delay Circuits for Radio Telephony," R. T. Holcomb, *Bell Lab. Record*, Vol. 9, No. 5, January, 1931, pp. 229–232.
11. "Acoustic Delay Circuits," W. P. Mason, *Bell Lab. Record*, Vol. 9, No. 9, May, 1931, pp. 430–432.
12. "The Time Factor in Telephone Transmission," O. B. Blackwell, *Bell Sys. Tech. Jour.*, Vol. 11, No. 1, January, 1932, pp. 53–66; *Bell Lab. Record*, Vol. 10, No. 5, January, 1932, pp. 138–143; *A.I.E.E., Trans.*, Vol. 51, 1932, pp. 141–147. *Elec. Engg.*, Vol. 50, No. 10, October, 1931, pp. 902–903 (Abstract).
13. "Two-Way Radio Telephone Circuits," S. B. Wright and D. Mitchell, *Bell Sys. Tech. Jour.*, Vol. 11, No. 3, July, 1932, pp. 368–382; *I.R.E., Proc.*, Vol. 20, No. 7, July, 1932, pp. 1117–1130.
14. "A Telephone System for Harbor Craft," W. K. St. Clair, *Bell Lab. Record*, Vol. 11, No. 3, November, 1932, pp. 62–66.
15. "Voice Frequency Control Terminals for Caribbean Radio Systems," W. A. MacMaster, *Bell Lab. Record*, Vol. 11, No. 12, August, 1933, pp. 369–374.
16. "Certain Factors Limiting the Volume Efficiency of Repeatered Telephone Circuits," L. G. Abraham, *Bell Sys. Tech. Jour.*, Vol. 12, No. 4, October, 1933, pp. 517–532.
17. "The 'Compandor'—An Aid Against Static in Radio Telephony," R. C. Mathes and S. B. Wright, *Elec. Engg.*, Vol. 53, No. 6, June, 1934, pp. 860–866; *Bell Sys. Tech. Jour.*, Vol. 13, No. 3, July, 1934, pp. 315–332.
18. "The Voice-Operated Compandor," N. C. Norman, *Bell Lab. Record.*, Vol. 13, No. 4, December, 1934, pp. 98–103.
19. "Ship-to-Shore Radio in Puget Sound Area," E. B. Hansen, *Elec. Engg.*, Vol. 54, No. 8, August, 1935, pp. 828–831.
20. "Marine Radio Telephone Service," F. A. Gifford and R. B. Meader, *Commun. & Broadcast Engg.*, Vol. 2, No. 10, October, 1935, pp. 9–11, 15. Tech. Digest in *Bell Sys. Tech. Jour.*, Vol. 14, No. 4, October, 1935, pp. 702–707.
21. "A Noise Reducer for Radio Telephone Circuits," N. C. Norman, *Bell Lab. Record*, Vol. 15, No. 9, May, 1937, pp. 281–285.
22. "Radio Telephone Noise Reduction by Voice Control at Receiver," C. C. Taylor, this issue of *Bell Sys. Tech. Jour.; Elec. Engg.*, August, 1937.

Radio Telephone Noise Reduction by Voice Control at Receiver *

By C. C. TAYLOR

In listening to speech transmitted over radio circuits, the noise arriving in the intervals between the signals may be annoying. There is also evidence that the intelligibility is reduced due to this noise shifting the sensitivity of the ear. Reducing the noise occurring in the intervals of no speech should therefore improve reception.

This paper gives the underlying requirements for a device to accomplish this type of noise reduction and describes the action of a typical "noise reducer." Laboratory and field tests are described which show that its use is equivalent to an improvement in signal-to-noise ratio which reaches a maximum value of about 5 db. It also reduces false operation of the voice-operated relays used on long radio telephone connections.

Introduction

IN transmitting speech over radio telephone circuits there are a number of conventional methods of increasing the signal with respect to the noise. Examples of such methods are the use of higher power, directive antennas, diversity reception and filters to narrow the received frequency band. In addition, there are other methods of a special character which reduce the effect of the noise interference with the speech transmission. One example of such a device limits the noise interference by eliminating the high peaks of noise of very short duration and depending upon the persistence of sensation of speech in the ear to bridge the gaps. Another method diminishes the noise in intervals of no speech. This is the method which will be discussed here.

Speech and Noise Considerations

Speech signals may be represented by a group or band of frequencies occupying a certain interval of time. In using the conventional method of narrowing the received frequency band, filters eliminate all noise outside the band actually required. In fact we sometimes go beyond this and remove some of the outer frequency components of

* Presented at the Pacific Coast Convention of A. I. E. E., Spokane, Washington, September 2, 1937. Published in *Elec. Engg.*, August, 1937.

speech which are weak and submerged in the noise and therefore contribute little or nothing to the intelligibility. Experiments have shown the effect on voice transmission of removing portions of the frequency range.[1] Articulation tests were used to afford a quantitative measure of the recognizability of received speech sounds. These show that the upper frequencies may be cut off down to about 3000 cycles without serious reduction in articulation. After such treatment, as the noise level increases, the weaker and less articulate sounds become more and more submerged in the noise and additional reduction in the detrimental effect of the noise is required.

In addition to the speech waves covering a frequency band they occupy intervals of time. The unoccupied intervals between the speech sounds contain noise. Reduction of the noise reaching the ear in these intervals has been found to result, under certain conditions, in an improvement in speech reception. This may possibly be explained by considering the characteristics of the ear.[1] It has been shown that noise present at the ear has the effect of shifting the threshold for hearing other sounds or has a deafening effect. That is, there is a reduction of the capacity of the ear to sense sounds in the presence of noise. For example, if a person has been listening to a noise for a certain period, his ear is made insensitive so that speech signals following are not so easily distinguished. The ear has a sensory build-up time, that is, a time needed for the noise to build up to a steady loudness. By reducing the noise in the intervals of no speech the average threshold shift seems to be diminished. Aside from this the presence of the noise tends to distract the attention from the perception of the speech. Removal of noise during the intervals of no speech tends to reduce this effect.

Requirements

In considering the elimination of the noise during these intervals it is necessary to bear in mind certain characteristics of speech.[2] Speech waves may be regarded as nonperiodic in that they start at some time, take on some finite values and then approximate zero again. In connected speech it is usually possible to approximately distinguish between sounds and to ascribe to each an initial period of growth, an intermediate period which in some cases approximates a steady state and then a final period of decay. The duration intervals of various syllabic sounds vary from about .03 to as much as .3 or .35 second. When noise is high the weaker initial and final sounds become obscured so that they contribute little to the intelligibility.

[1] See end of paper for references.

In connected speech, silent intervals occupy about one-fifth to one-third of the total time. Also there are frequent intervals when the sounds are rather weak. However, if we attempt to suppress noise during all these intervals, experience shows that the suppression becomes too obvious, and the speech is apt to sound mutilated. For this reason the function of any device to be used for reduction of noise in the intervals between speech is to operate rather quickly to remove suppression and pass the speech and approximately to sustain this condition for sufficient periods to override weaker intervals so that obvious speech distortion does not occur.

To reduce the noise in the intervals between speech it is necessary to depend for control upon either the speech itself or upon some auxiliary signal usually under the control of the speech at some point in the circuit where the signal-to-noise ratio is better. This latter condition is illustrated on a circuit where the carrier is transmitted only during speech intervals. The carrier then acts as an auxiliary signal which operates a device at the receiver to remove loss.[3, 4] The device to be discussed below utilizes the speech itself at the receiver to perform this function.

In using the speech in this way it is obvious that control can be accomplished only when the speech energy sufficiently exceeds the noise energy so that the presence of the speech is distinguishable. The device could operate abruptly as, for example, a relay which removes a fixed loss in the operated position and restores it when non-operated. Experience indicates that the use of such a device makes the suppression too obvious if it is to follow the speech sounds closely. It is desirable, then, to perform this reduction by more or less gradually removing loss as the speech increases to accentuate the difference between levels of speech sounds and levels of noise which occur in the gaps between speech.

Noise Reducer

This kind of performance has been secured in a device known as a noise reducer. A comparison of the action of the noise reducer and a relay having similar maximum loss is shown in Fig. 1. This figure shows the input-output characteristics of these devices over the voice amplitude range to which they are subjected on a radio circuit. The noise reducer may be likened to a relay with a variable loss, the loss not varying instantaneously but over a short period of time. The loss, for any short period, may be any value within the loss range and the device has, therefore, been likened to an elastic or shock absorbing relay.

The noise reducer has no loss for strong inputs, considerable loss for weak inputs and changes this loss gradually over a short interval of time. It introduces loss in the absence of speech but reduces this loss in proportion to the amplitude and duration of waves impressed upon it. The time required for the loss change is such that abruptness of noise change is absent and very short impulses of static do not effectively control the loss. This contrasts with a very fast limiter acting on high-peak crashes only.

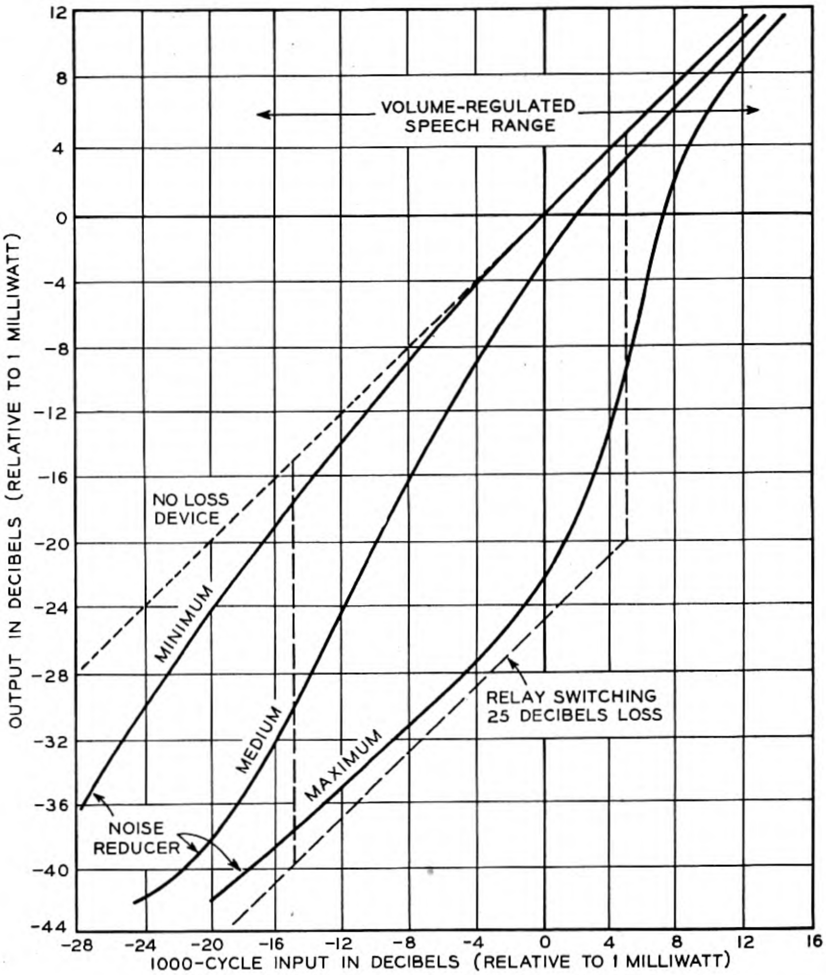


Fig. 1—Input-output comparison of noise reducer and voice-operated relay.

The noise may control the loss if its average amplitude is strong enough. Therefore, the control is made adjustable so that the noise waves are not permitted to control for any noise condition within the range of usefulness of this device. Thus the noise in the absence of speech is always reduced and the portions of the initial and decay periods of the speech sounds which are also reduced vary with this adjustment for noise intensity. Of course, if the speech-to-noise ratio becomes too small or if other transmission conditions interfere, an improvement becomes impossible.

Circuit Arrangement

Figure 2 shows the circuit of the noise reducer in simplified schematic form.[5] Incoming waves pass from left to right through the fixed pad, the vario-losser and the amplifier to the output. At the input, part of these waves pass through the reduction control branch circuit which includes a variable resistor, an amplifier and a rectifier. The direct current produced by the rectifier is applied through the condenser and resistance filter to the copper-oxide losser circuit. For current below a threshold value, no appreciable change occurs in the losser and the loss introduced is about 20 db. As input increases, rectified current reaches a value where the loss begins to change rapidly. It becomes 0 db at an input about 20 db above the point at which the loss starts to change. The design is such that the loss remains substantially constant for higher inputs.

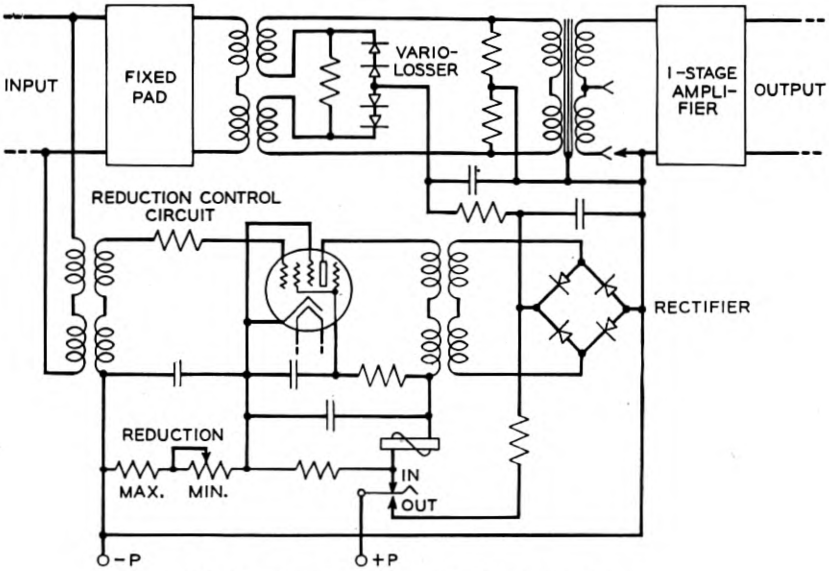


Fig. 2—Simplified schematic of noise reducer.

The vario-losser makes use of the resistance variation with current of copper-oxide rectifier disks. This variable resistance shunts a fixed resistance in series with the windings of a repeating coil as shown in Fig. 2. The maximum loss is determined by the fixed resistance when small current is flowing through the disks while the varying loss is determined by the shunting copper-oxide resistance which decreases rapidly with increasing current above a threshold value until a low value is reached. The minimum loss is limited by the output of the control tube approaching a maximum and the shunting resistance becoming so small that additional decrease affects the loss inappreciably.

The variable resistor setting in the reduction control circuit determines the input amplitude at which reduction begins and therefore the point above which the loss remains substantially constant. If there is a difference in amplitude between speech and noise, the reduction control may be so adjusted that the noise on the circuit, when no speech is present, is appreciably reduced. The action then is as follows: In the absence of speech, noise is reduced usually the maximum value of 20 db; during intervals of lower speech amplitudes the loss decreases in proportion to the increase in amplitude, and during speech of high amplitude both noise and speech are transmitted without loss. As the noise encroaches upon the range of speech amplitude, it becomes necessary to reduce greater amplitudes, thereby also further reducing the weaker parts of speech.

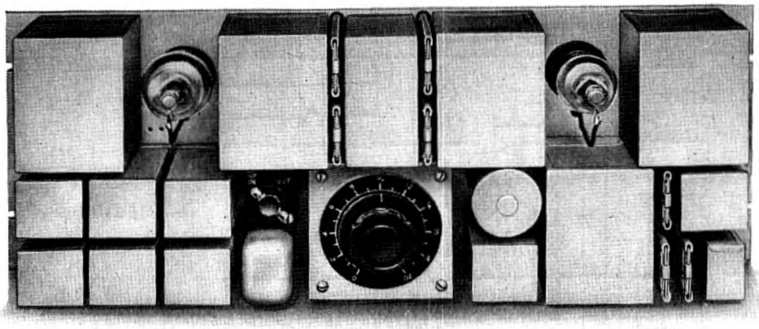


Fig. 3—noise reducer panel.

The noise reducer is contained on a 7¼ inch panel for relay rack mounting. Figure 3 gives a front view. The panel contains the reduction control resistor and an In-Out key which, in the Out position, gives the device a fixed loss. Both resistor and key may be duplicated external to the panel with the wiring arranged to give remote control.

Characteristics

Figure 4 gives the 1000-cycle input-loss characteristic for three settings of the reduction control. For any setting, there is an input volume above which the loss remains constant, while for volumes below this the loss increases with decreasing input until the maximum loss is reached. The volume regulated speech range encountered on radio circuits at some point in the circuit which is 5 db above reference volume as measured on a volume indicator is indicated as extending from + 13 db to − 17 db referred to 1 milliwatt for the purpose of showing approximate corresponding speech amplitudes.

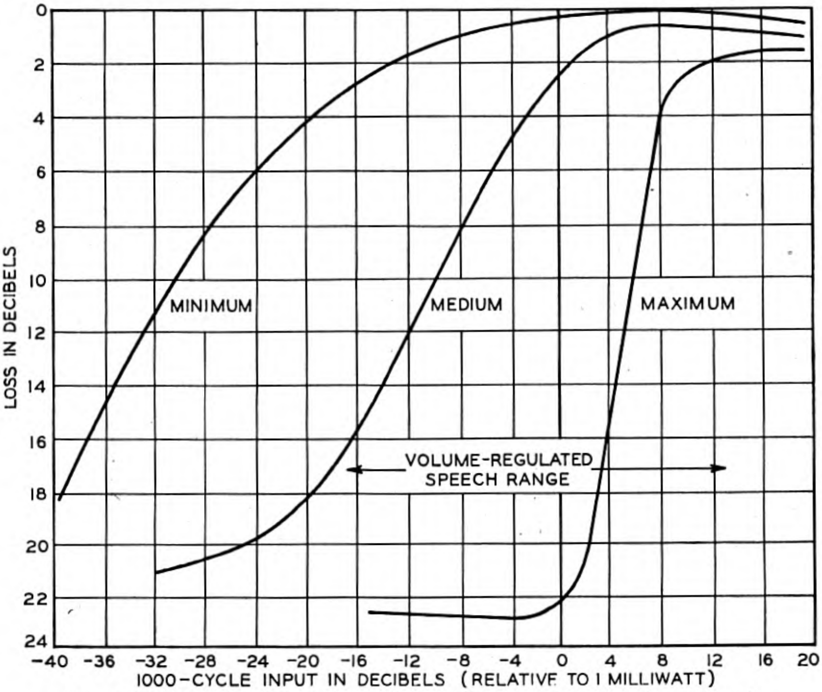


Fig. 4—Loss versus input for several settings of the reduction control.

Figure 5 shows oscillograms giving the input and output characteristics of noise for maximum reduction and of speech for maximum, medium and minimum reduction. The upper trace is the input and the lower trace the output. The middle trace is not used. It will be noted by inspecting the In and Out traces at the beginning and ending of the word "bark" that there is some distortion in speech for the maximum reduction condition, but very little distortion for minimum reduction. Maximum reduction would be used only in case of high noise where this distortion is less objectionable than the noise.

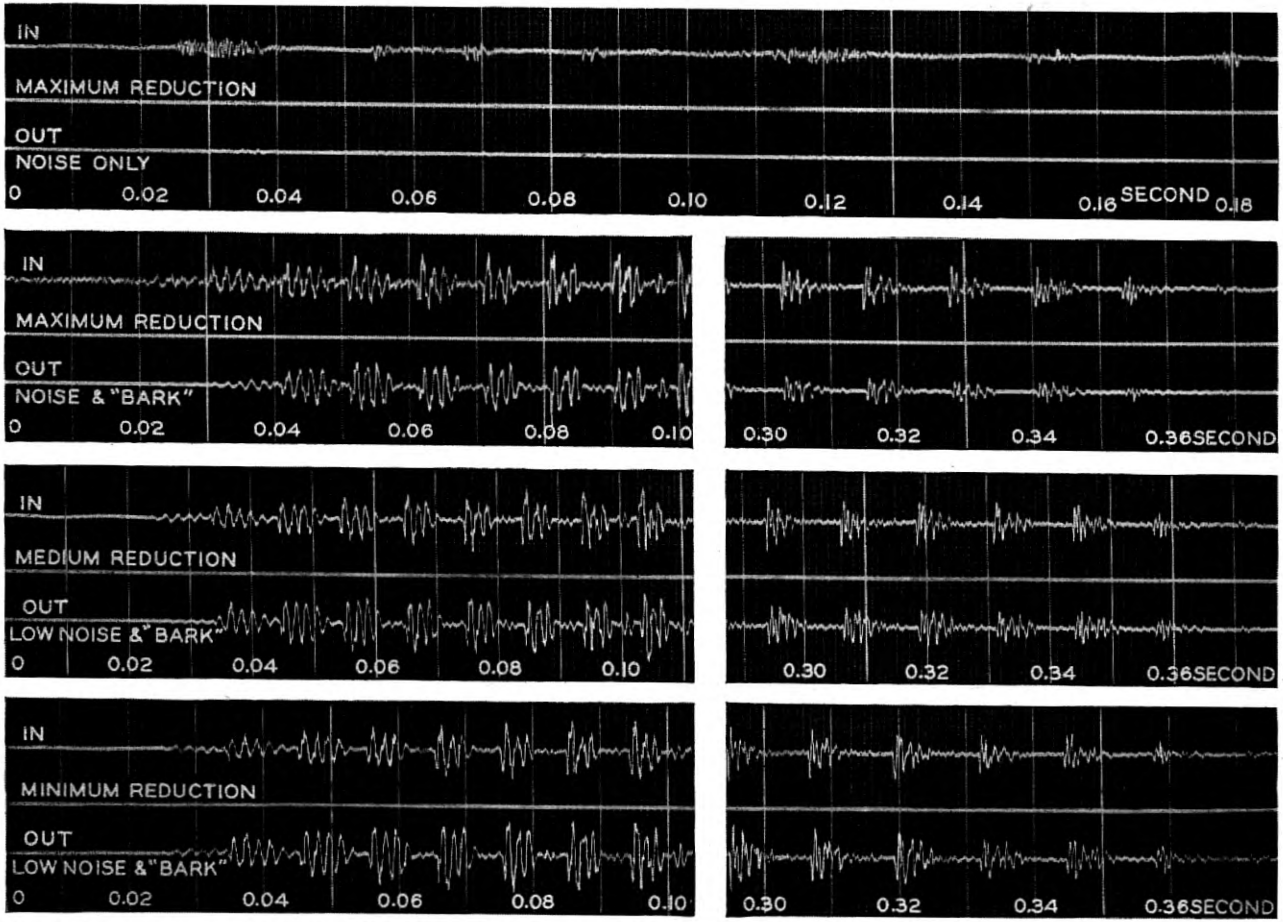


Fig. 5—Input and output for: (1) high noise with maximum reduction; (2) high noise with the beginning and ending of the word "bark"—maximum reduction; (3) low noise, the word "Bark"—medium reduction; (4) low noise, the word "bark"—minimum reduction.

PERFORMANCE

Laboratory tests have been made in an attempt to evaluate the advantages to be gained by the use of the noise reducer. It was shown that, for the rather limited and controlled conditions which were tested, definite advantage can be observed in judgment tests of the effectiveness of speech transmission through noise with and without the noise reducer. This advantage is of the order of magnitude of 3 to 5 db at the border line between commercial and uncommercial conditions on the noisy circuit.

This figure is in approximate agreement with results obtained from records of performance on commercial connections. A curve is available which shows the approximate relation between percentage lost circuit time and transmission improvement for a long-range short-wave radio telephone circuit.[6] From the records of lost circuit time as affected by the noise reducer use, an improvement of 4 db is obtained from this curve.

Observations were made and records kept for twelve months of the use of the device at the land terminal of the high seas ship-to-shore circuit and for shorter periods on New York-London circuits. These observations indicate that the noise reducer most satisfactorily reduces objectionable effects where the interference consists of noise of a fairly steady character. As might be expected it is somewhat less effective on crashy static. If the noise is very low there is no improvement; as the noise increases the benefit increases up to a certain point; when the noise amplitudes begin to approach too closely the peak amplitudes of the voice waves it becomes impossible to distinguish between them without producing objectionable speech distortion and there is again no advantage. Where volume fading is present there is a tendency to accentuate the volume changes and it becomes necessary to adjust the reduction control to limit this. Otherwise this effect may offset the possible noise improvement. The operating practice is to adjust the reducer control circuit for each noise or transmission condition so that optimum reception as judged by the technical operator is obtained. The general rule is to use the minimum reduction possible.

USE OF NOISE REDUCER WITH VOICE SWITCHED CIRCUITS

On radio telephone circuits for connection to the land telephone system, control terminal equipment is used at the junction of the land lines and the two one-way radio channels (one transmitting the other receiving) necessary for two-way communication. In making this connection a widely used method is one in which the two-wire land circuit is normally connected to the receiving radio channel and is

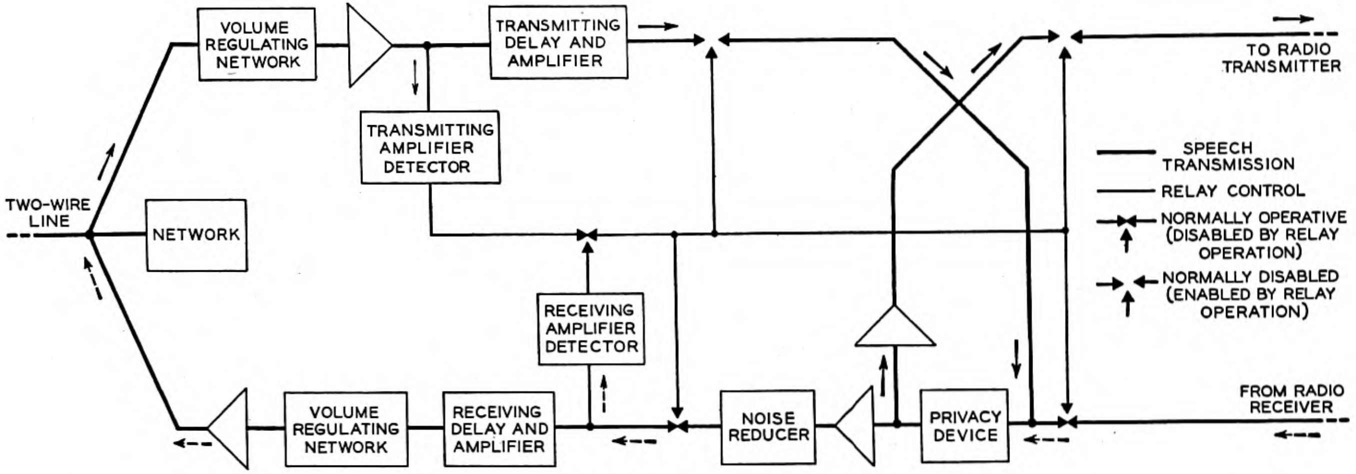


Fig. 6—Application of noise reducer to radio control terminal.

switched to the transmitting channel when the land subscriber talks. This switching is done by voice-operated relays.[7, 8] The noise reducer in addition to improving the intelligibility of the speech received protects these voice-operated relays against false operation by the received noise.

Figure 6 shows the application of the noise reducer to such a control terminal. Speech entering the terminal from the left goes through the upper branch of the circuit, with volume regulating means and privacy apparatus, to the radio transmitter. Speech received from the distant terminal enters at the lower right from the radio receiver and proceeds through the privacy apparatus, the noise reducer, receiving regulating network and amplifier to the two-wire line. Outgoing speech operates the transmitting path and disables the receiving path. Incoming speech operates the receiving amplifier detector, which disables the transmitting amplifier detector, thus preventing singing and reradiation of received waves.

Without the noise reducer the receiving relay may be operated by noise in the receiving path and such operation to an excessive extent will interfere with outgoing speech. To avoid this effect, it is customary to reduce its sensitivity so that noise may not operate it. This results in the weaker speech parts also failing to operate the receiving relay. This weak speech and noise returned to the transmitting path through the land line connection may be strong enough to operate the transmitting relays and thus cut off incoming speech. This is avoided by reducing the volume to the land line. Therefore, any device which reduces noise in the receiving path in the absence of speech effects an improvement not only in the switching operation but also in the received volume. By placing the noise reducer in the receiving path false operation is diminished and volume increases of 5 to 15 db are realized. The noise reducer is applied to the receiving side of the terminal beyond the privacy apparatus so that it does not introduce any distortion in the privacy portion of the circuit. It is placed ahead of the receiving amplifier detector, thereby reducing noise between words which might affect the operation of this relay apparatus.

Summary

The noise reducer, which is a voice controlled variolosser with limited and controllable action, has been provided for use on short-wave radio telephone circuits and has proved to be a valuable and relatively inexpensive means of securing noise reduction. Improved reception is obtained for many of the transmission conditions experienced on such circuits. This results in better intelligibility to the subscriber, greater margin in the operation of two-way radio telephone circuits and a reduction of difficulties in the wire plant caused by connection to noisy radio circuits.

References

1. "Speech and Hearing," H. Fletcher, D. Van Nostrand Co., 1929.
2. "Effects of Phase Distortion on Telephone Quality," J. C. Steinberg, *Bell Sys. Tech. Jour.*, July, 1930.
3. "Ship-to-Shore Radio in Puget Sound Area," E. B. Hansen, *Elec. Engg.*, Vol. 24, No. 8, August, 1935.
4. "A Telephone System for Harbor Craft," W. K. St. Clair, *Bell Laboratories Record*, November, 1932.
5. "A Noise Reducer for Radio Telephone Circuits," N. C. Norman, *Bell Laboratories Record*, May, 1937.
6. "The Reliability of Short-Wave Radio Telephone Circuits," R. K. Potter and A. C. Peterson, Jr., *Bell Sys. Tech. Jour.*, July, 1934.
7. "Two-Way Radio Telephone Circuit," S. B. Wright and D. Mitchell, *Bell Sys. Tech. Jour.*, July, 1932.
8. "The Vodas," S. B. Wright, this issue of the *Bell Sys. Tech. Jour.; Elec. Engg.*, August, 1937.

Transmitted Frequency Range for Circuits In Broad-Band Systems

By H. A. AFFEL

IN utilizing the broad frequency ranges which the newer carrier systems can transmit the telephone engineer has a problem of choice in band width per channel to be allotted to speech currents. A sufficient width is vital to faithful speech reproduction; and desire for better telephone service always recommends an increase in band width over past practice. A reasonable balance, however, must obtain between various economic factors; and there must always be considered the relation between a proposed system and the other parts of the telephone plant, and also the trend of the art.

The message band widths and the channel spacing which have been chosen by the Bell System for various new systems are summarized and discussed in this paper. These systems are expected to play a large part in the future growth of its long distance plant; and the reasons underlying this choice may therefore be of general interest.

Different broad band systems are under development: A 12-channel system for use on open-wire lines employing frequencies up to 140,000 cycles, a 12-channel system for use on 19-gauge pairs in existing toll cables using frequencies up to 60,000 cycles, and a coaxial system capable of transmitting frequencies up to a million cycles or more, from which it is proposed to obtain 240 or more channels.

In the different systems noted above, terminal apparatus is employed which has many common features: The different channels are uniformly spaced at 4000-cycle intervals; the same band filters are used in the ultimate channel selecting circuits; and the derived voice circuit band widths are substantially identical for all channels of all systems. The transmission frequency characteristic of a single link of such systems, in accordance with present designs, is shown on Fig. 1. A curve for five similar links connected in tandem is also indicated. Based on a 10 db cutoff as compared with 1000-cycle transmission, a single-link band extends from approximately 150 to 3600 cycles, and a five-link band extends from about 200 to 3300 cycles.

There is, of course, no fixed relationship between the channel spacing and the frequency range of the derived voice-frequency circuit. This is largely a matter of economics in the design of a particular system.

The 4000-cycle channel spacing would permit obtaining a narrower band width with some simplification in the selecting circuits. With further development in selecting circuits, it is believed that it would permit obtaining a somewhat wider band or, if desired, a reduction in the cost of apparatus, maintaining the same band.

The band chosen initially for the new systems is believed to be a desirable and forward-looking step in the direction of improving the quality of speech transmission, a continuing trend which is as old as telephony itself. Figure 2 shows typical band characteristics which mark the progress of transcontinental telephony since 1915. For shorter distances, the band widths have, of course, generally been wider than indicated on this series of curves. In the case of carrier systems the band depends on the number of links. The curve shown for 1937 is for the broad-band systems, estimated on the basis of a three-link connection.

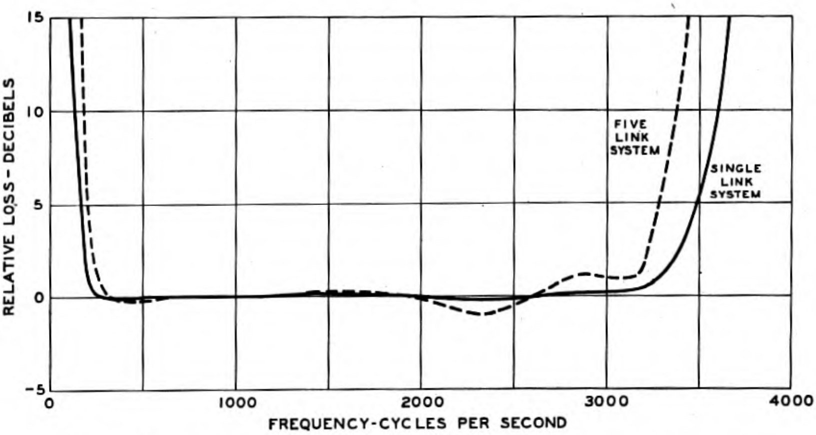


Fig. 1—Transmission frequency characteristics of broad-band systems.

The increase in band width is achieved without material increase in cost, since in situations which favor their use, broad-band systems provide circuits which are substantially more economical than other alternatives, and the improvement can therefore be obtained by giving up only a small portion of the savings which the systems themselves make possible. If, as in some older types of systems, it had been chosen to maintain a standard of 250 to 2750 cycles for a single-link connection in the broad band systems, this could have been accomplished by the use of a channel frequency spacing of about 3000 cycles. The wider transmission band is therefore obtained by a sacrifice in the ratio of approximately 3:4 in the number of channels obtained within a given frequency range. However, this does not mean a 4:3 increase in the cost per circuit. The amount is considerably less than this—depending somewhat on the type of system. In the proposed coaxial system, which appears to be a favorable example, where the attenuation increases roughly as the square root of the frequency, a frequency band increased by one-third means that for repeaters of a given type and amplification the number of repeaters is multiplied by approximately $\sqrt{\frac{4}{3}}$; that is to say, approximately 15 per cent more repeaters are required. Furthermore, the line and terminal apparatus costs are not changed in a case of this kind, and since they constitute a major part of the total cost, the net increase in cost for the wider band width will be considerably less than 15 per cent—about five per cent in the case of the longer systems where the terminal apparatus costs are a small factor, and only a per cent or two in the case of the very short systems where the terminal apparatus costs predominate.

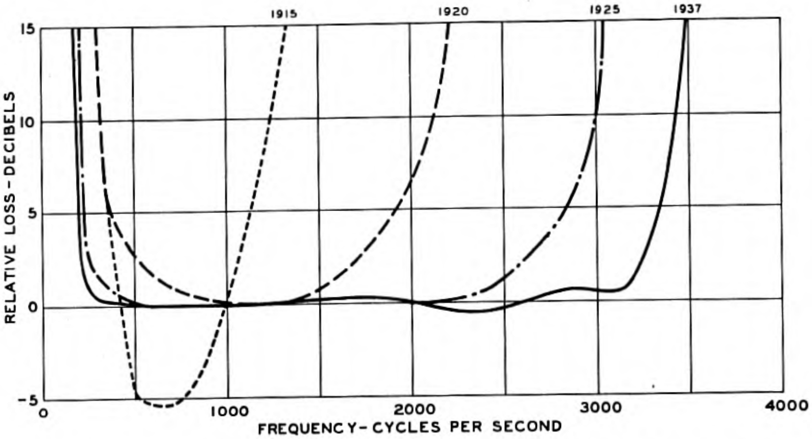


Fig. 2—Representative transmission frequency characteristics of 3000-mile toll circuits.

In the ideal case, using substantially perfect transmitters and receivers, articulation is improved as the upper limit in frequency transmission is raised, as shown in Fig. 3. The increase in transmission performance, which a step from 2750 to 3300 cycles, or 3600 cycles for a single link, makes possible, is evidently still on the part of the band width-articulation relationship where a measurable increase in articulation may be expected. An improvement in band width accordingly reduces the effort needed to interchange ideas, since fewer repetitions occur and attention can be somewhat relaxed. It also enhances the naturalness of the received speech, and so makes the conversation more pleasing as well as easier. It should be noted also that the proposed broad-band systems will transmit frequencies approximately 50 to 100 cycles lower than earlier systems, which, while not contributing appreciably to articulation, has the effect of increasing naturalness.

When applied in the telephone plant, the resultant effect of a given increase in band width will of course depend on the other parts of the circuit, and the transmission characteristics of the transmitters and receivers. Improved transmitters and receivers are now being applied rapidly in the Bell System. They have much better transmission characteristics than earlier types and an effective upper frequency of transmission for the new station set which is well above 3000 cycles, as shown on Fig. 4.

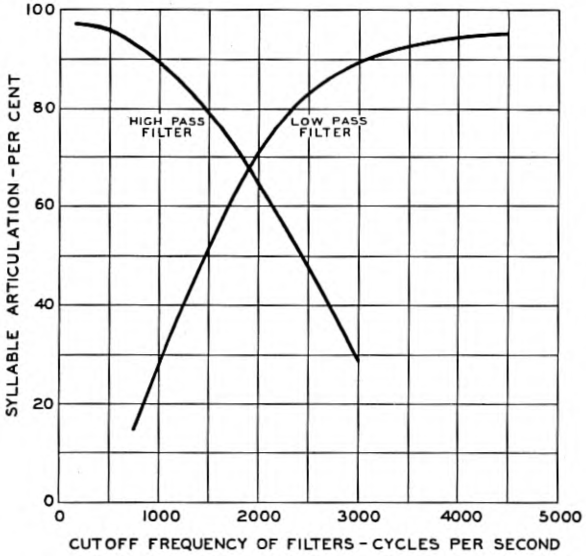


Fig. 3—Effect of cutoff frequency on syllable articulation.

The toll connecting trunks are important links in a typical overall connection, and here also there has been a continued trend to provide wider band circuits. Figure 5 shows the transmission frequency characteristics of representative types of toll connecting trunks which are being commonly installed at present. Both non-loaded and loaded trunks are shown on the figure. Of course, in the non-loaded case, there is no definite cutoff frequency. The curve for the loaded trunk shows a reasonably long trunk having a 5 db loss at 1000 cycles (6.4 miles). In practice, of course, the trunk length may vary from a fraction of a mile to 10 miles or more, with a corresponding effect on the transmission characteristic. It will be noted that the effective

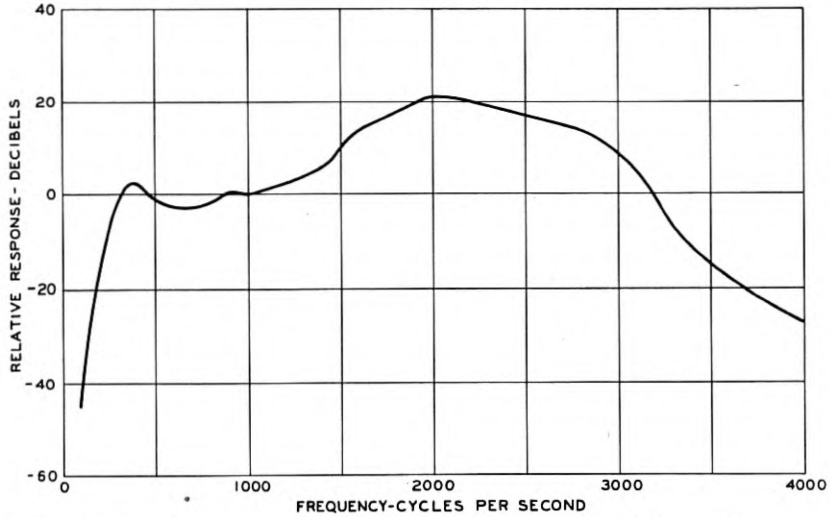


Fig. 4—New station-set characteristics (including two one-mile 24-gauge loops connected by distortionless trunk).

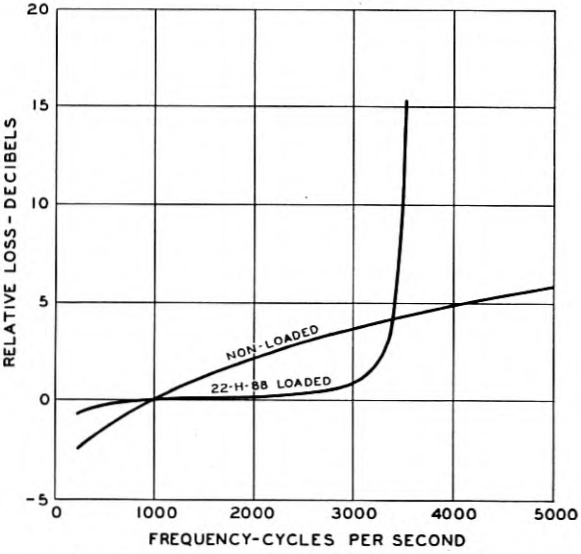


Fig. 5—Toll connecting trunk characteristics.

cutoff of the loaded trunk shown is about 3500 cycles based on a 10 db cutoff point. Other types of loading, which will also be employed, will have still higher cutoff points. Evidently the band widths of the broad-band circuits, toll connecting trunks, and new station sets are well matched.

Laboratory and field tests have been made with circuits simulating the cutoff of the new broad-band systems and using various types of station sets, including the new standard. These indicate that raising the cutoff from 2750 cycles to 3600 cycles is equivalent to making a reduction of 3 to 4 db in the net overall loss of the circuit. Raising the cutoff from 2750 cycles to 3300 cycles is equivalent to a lesser reduction. With older types of instruments which reproduce speech less faithfully, this difference is also less, and of course, with instruments providing transmission up to considerably higher frequencies, the difference is greater.

It will be appreciated, of course, that the wider speech band which will be made available in the new broad-band systems will not be fully effective in all telephone connections unless other toll circuits and toll connecting trunks and station sets are provided with improved transmission frequency characteristics. From a practical standpoint it is obvious that in a large telephone plant improvements cannot be made in all parts at one time. They must be introduced gradually as new systems and apparatus are applied, and with a far-sighted concern for future trends.

The Dielectric Properties of Insulating Materials

By E. J. MURPHY and S. O. MORGAN

This paper gives a qualitative account of the way in which dielectric constant and absorption data have been interpreted in terms of the physical and chemical structure of materials. The dielectric behavior of materials is determined by the nature of the polarizations which an impressed field induces in them. The various types of polarization which have been demonstrated to exist are listed, together with an outline of their characteristics.

I. Outline of the Physico-Chemical Interpretation of the Dielectric Constant

THE development of dielectric theory in recent years has been along such specialized lines that there is need of some correlation between the newer and the older theories of dielectric behavior to keep clear what is common to both, though sometimes expressed in different terms. The purpose of the present paper is to outline in qualitative terms the way in which the dielectric constant varies with frequency and temperature and to indicate the type of information regarding the structure of materials which can be obtained from the study of the dielectric constant.

The important dielectric properties include dielectric constant (or specific inductive capacity), dielectric loss, loss factor, power factor, a.c. conductivity, d.c. conductivity, electrical breakdown strength and other equivalent or similar properties. The term *dielectric behavior* usually refers to the variation of these properties with frequency, temperature, voltage, and composition.

In discussing the dielectric properties and behavior of insulating materials it will be necessary to use some kind of model to represent the dielectric. The success of wave-mechanics in explaining why some materials are conductors and others dielectrics suggests that it might be desirable to use a quantum-mechanical model even in a general outline of the characteristics of dielectrics, but for the aspects of the theory of dielectric behavior with which we are immediately concerned here the behavior predicted is essentially the same as that derived on the basis of classical mechanics. However, in the course of the description of the frequency-dependence of dielectric constant we shall have occasion to make a comparison between the dispersion and absorption curves for light and those for electromagnetic disturbances in the electrical (i.e., radio and power) range of frequencies. The difficulty is then met that the quantum-mechanical model is the customary medium of description of the absorption of light. But, since the references to optical properties will be only incidental and for comparative purposes, there is little to be lost, even in this domain in which quantum-mechanical concepts are the familiar medium of description, in using the pre-quantum theory concepts of dispersion and absorption processes. Thus a model operating on the basis of classical mechanics and the older conceptions of atomic structure will be sufficient for our present purposes.

On the wave-mechanical theory of the structure of matter a dielectric is a material which is so constructed that the lower bands of allowed energy levels are completely full at the absolute zero of temperature (on the Exclusion Principle) and at the same time isolated from higher unoccupied bands by a large zone of forbidden energy levels.[1] Thus conduction in the lower, fully occupied bands is impossible because there are no unoccupied energy levels to take care of the additional energy which would be acquired by the electrons from the applied field, while the zone of forbidden energy levels is so wide that there is only a negligible probability that an electron in the lower band of allowed levels will acquire enough energy to make the transition to the unoccupied upper band where it could take part in conduction. The bound electrons in a completely filled and isolated band of allowed levels can, however, interact with the applied electric field by means of the slight modifications which the applied field makes in the potential structure of the material and hence in the allowed levels.

On the other hand in the older theory of the structure of matter the essential condition which makes a material a dielectric is that the electrons and other charged particles of which it is composed are held in equilibrium positions by constitutive forces characteristic of the structure of the material. When an electric field is applied these charges are displaced, but revert to their original equilibrium positions when the field is removed. In this account of the behavior of dielectrics this model will be sufficient, but no essential change in the relationships which will be discussed here would result if a translation were made to a model based upon quantum-mechanics.

When an electric field is impressed upon a dielectric the positive and negative charges in its atoms and molecules are displaced in opposite directions. The dielectric is then said to be in a polarized

[1] Cf., for example, Gurney, "Elementary Quantum Mechanics," Cambridge (1934); Herzfeld, "The Present Theory of Electrical Conduction," *Electrical Engineering*, April 1934.

condition, and since the motion of charges of opposite sign in opposite directions constitutes an electric current there is what is called a *polarization current* or *charging current* flowing while the polarized condition is being formed.

For the case of a static impressed field a charging current flows in the dielectric only for a certain time after application of the field, the time required for the dielectric to reach a fully polarized condition. If the material is not an ideal dielectric, but contains some free ions, the current due to a static impressed field does not fall to zero but to a constant value determined by the conductivity due to free ions. More important than the static is the alternating current case, where the potential is continually varying and where, consequently, there must be a continuously varying current.

The dielectric behavior of different materials under different conditions is reflected in the characteristics of the charging or polarization currents, but since polarization currents depend upon the applied voltage and the dimensions of condensers it is inconvenient to use them directly for the specification of the properties of materials. Eliminating the dependence upon voltage by dividing the charge by the voltage, we have the capacity ($C = Q/V$); and the dependence upon dimensions may be eliminated by using the dielectric constant, defined as $\epsilon = C/C_0$, where C is the capacity of the condenser when the dielectric material is between its plates and C_0 is the capacity of the same arrangement of plates in a vacuum. The dielectric constant is then a property of the dielectric material itself.

The term "dielectric polarization" is used to refer to the polarized condition created in a dielectric by an applied field of either constant or varying intensity. The *polarizability* is one of the quantitative measures of the dielectric polarization; it is defined as the electric moment per unit volume induced by an applied field of unit effective intensity. Another quantitative measure of the dielectric polarization is the *molar polarization;* this is a quantity which is a measure of the polarizability of the individual molecule, whatever the state of aggregation of the material.

The concept of polarizability is as fundamental to, and plays about the same role in, the theory of dielectric behavior as does the concept of free ions in the theory of electrolytic conduction. Just as the conductivity of a material is a measure of the product of the number of ions per unit cube and their average *velocity* in the direction of a unit applied field, so the polarizability is a measure of the number of bound charged particles per unit cube and their average *displacement* in the direction of the applied field.

In the early investigations of dielectrics two distinct types of charging current were recognized, the one in which the charging or discharging of a condenser occurred practically instantaneously and the other in which a definite and easily observable time was required. A charge accumulating in a condenser in an unmeasurably short time was variously referred to as the instantaneous charge or geometric charge or the elastic displacement. The current by which this charge is formed was called the instantaneous or geometric charging current, and similarly the terms *instantaneous dielectric constant* or *geometric dielectric constant* were used to describe the property of the medium giving rise to the effect between the condenser plates. An even wider variety of names has been used for the part of the charge which formed or disappeared more slowly. Among these are residual charge, reversible absorption, inelastic displacement, viscous displacement and anomalous displacement. The modern theory still recognizes these two distinct types of condenser charges and charging currents but the simple descriptive designations *rapidly-forming or instantaneous polarizations* and *slowly-forming or absorptive polarizations* will be adopted here, as they seem sufficient and to be preferred to terms which have more specialized connotations as to the mechanism upon which the behavior depends. The properties of these two types of charging currents and the dielectric polarizations corresponding to them appear prominently in the theories of dielectric behavior.

The total polarizability of the dielectric is the sum of contributions due to all of the different types of displacement of charge produced in the material by the applied field. Constitutive forces characteristic of the material determine both the magnitude of the polarizability and the time required for it to form or disappear. The quantitative measure of the time required for a polarization to form or disappear is called the *relaxation-time*. In the following a description will be given of the physical processes involved in the formation of dielectric polarizations, indicating the effect of chemical and physical structure upon the two quantities, magnitude and relaxation-time, which determine many of the properties of dielectric polarizations of the slowly-forming or absorptive type.

The magnitude of the polarizability k of a dielectric can be expressed in terms of a directly measurable quantity, the dielectric constant ϵ, by the relation

$$k = \frac{3}{4\pi}\frac{(\epsilon - 1)}{(\epsilon + 2)} \cdot$$

It is sometimes convenient to use the polarizability and the dielectric constant interchangeably in the qualitative discussion of the magnitude of the dielectric polarization. In dealing with alternating currents the fact that polarizations of the absorptive type require a time to form which is often of the same order of magnitude as, or greater than, the period of the alternations, results in the polarization not being able to form completely before the direction of the field is reversed. This causes the magnitude of the dielectric polarization and dielectric constant to decrease as the frequency of the applied field increases. An example of this variation of the dielectric constant with frequency is shown in the radio and power frequency section of the curve plotted in Fig. 1. It is often convenient to refer to the mid-point of the decreasing dielectric constant-frequency curve as the *relaxation-frequency;* this frequency f_m is very simply related to the relaxation-time τ, for the theory of these effects shows that $f_m = 1/2\pi\tau$.

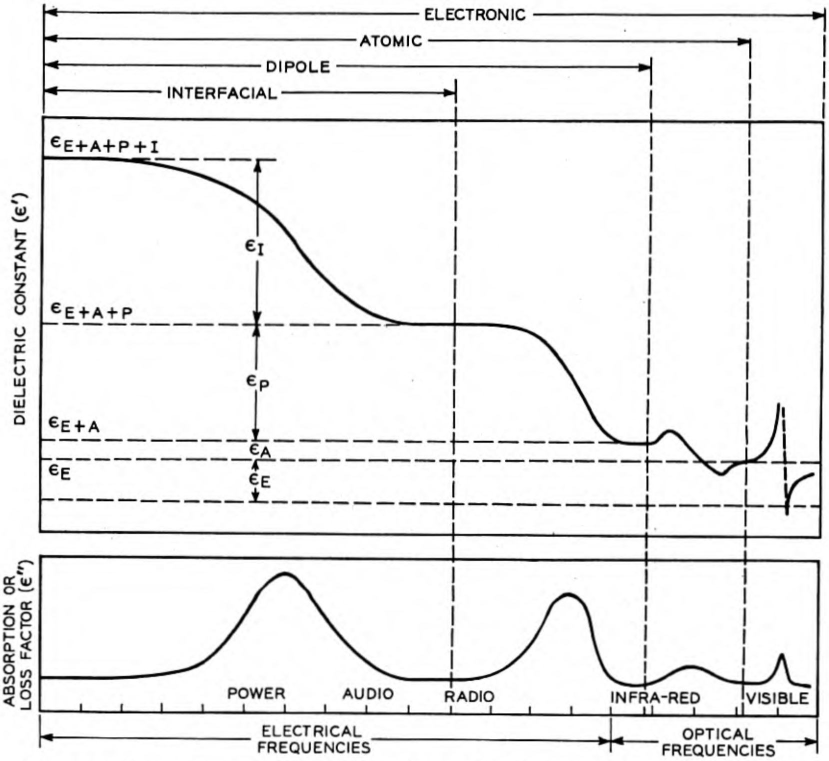


Fig. 1—Schematic diagram of variation of dielectric constant and dielectric absorption with frequency for a material having electronic, atomic, dipole and interfacial polarizations.

Various types of polarization can be induced in dielectrics: There should be an electronic polarization due to the displacement of electrons with respect to the positive nuclei within the atom; an atomic polarization due to the displacement of atoms with respect to each other in the molecule and in certain ionic crystals, such as rock salt, to the displacement of the lattice ions of one sign with respect to those of the opposite sign; dipole polarizations due to the effect of the applied field on the orientations of molecules with permanent dipole moments; and finally interfacial (or ionic) polarizations caused by the accumulation of free ions at the interfaces between materials having different conductivities and dielectric constants.

Electronic Polarizations

A classification of dielectric polarizations into rapidly-forming or instantaneous polarizations and slowly-forming or absorptive polarizations has been made. Instantaneous polarizations may be thought of as polarizations which can form completely in times less than say 10^{-10} seconds, that is, at frequencies greater than 10^{10} cycles per second or wave-lengths of less than 1 centimeter, and so beyond the range of conventional dielectric constant measurements. The *electronic polarizations* are due to the displacement of charges within the atoms, and are the most important of the instantaneous polarizations. The polarizability per unit volume due to electronic polarizations may be considered to be a quantity which is proportional to the number of bound electrons in a unit volume and inversely proportional to the forces binding them to the nuclei of the atoms.

The effect of number of electrons and binding force is illustrated by a comparison of the values for the polarizability per unit volume of different gases; for the number of molecules per unit volume is independent of the composition of the gas. Thus, although a c.c. of hydrogen with two electrons per molecule has the same number of electrons as a c.c. of helium, which is an atomic gas with two electrons per atom, the quantity $\epsilon - 1$, that is the amount by which the dielectric constant is greater than that of a vacuum, is nearly four times as large for hydrogen as for helium. This shows that in hydrogen the electrons are in effect less tightly bound to the nucleus than in helium, resulting in a larger induced polarization. Nitrogen has a larger dielectric constant than either hydrogen or helium because it has 14 electrons per molecule. Some of these are tightly bound as in helium and some are more loosely bound as in hydrogen.

The dielectric constant of *liquid* nitrogen is 1.43, which is much higher than the value 1.000600 for the gas. This is due to the fact that the number of molecules, and consequently of bound charges, per unit volume is much greater in the liquid than in the gas. However, the molar polarization, a quantity which is corrected for variations in density, is the same for liquid as for gaseous nitrogen.

The time required for the applied field to displace the electrons within an atom to new positions with respect to their nuclei is so short that there is no observable effect of time or frequency upon the value of the dielectric constant until frequencies corresponding to absorption lines in the visible or ultra-violet spectrum are reached. For convenience in this discussion the frequency range which includes the infra-red, visible and ultra-violet spectrum will be called the *optical frequency range* while that which includes radio, audio and power frequencies will be called the *electrical frequency range.* For all frequencies in the electrical range the electronic polarization is independent of frequency and for a given material contributes a fixed amount to the dielectric constant, but at the frequencies in the optical range corresponding to the absorption lines in the spectrum of the material, the dielectric constant, or better the refractive index, changes rapidly with frequency, and absorption appears. (The justification for using refractive index n and dielectric constant ϵ interchangeably for the qualitative discussion of the properties of dielectric polarizations follows from the relation, $\epsilon = n^2$, which is known as Maxwell's rule. This is a general relationship based upon electromagnetic theory and is applicable whenever ϵ and n are measured at the same frequency no matter how high or low it may be.)

The electronic polarization of a molecule may be regarded as an additive property of the atoms or of the atomic bonds in the molecule, and may be calculated for any dielectric of known composition with sufficient accuracy for most purposes. Within any one chemical class of compounds such as, for example, the saturated hydrocarbons or their simple derivatives, in which all of the bonds are $C-H$, $C-C$ or $C-X$, the calculated values agree with the measured to within a few per cent. For other classes of compounds—for example, benzene, in which there are both single and double bonds such calculations must be corrected for the fact that some of the valence electrons have their binding forces and hence their polarizabilities altered in the double bond as compared to the single bond. Such values of electronic polarization, usually called atomic refractions, have been determined for all of the different types of bonds from the vast amount of experimental study of refractive indices of organic and inorganic compounds.

In some materials the electronic polarization is the only one of importance. For example, in benzene the dielectric constant is the same at all frequencies in the electrical range and is equal to the square of the optical refractive index. This must mean that the only polarizable elements of consequence in C_6H_6 are electrons which are capable of polarizing as readily in the visible spectrum, where the refractive index is measured, as at lower frequencies where dielectric constant is measured. The refractive index in the visible spectrum provides the means of determining the magnitude of electronic polarizations, for other types of polarization are usually of negligible magnitude when the frequency of the impressed field lies in the visible spectrum. For materials having only electronic polarizations the dielectric properties are very simply dependent upon the chemical composition and the temperature, and are independent of frequency in the electrical frequency range. In many materials, however, there are also other polarizations which can form at low frequencies but not at high; these are characterized by more complex dielectric behavior.

Atomic Polarizations

Included among the polarizations which may be described as instantaneous by comparison with the order of magnitude of the periods of alternation of the applied field in the electrical frequency range are those arising from the displacement of the ions in an ionic crystal lattice (such as rock salt) or of atoms in a molecule or molecular lattice. In some few materials, for example the alkali halides, sufficient study has been made of the infra-red refractive index to provide data on the atomic polarizations, but for most substances little is known about them. What is known has in part been inferred from infra-red absorption spectra and in part from the infra-red vibrations revealed by studies of the Raman effect.

Atomic polarizations are distinguished from electronic polarizations by being the part of the polarization of a molecule which can be attributed to the relative motion of the atoms of which it is composed. The atomic polarizations may be attributed to the perturbation by the applied field of the vibrations of atoms and ions having their characteristic or resonance frequencies in the infra-red. Atomic polarizations may be large for substances such as the alkali halides and other inorganic materials, but are usually negligible for organic materials. The exact value of the time required for the formation of atomic polarizations is unimportant in the electric range of frequencies with which we are primarily concerned. The essential thing is that atomic polarizations begin to contribute to ϵ(or n^2) at frequencies below approximately 10^{14} seconds—that is, in the near infra-red and that below about 10^{10} cycles per second, where the optical and electrical frequency ranges merge, atomic polarizations contribute a constant amount to ϵ(or n^2) for a given material. The atomic polarization is determined as the difference between the polarization which is measured at some low infra-red or high electric frequency and the electronic polarization as determined from refractive index measurements in the visible spectrum.

The electronic and atomic polarizations are considered to comprise all of the so-called instantaneous polarizations; that is, the polarizations which form completely in a time which is very short as compared with the order of magnitude of the periods of applied fields in the electrical range of frequencies.

The Debye Orientational Polarization

The remaining types of polarization are of the "absorptive" kind, characterized by relaxation-times corresponding to "relaxation-frequencies" in the electrical range of frequencies. These polarizations include the important type which is due to the effect of the applied field on the orientation of molecules with permanent electric moments, the theory of which was developed by Debye. Among the other possible polarizations of the absorptive type are those due to interfacial effects or to ions which are bound in various ways.

Debye,[2] in 1912, suggested that the high dielectric constant of water, alcohol and similar liquids was due to the existence of permanent dipoles in the molecules of these substances. The theory which Debye based upon this postulate opened up a new field for experimental investigation by providing a molecular mechanism to explain dielectric behavior which fitted into and served to confirm the widely held views of chemical structure. Debye postulated that the molecules of all substances except those in which the charges are symmetrically located possess a permanent electric moment which is characteristic of the molecule. In a liquid or gas these molecular dipoles are oriented at random and therefore the magnitude of the polarization vector is zero. When an electric field is applied, however, there is a tendency for the molecules to align themselves with their dipole axes in the direction of the applied field, or, put in another way, to spend more of their time with their dipole axes in the direction of the field than in the opposite direction. This dipole polarization is superimposed upon the electronic and atomic polarizations which are also induced by the field. The theory as developed by Debye accounts for the observed difference between the temperature and frequency dependence of the dipole polarizations and the instantaneous polarizations. While the latter are present in all dielectrics, the dipole polarizations can

[2] P. Debye, *Phys. Zeit.*, *13*, 97, (1912); *Verh. d. D. phys. Ges.*, *15*, 777 (1913).

occur only in those made up of molecules which are electrically asymmetrical.

Polar molecules (that is molecules with permanent electric moments) are, by definition, those in which the centroid of the negative charges does not coincide with the centroid of the positive charges, but falls at some distance from it. All materials must be classed either as polar or non-polar, the latter class including those which are electrically symmetrical. Some simple examples of non-polar molecules are H_2, N_2, O_2, CH_4, CCl_4 and C_6H_6. In these molecules each C — H, C — Cl or other bond may be regarded as having a vector dipole moment of characteristic magnitude located in the bond. Where the sum of these vector moments is zero the molecule will be non-polar. Both CH_4 and CCl_4 meet this requirement but CH_3Cl is polar because the C — Cl vector moment is considerably greater than the resultant of the three C — H vectors. (See Fig. 2.) Polar molecules are the rule and non-polar the exception.

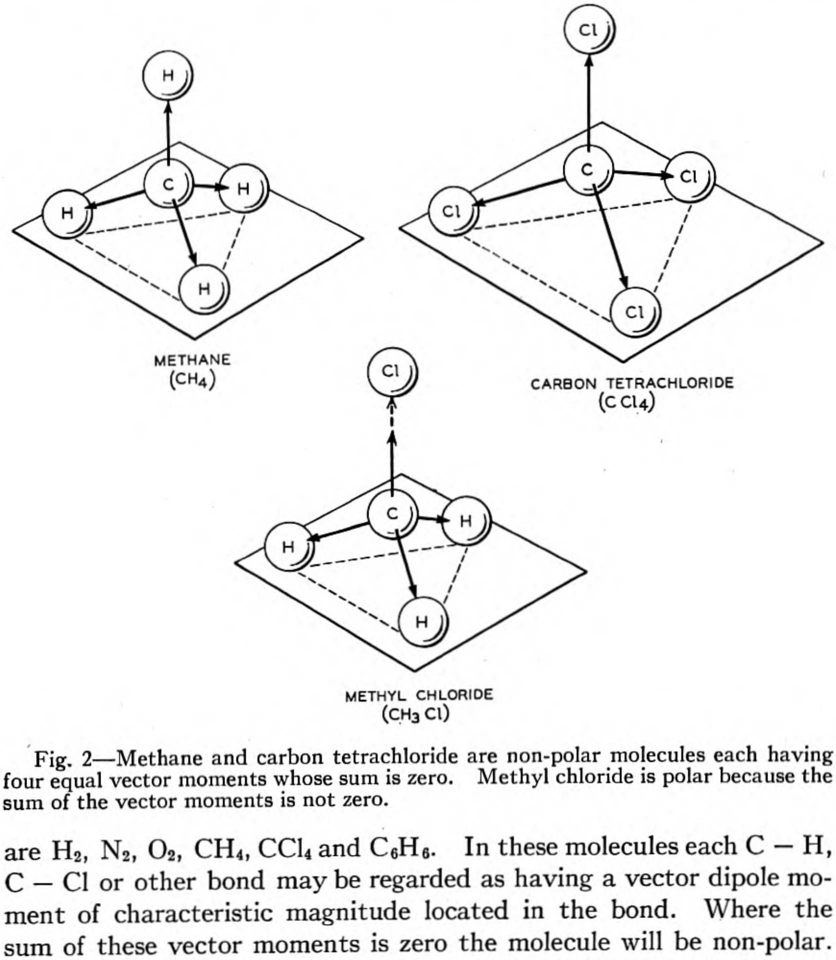


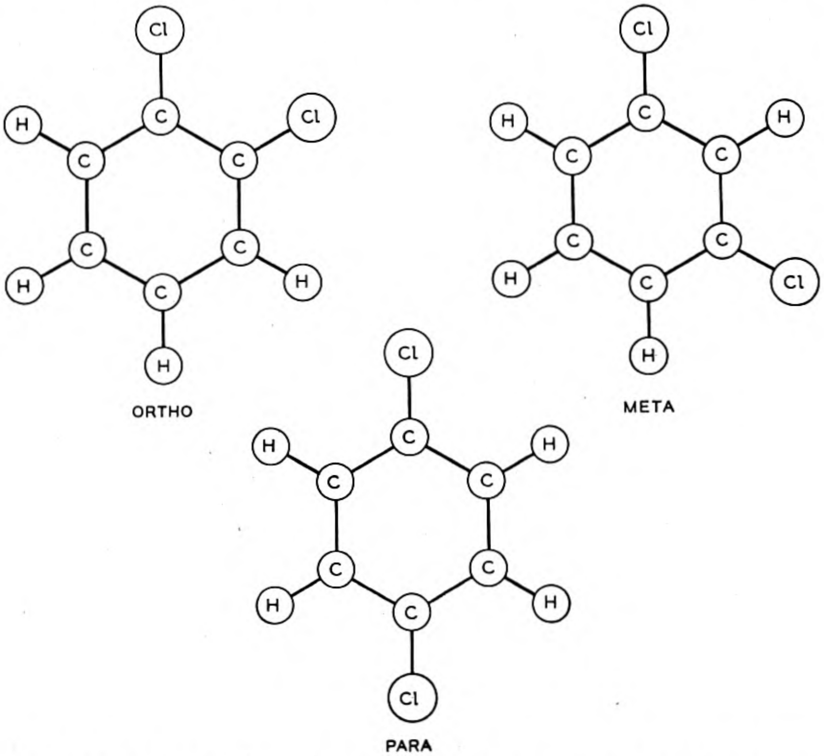
Fig. 2—Methane and carbon tetrachloride are non-polar molecules each having four equal vector moments whose sum is zero. Methyl chloride is polar because the sum of the vector moments is not zero.

In the discussion of dipole polarizations it has frequently been pointed out that non-polar materials usually obey the general relationship $\epsilon = n^2$ whereas for polar materials such as H_2O, NH_3 and HCl this rule is apparently not obeyed. Water, for example, has $n^2 = 1.7$ and $\epsilon = 78$. This apparent discrepancy arises because the refractive index as measured in the *visible* spectrum is usually compared with the dielectric constant as measured in the electric range of frequencies. Non-polar materials usually have only electronic polarizations and these can form both in the optical and in the electrical frequency ranges, but the dipole polarizations can form and contribute to the dielectric constant only in the electrical frequency range; this is the most frequent source of the above mentioned discrepancy. The general relationship $\epsilon = n^2$ should apply for any material at any frequency provided ϵ and n are measured at the same frequency. The refractive index of water when measured with electric waves,[3] for example, at a million cycles, is found to be slightly less than 9, the square of which agrees very well with the observed value $\epsilon = 78$. However, it does not always follow that when $\epsilon > n^2$ the molecules of which the material is composed have permanent dipole moments, for this condition can also result from the presence of any slowly-forming or absorptive polarization or of a large atomic polarization. Experimental investigations based upon the Debye theory have shown, however, that in the case of water and many other familiar compounds the orientation of dipole molecules actually accounts for the high dielectric constant.

The Debye theory shows that the magnitude of the dipole polarization of a material is proportional to the square of the electric moment of the molecule, which, as has been pointed out, may be regarded as the vector sum of a number of constituent moments characteristic of the individual atoms or radicals of which the molecule is composed, or alternatively, of the bonds which bind these atoms into molecules or more complex aggregates. The very great amount of experimental study of the Debye theory has shown that the NO_2 and CN groups have the largest group moments while CO, OH, NH_2, Cl, Br, I and CH_3 have progressively smaller group moments. The value 34 for the dielectric constant of nitrobenzene ($C_6H_5NO_2$), as against 5.5 for chlorobenzene (C_6H_5Cl), 2.8 for methyl benzene ($C_6H_5CH_3$) and 2.28 for benzene (C_6H_6), which is non-polar, are evidence of the large differences in the magnitudes of these group moments and the large part that dipole moments can play in determining the dielectric constant.

[3] Drude, "Physik des Aethers," Stuttgart (1894), p. 486.

Another point regarding molecular structure shown by such studies is that it is not only the presence of polar groups in the molecule but also their position which determines the electric moment of the molecule. This is nicely illustrated by the dichlorobenzenes, of which there are three isomers. As is shown in Fig. 3, ortho-dichlorobenzene, having the two substituent groups in adjacent positions, is the most asymmetrical of the three compounds, and consequently has the highest electric moment, $\mu = 2.3$. The meta compound has about the same moment as monochlorobenzene, $\mu = 1.55$. The para compound, however, is symmetrical and has zero electric moment because the Cl atoms are substituted on opposite sides of the benzene ring so that their vector moments cancel. These values of electric moment are reflected in the values of dielectric constant which are respectively 10, 5.5 and 2.8 for the three isomeric dichlorobenzenes.

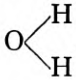


ORTHO

META

PARA

Fig. 3—Ortho dichlorobenzene being the more asymmetrical has a higher electric moment than the meta isomer; the para isomer which is symmetrical has zero electric moment.

Dielectric studies of this kind have also shown, for example, that H_2O is not a symmetrical linear molecule, $H - O - H$, but rather a triangular structure . CO_2 on the other hand, being non-polar, is determined to be a linear molecule $O = C = O$. Thus, dielectric measurements interpreted by the Debye theory have become established as one of the standard means of studying molecular structure.

Since dipole polarizations depend upon the relative orientations of molecules, rather than upon the displacement of charges within the atom or molecule, the time required for a polarization of this type to form depends upon the internal friction of the material. Debye expressed the time of relaxation of dipole polarizations in terms of the internal frictional force by the equation:

$$\tau = \frac{\zeta}{2kT} = \frac{8\pi\eta a^3}{2kT},$$

where ζ is the internal friction coefficient, η is the coefficient of viscosity, a the radius of the molecule and T the absolute temperature.[4] This latter expression, because it depends on Stokes' law for a freely falling body, is rigidly applicable only to gases or possibly to dilute solutions of polar molecules in non-polar solvents in which the polar molecules are far enough apart that they exert no appreciable influence on each other.

Applying this equation to the calculation of the relaxation-time of the orientational polarizations in water at room temperature we obtain $\tau = 10^{-10}$ seconds, assuming a molecular radius of 2×10^{-8} cm. and taking the viscosity as 0.01 poises.[5] The relaxation-frequency corresponding to this relaxation-time is about 1.6×10^9 cycles/sec., agreeing with the results of experimental studies on water which show that in the range of frequencies extending from 10^9 to 10^{11} cycles the dielectric constant decreases from its high value to a value approximately equal to the square of the refractive index. Thus the drop in dielectric constant occurs in the frequency range which corresponds to the calculated value of the relaxation-time.

Similar experiments on dilute solutions of alcohols[6] in non-polar solvents yield values of τ of about 10^{-9} seconds. The shortest relaxation-times which dipole polarizations can have are probably not

[4] P. Debye, "Polar Molecules," Chem. Cat. Co., 1929, p. 85.

[5] The viscosity of a liquid in poises is given by the force in dynes required to maintain a relative tangential velocity of 1 cm./sec. between two parallel planes in the liquid each 1 cm.^{2} in area and 1 cm. apart, the distance being measured normal to their surfaces.

[6] R. Goldammer, *Phys. Zeit.*, *33*, 361 (1932).

much less than the order of 10^{-11} seconds, since in general either the internal friction or the molecular radius of materials having polar molecules will be greater than those of water, resulting in longer relaxation-times. No long-time limit can be placed on the relaxation-times which dipole polarizations may have, for they are limited only by the values which the internal friction can assume. For materials, such as glycerine, which tend to become very viscous at low temperatures the time of relaxation of the dipoles may be a matter of minutes. Studies of the dielectric constant of crystalline solids, to be discussed in a later paper, show also that in some cases polar molecules are able to rotate even in the crystalline state, where the ordinary coefficient of viscosity has no meaning because the materials do not flow. In connection with the dielectric properties we are concerned only with the ability of the polar molecules to undergo rotational motion and it is likely that in these solids, which constitute a special class, the internal frictional force opposing rotation of the molecules is small even though the forces opposing translational motion may be very large. The particular equation for the calculation of the time of relaxation given above obviously does not apply to solids.

In discussing the three types of polarizations which have been considered thus far, it has been pointed out that the magnitude of the dielectric constant depends upon the polarizability of the material. Each type of polarization makes a contribution to the dielectric constant if the measuring frequency is considerably below its relaxation-frequency. However, if the frequency of the applied field used for measuring the dielectric constant is too high the presence of polarizations with low relaxation-frequencies will not be detected. Thus the refractive index of water in the visible spectrum is 1.3 and therefore gives no evidence whatever of the presence of permanent dipoles. This is due to the fact that the H_2O molecules do not change their orientations rapidly enough to allow fields which alternate in direction as rapidly as those of light to cause an appreciable deviation from the original random orientation which prevails in the absence of an applied field.

The band of frequencies in which the dielectric constant decreases with increasing frequency because of inability of the polarization to form completely in the time available during a cycle, is called a region of absorption or of *anomalous dispersion.* The discussion of this characteristic of dielectric materials forms an important part of dielectric theory. The term anomalous dispersion is no doubt usually thought of in connection with the anomalous dispersion of light: when the refractive index of light decreases with increasing frequency the material is said to display anomalous dispersion in the range of frequencies concerned. However, in a paper published in 1898 Drude [7] applied this term to the decrease of dielectric constant with increasing frequency in the electrical range of frequencies. The justification for this extension of the original application of the term is very direct for electromagnetic theory shows that the dielectric constant and the refractive index of a material are connected by the general relationship $\epsilon = n^2$ whatever the frequency of the electromagnetic disturbance. As the dispersion of light by a prism is due to the variation of its refractive index with frequency, the use of the expression *anomalous dispersion* to refer to the decrease of dielectric constant with increasing frequency is consistent and has become generally accepted.

Interfacial Polarizations

The polarizations thus far considered are the main types to be expected in a *homogeneous* material. They depend upon the effect of the applied field in slightly displacing electrons in atoms, in slightly distorting the atomic arrangement in molecules and in causing a slight deviation from randomness in the orientation of polar molecules. The remaining types of polarization are those resulting from the *heterogeneous* nature of the material and are called *interfacial polarizations*. Interfacial polarizations must exist in any dielectric made up of two or more components having different dielectric constants and conductivities except for the particular case where $\epsilon_1\gamma_2 = \epsilon_2\gamma_1$, γ being the conductivity [8] and the subscripts referring to the two components. Heterogeneity in a dielectric may be due to a number of causes, and in the case of practical insulating materials is probably the rule rather than the exception. Impregnated paper condensers and laminated plastics are obvious examples of heterogeneous dielectrics. Paper is itself a heterogeneous dielectric, consisting of water and cellulose. In all probability the plastic resins are also heterogeneous, and certainly so if they contain fillers. Ceramics, being mixtures of crystalline and glassy phases, are also heterogeneous.

The simplest case of interfacial polarization is that of the *two-layer* dielectric, that is, a composite dielectric made up of two layers, the dielectric constants and conductivities of which are different. Maxwell showed that in such a system the capacity was dependent upon the charging time. This is due to the accumulation of charge at the interface between the two layers, for this charge must flow through a

[7] P. Drude, *Ann. d. Physik, 64*, 131 (1898), "Zur Theorie der anomalien elektrischen Dispersion."

[8] In this expression γ represents the total a.c. conductivity, a quantity which depends on the frequency.

layer of dielectric whose resistance may be high enough that the interface does not become completely charged during the time allowed for charging. For the alternating current case this implies a decrease of capacity with increasing frequency, which is equivalent to the anomalous dispersion which has been described for the case of dipole polarizations. It should be particularly emphasized that the term *anomalous dispersion* describes a type of variation of dielectric constant with frequency which can be produced by a number of different physical mechanisms.

The two-layer dielectric is of less interest than a generalization of this type of polarization which includes heterogeneous systems composed of particles of one dielectric dispersed in another. This type of heterogeneous dielectric is of considerable importance since such systems represent the actual structure of many practical dielectrics. Such a generalization of the two-layer dielectric has been made by K. W. Wagner [9] who developed the theory for the case of spheres of relatively high conductivity dispersed in a continuous medium of low conductivity. The conditions for the existence of an interfacial polarization are, as in the two-layer case, that $\epsilon_1\gamma_2 \neq \epsilon_2\gamma_1$, where the symbols have the significance just given. This type of polarization, which is variously referred to as an interfacial polarization, an ionic polarization and a Maxwell-Wagner polarization, shows anomalous dispersion like other absorptive polarizations. When the particle size is small as compared with the electrode separation it may be treated as a uniformly distributed polarization.

The magnitude and time of relaxation of interfacial polarizations are determined by the differences in ϵ and γ of the two components. There is a widely prevalent opinion that this type of polarization always has such long relaxation-times as to be observed only at very low frequencies. While this is true for mixtures of very low-conductivity components, the general equations show that for the case where one component has a high conductivity—for example equal to that of a salt solution—the dispersion may occur in the radio frequency range.

Several special types of interfacial polarization have been proposed to explain the dielectric properties of various non-homogeneous dielectrics where something regarding the nature of the inhomogeneity is known. The dielectric constant of cellulose, for example, receives a contribution from an interfacial polarization due to the water and dissolved salt which it contains. Experimental evidence indicates that an aqueous solution of various salts is distributed through the

[9] K. W. Wagner, *Arch. f. Elektrotechn.*, *2*, pp. 374 and 383.

cellulose in such a way as to form a reticulated pattern which may correspond to the pattern formed by the micelles or to some feature of it. An interesting feature of this structure is that the conductance of the aqueous constituent can be changed by varying the moisture content or the salt content of the material and the effect on the dielectric constant observed.[10]

Frequency Dependence of Dielectric Constant

As has been pointed out, each of the different types of polarization may contribute to the dielectric constant an amount depending upon the polarizability and its time of relaxation. The upper curve in Fig. 1 shows schematically the variation of the dielectric constant (or of the square of the refractive index) for a hypothetical material possessing an interfacial polarization with relaxation-frequency in the power range, a dipole polarization with relaxation frequency in the high radio frequency range and atomic and electronic polarizations with dispersion regions in the infra-red and visible respectively. If polarizability were plotted, instead of ϵ (or n^2), the curves would be of the same general form but of different magnitudes, because of a relationship between the two given earlier.

At the low-frequency side of Fig. 1, the dielectric constant curve has its highest value, usually called the static or zero-frequency dielectric constant. Here all of the polarizations have time to form and to contribute their full amount to the dielectric constant. With increasing frequency ϵ begins to decrease as the relaxation-frequency of the *interfacial* polarization is approached and reaches a constant lower value (called the infinite-frequency dielectric constant) when the applied frequency is sufficiently above the relaxation-frequency of the polarization that it has not time to form appreciably. It is this decrease of ϵ with frequency which is called anomalous dispersion. The horizontal arrows across the top of Fig. 1 indicate the frequency region in which the various types of polarizations indicated are able to form and contribute to the dielectric constant.

At still higher frequencies we see that ϵ again decreases as the relaxation-frequency of the *dipole* polarization is approached, and again reaches a constant lower value as the frequency becomes too high for the field to affect appreciably the orientation of dipoles. This second region of anomalous dispersion is similar to the first, which was due to interfacial polarizations. It has been shown as occurring at a higher frequency, but it should be emphasized that the frequency ranges chosen to illustrate anomalous dispersion in Fig. 1

[10] Murphy and Lowry, *Jour. Phys. Chem.*, *34*, 594 (1930).

are purely arbitrary. Anomalous dispersion due to dipole polarizations has been observed at power frequencies while that due to interfacial polarizations has been observed at radio frequencies. The two types of polarizations may in fact give rise to anomalous dispersion in the same frequency range in a given dielectric.

Proceeding to still higher frequencies in Fig. 1 other regions of dispersion appear in the infra-red and visible spectrum. These regions show a combination of normal optical dispersion, in which the dielectric constant, or better now the refractive index, increases with frequency, and anomalous dispersion in which it decreases. The dispersion in the visible range of frequencies is predominantly normal (anomalous dispersion being confined to relatively narrow frequency bands) whereas in the electrical range the reverse is true, normal dispersion not being observed; the infra-red represents an intermediate region. Dipole and interfacial polarizations are not represented in the dispersion in the optical range, the dielectric constant (or refractive index) in the visible being due to electronic polarizations and in the infra-red to electronic and atomic polarizations.

The curves plotted in Fig. 1 are merely schematic and the relative magnitudes of the different contributions to the dielectric constant are therefore arbitrary. However, experimental results indicate that the contribution ϵ_E of the electronic polarization to the dielectric constant is limited to values between 2 and 4 except for certain inorganic materials, since very few organic solids or liquids are known which have refractive indices in the visible spectrum which are greater than 2 or less than 1.4. The contribution ϵ_A of atomic polarizations to the dielectric constant is in general small and is usually negligible, as has been indicated on the curve, although the possibility exists of special cases occurring in which the infra-red refractive indices are very high. The contributions ϵ_P and ϵ_I of dipole and interfacial polarizations to the dielectric constant may vary greatly from one material to another, depending upon the symmetry of the molecule and the physical structure of the dielectric. From the above mentioned limitations on the contribution to the dielectric constant which can be expected from electronic and atomic polarizations, it is apparent that the explanation of values of ϵ higher than 3 to 4, at least in organic materials, requires the existence of some absorptive polarization such as arises from dipoles or interfacial effects. Thus all of the liquids which have high dielectric constants such as H_2O (78), alcohol (24), nitrobenzene (34) have been shown to contain polar molecules.

The lower part of Fig. 1 shows a maximum in the *absorption* for each type of dielectric polarization. The absorption, at least in the electrical frequency range, is due to the dissipation of the energy of the field as heat because of the friction experienced by the bound charges or dipoles in their motion in the applied field in forming the polarizations. The theory of dispersion shows that the dielectric constant and absorption are not independent quantities but that the absorption curve can be calculated from the dielectric constant vs. frequency curve and vice versa. The absorption maximum is greatest for those materials showing the greatest change in dielectric constant in passing through the dispersion region. Thus a material having a high dielectric constant must have a large dielectric loss at the frequency at which ϵ has a value half way between its low and high-frequency values.

Though the quantum theory is necessary for the explanation of many optical and electrical phenomena a simple explanation, sufficient for our purposes, of the general form of the curves of dielectric constant vs. frequency in the infra-red and visible spectrum may be given in terms of the Lorentz theory of optical dispersion. In this theory the form of the dispersion curves depends upon the variation with frequency of the relative importance of the inertia of the typical electron and of the frictional forces and restoring forces acting upon it. For electronic polarizations the frictional or dissipative force is negligible, except in the narrow frequency interval included in the absorption band, and the inertia and restoring force terms predominate. For the atomic polarizations the frictional force is larger and the absorption region extends over a wider interval of frequencies. For dipole and interfacial polarizations the influence of inertia is entirely negligible as compared with the frictional or dissipative forces so that in effect these polarizations may be thought of as aperiodically damped.

Temperature Dependence of Dielectric Constant

The dielectric constant of a material is a constant only in the exceptional case. Besides the variation with frequency which has been considered the dielectric constant varies with temperature. Electronic polarizations may be considered to be unaffected by the temperature. The refractive index does indeed change with temperature but this is completely accounted for by the change of density, and the molar refraction is independent of temperature. The atomic and ionic vibrations are, however, affected by temperature, the binding force between ions or atoms being weakened by increased temperature. This factor of itself would yield a positive temperature coefficient for the atomic polarizations but the decrease in density with the increase in temperature acts in the opposite direction. The result is that calculation of the temperature coefficient of atomic polarizations usually yields zero or slightly positive values. What experimental data there are indicate small positive temperature coefficients for atomic polarizations.

One of the principal achievements of the Debye theory of dipole polarizations has been the manner in which it explains the large negative temperature coefficients of polarization of many liquids. Debye showed that the variation of polarization with temperature could be expressed by the relation $P = A + (B/T)$, in which the constant A is a measure of the instantaneous polarizations which are independent of temperature and B is a measure of the dipole polarizations. In a liquid or gas the molecules are continuously undergoing both translational and rotational motion, and the result of this thermal motion is to maintain a random orientation of molecules. The action of the electric field in aligning the dipoles is opposed by the thermal motion which acts as an influence tending to keep them oriented at random. As the temperature decreases, the thermal energy becomes smaller and the dipole polarization becomes larger, resulting in a negative temperature coefficient of dielectric constant.

The effect of temperature upon interfacial polarizations has not been experimentally investigated to an extent at all comparable with that of dipole polarizations. However, interest in the interfacial or ionic type of polarization has increased considerably in the past few years, and it has applications of some importance. Among these is diathermy which is becoming of considerable importance as a therapeutic agency.

The foregoing qualitative description of the behavior of the dielectric constant and the type of information regarding molecular structure which has been derived from it will be followed in the next section by the derivation of some of the quantitative relationships which are common to all polarizations of the absorptive type.

Variable Frequency Electric Circuit Theory with Application to the Theory of Frequency-Modulation

By JOHN R. CARSON AND THORNTON C. FRY

In this paper the fundamental formulas of variable frequency electric circuit theory are first developed. These are then applied to a study of the transmission, reception and detection of frequency modulated waves. A comparison with amplitude modulation is made and quantitative formulas are developed for comparing the noise-to-signal power ratio in the two modes of modulation.

FREQUENCY modulation was a much talked of subject twenty or more years ago. Most of the interest in it then centered around the idea that it might afford a means of compressing a signal into a narrower frequency band than is required for amplitude modulation. When it was shown that not only could this hope not be realized,* but that much wider bands might be required for frequency modulation, interest in the subject naturally waned. It was revived again when engineers began to explore the possibilities of radio transmission at very short wave lengths where there is little restriction on the width of the frequency band that may be utilized.

During the past eight years a number of papers have been published on frequency modulation, as reference to the attached bibliography will show. That by Professor E. H. Armstrong † deals with this subject in comprehensive fashion. In his paper the problem of discrimination against extraneous noise is discussed, and it is pointed out that important advantages result from a combination of wide frequency bands together with severe amplitude limitation of the received signal waves. His treatment is, however, essentially nonmathematical in character, and it is therefore believed that a mathematical study of this phase of the problem will not be unwelcome. This the present paper aims to supply by developing the basic mathematics of frequency modulation and applying it to the question of noise discrimination with or without amplitude limitation.

The outstanding conclusions reached in the present paper, as regards discrimination against noise by frequency modulation, may be briefly summarized as follows:

* See Bibliography, No. 1.
† See Bibliography, No. 12.

(1) To secure any advantage by frequency modulation as distinguished from amplitude modulation, the frequency band width must be much greater in the former than in the latter system.

(2) Frequency modulation in combination with severe amplitude limitation for the received wave results in substantial reduction of the noise-to-signal power ratio. Formulas are developed which make possible a quantitative estimate of the noise-to-signal power ratio in frequency modulation, with and without amplitude limitation, as compared with amplitude modulation.

It is a pleasure to express our thanks to several colleagues who have been helpful in various ways: to Dr. Ralph Bown who in a brief but very incisive memorandum, which was not intended to be a mathematical study, disclosed all the essential ideas of the quasi-stationary method of attack; to Mr. J. G. Chaffee,* who has been conducting experimental work on frequency modulation in these Laboratories for some years past, by means of which quantitative checks on the accuracy of some of the principal results have been possible; and to various associates, especially Mr. W. R. Bennett and Mrs. S. P. Mead, for detailed criticism of certain portions of the work.

I

In the well-known steady-state theory of alternating currents, the e.m.f. and the currents in all the branches of a network in which the e.m.f. is impressed involve the time t only through the common factor $e^{i\omega t}$ where $i = \sqrt{-1}$ and ω is the *constant* frequency. To this fact is attributable the remarkable simplicity of alternating current theory and calculation, and also the fact that the network is completely specified by its complex admittance $Y(i\omega)$. Thus, if the e.m.f. is $Ee^{i\omega t}$, the steady-state current is

$$I_{ss} = EY(i\omega)e^{i\omega t}. \tag{1}$$

In the present paper we shall deal with the case where the frequency is *variable*, and write the impressed e.m.f. as

$$E \exp\left(i \int_0^t \Omega(t)dt \right). \tag{2}$$

$\Omega(t)$ will be termed the *instantaneous* frequency. This agrees with the usual definition of frequency when Ω is a constant; it is the rate of change of the phase angle at time t; and in addition the interval T between adjacent zeros of $\sin \int \Omega(t)dt$ or $\cos \int \Omega(t)dt$ is approximately $\pi/\Omega(t)$ in cases of practical importance.

* See Bibliography, No. 11.

Instead of dealing with an arbitrary instantaneous frequency $\Omega(t)$ we shall suppose that

$$\Omega(t) = \omega + \mu(t), \tag{3}$$

where ω is a constant and $\mu(t)$ is the variable part of the instantaneous frequency. In practical applications $\mu(t)$ will be written as $\lambda s(t)$ where λ is a real parameter and the mean square value $\overline{s^2}$ of $s(t)$ is taken as equal to 1/2. Other restrictions on $\mu(t)$ will be imposed in the course of the theory to be developed in this paper. Fortunately these restrictions do not interfere with the application of the theory to important problems.

The steady-state current as given by (1) varies with time in precisely the same way as the impressed e.m.f. When the frequency is variable this is no longer true. On the other hand, formula (1) suggests a "quasi-stationary" or "quasi-steady-state current" component, I_{qss}, defined by the formula

$$I_{qss} = EY(i\Omega)\cdot\exp\left(i\int_0^t \Omega dt\right), \tag{4}$$

which corresponds exactly to (1) with the distinction that the admittance is now an explicit function of time. We are thus led to examine the significance of I_{qss} as defined above and the conditions under which it is a valid approximate representation of the actual response of the network to a variable frequency electromotive force, as given by (2).

We start with the fundamental formula of electric circuit theory.[1] Let an e.m.f. $F(t)$ be impressed at time $t = 0$, on a network of indicial admittance $A(t)$; then the current $I(t)$ in the network is given by

$$I(t) = \int_0^t F(t-\tau)\cdot A'(\tau)d\tau. \tag{5}$$

Here $A'(t) = d/dt\cdot A(t)$ and it is supposed that $A(0) = 0$. (This restriction does not limit our subsequent conclusions and is introduced merely to simplify the formulas. Furthermore $A(0)$ is actually zero in all physically realizable networks.)

Omitting the superfluous amplitude constant E we have

$$\begin{aligned} F(t) &= \exp\left(i\int_0^t \Omega dt\right) \\ &= \exp\left(i\omega t + i\int_0^t \mu dt\right), \end{aligned} \tag{6}$$

[1] See J. R. Carson, "Electric Circuit Theory and Operational Calculus," p. 16.

$$\begin{aligned} F(t - \tau) &= \exp\left[i(t - \tau)\omega + i\int_0^{t-\tau} \mu d\tau_1 \right] \\ &= \exp\left[i(t - \tau)\omega + i\int_0^{t} \mu d\tau_1 - i\int_{t-\tau}^{t} \mu d\tau_1 \right] \\ &= \exp\left[i\Omega(t) \right]\cdot\exp\left[- i\omega\tau - i\int_0^{\tau} \mu(t - \tau_1) d\tau_1 \right]. \qquad (7) \end{aligned}$$

Substituting this expression in (5) for $F(t - \tau)$ and writing

$$\exp\left(- i\int_0^{\tau} \mu(t - \tau_1) d\tau_1 \right) = M(t, \tau), \qquad (8)$$

we have for the current in the network

$$I = e^{i\int \Omega dt}. \int_0^t M(t, \tau) e^{-i\omega\tau} A'(\tau) d\tau. \qquad (9)$$

We now split the integral into two parts, thus:

$$\int_0^t = \int_0^\infty - \int_t^\infty.$$

The second integral on the right represents an initial transient which dies away for sufficiently large values of time, t, while the infinite integral represents the total current, I, for sufficiently large values of t. We have therefore

$$\begin{aligned} I &= e^{i\int \Omega dt}. \int_0^\infty M(t, \tau) e^{-i\omega\tau} A'(\tau) d\tau + I_T \qquad (10) \\ &= Y(i\omega, t) e^{i\int \Omega dt} + I_T, \end{aligned}$$

where

$$Y(i\omega, t) = \int_0^\infty M(t, \tau) e^{-i\omega\tau} A'(\tau) d\tau. \qquad (11)$$

The transient current,[2] I_T, is then given by

$$I_T = e^{i\int \Omega dt} \int_t^\infty M(t, \tau) e^{-i\omega\tau} A'(\tau) d\tau. \qquad (12)$$

The foregoing formulas correspond precisely with the formulas for a constant frequency impressed e.m.f.; these are

$$I_{ss} = e^{i\omega t} \int_0^\infty e^{-i\omega\tau} A'(\tau) d\tau, \qquad (10a)$$

[2] Hereafter the transient term I_T of (10) will be consistently neglected and the symbol I will refer only to the quasi-stationary current.

$$Y(i\omega) = \int_0^\infty e^{-i\omega\tau}A'(\tau)d\tau, \tag{11a}$$

$$I_T = e^{i\omega t}\int_t^\infty e^{-i\omega\tau}A'(\tau)d\tau, \tag{12a}$$

to which the more general formulas reduce when $\mu = 0$ and consequently $M = 1$.

We have now to evaluate $Y(i\omega, t)$ as given by (11). We shall assume tentatively, at the outset, that $\mu = \lambda s(t)$ has the following properties:

$$\lambda s(t) \ll \omega \quad \text{for all values of } t,$$
$$-1 \leq s(t) \leq 1,$$
$$-1 \leq \int_0^t s dt \leq 1.$$

With these restrictions the instantaneous frequency lies within the limits $\omega \pm \lambda$.

Let us now replace $M(t, \tau)$ by the formal series expansion

$$M(t, \tau) = M(t, 0) + \frac{\tau}{1!}\left[\frac{\partial}{\partial\tau}M(t, \tau)\right]_{\tau=0} + \frac{\tau^2}{2!}\left[\frac{\partial^2}{\partial\tau^2}M(t, \tau)\right]_{\tau=0} + \cdots, \tag{13}$$

which converges in the vicinity of all values of t for which s has a complete set of derivatives. Then, if we write

$$\left[\frac{\partial^n}{\partial\tau^n}M(t, \tau)\right]_{\tau=0} = (-i)^n C_n(t) \tag{13a}$$

and substitute (13) in (11), we get

$$Y(i\omega, t) = \int_0^\infty e^{-i\omega\tau}A'(\tau)d\tau + \sum_1^\infty (-i)^n C_n \int_0^\infty \frac{\tau^n}{n!}e^{-i\omega\tau}A'(\tau)d\tau. \tag{14}$$

From (11a) it follows at once that

$$\int_0^\infty \frac{\tau^n}{n!}e^{-i\omega\tau}A'(\tau)d\tau = \frac{i^n}{n!}\frac{d^n}{d\omega^n}Y(i\omega), \tag{15}$$

so that

$$Y(i\omega, t) = Y(i\omega) + \sum_1^\infty \frac{1}{n!}C_n(t)\frac{d^n}{d\omega^n}Y(i\omega). \tag{16}$$

The coefficients C_n are easily evaluated from (8) and (13a); they are [3]

$$\begin{aligned} C_1 &= \mu(t), \\ C_2 &= \mu^2 - i\frac{d}{dt}\mu, \qquad (17) \\ &\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot \\ C_{n+1} &= \left(\mu - i\frac{d}{dt}\right)C_n. \end{aligned}$$

Now consider the quasi-stationary admittance $Y(i\Omega)$. Writing $\Omega = \omega + \mu(t)$ and expanding as a power series, we have (assuming that the series is convergent)

$$Y(i\Omega) = Y(i\omega) + \sum_1^\infty \frac{\mu^n}{n!}\frac{d^n}{d\omega^n}Y(i\omega). \qquad (18)$$

From (16), (17) and (18) we have at once

$$Y(i\omega, t) = Y(i\Omega) + \sum_2^\infty \frac{1}{n!}D_n(t)\frac{d^n}{d\omega^n}Y(i\omega), \qquad (19)$$

where

$$\begin{aligned} D_2 &= -\,i\frac{d}{dt}\mu(t), \\ D_3 &= -\,i3\mu\frac{d\mu}{dt} - \frac{d^2}{dt^2}\mu, \qquad (20) \\ &\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot \\ D_{m+1} &= C_{m+1} - \mu^{m+1}. \end{aligned}$$

Consequently, the total current, after initial transients have died away, is given by

$$\begin{aligned} I &= I_{qss} + \Delta(t) \\ &= \exp\left(i\int_0^t \Omega dt\right)\cdot\left[\,Y(i\Omega) - \frac{i}{2!}\frac{d\mu}{dt}\frac{d^2Y}{d\omega^2}\right. \\ &\qquad\qquad \left. - \frac{1}{3!}\left(i3\mu\frac{d\mu}{dt} + \frac{d^2\mu}{dt^2}\right)\frac{d^3Y}{d\omega^3} + \cdots\right]. \qquad (21) \end{aligned}$$

We have thus succeeded in expressing the response of the network in terms of the quasi-stationary current

$$I_{qss} = Y(i\Omega)\cdot\exp\left(i\int \Omega dt\right) \qquad (22)$$

[3] From these recursion formulas C_n can be derived in the compact form

$$\begin{aligned} C_n &= \left(\mu - i\frac{d}{dt}\right)\left(\mu - i\frac{d}{dt}\right)\cdots\left(\mu - \frac{d}{dt}\right)\mu \\ &= \left(\mu - i\frac{d}{dt}\right)^{n-1}\mu \quad \textit{symbolically.} \end{aligned}$$

and a correction series Δ, which depends on the derivatives of the steady-state admittance $Y(i\omega)$ with respect to frequency and the derivatives of the variable frequency $\mu(t)$ with respect to time.

If the parameter λ is sufficiently large and the derivatives of s are small enough so that C_n may be replaced by the two leading terms, we get

$$C_n = \mu^n - i\,\frac{(n-1)n}{2}\,\mu'\mu^{n-2}, \quad \mu' = \frac{d\mu}{dt}\,\cdot$$

Then by (16) and (18)

$$\begin{aligned} Y(i\omega, t) &= Y(i\Omega) - \frac{i\mu'}{2}\sum_{2}^{\infty}\frac{\mu^{n-2}}{(n-2)!}\frac{d^n}{d\omega^n}\,Y(i\omega) \\ &= Y(i\Omega) - \frac{i\mu'}{2}\frac{\partial^2}{\partial\mu^2}\,Y(i\Omega) \\ &= Y(i\Omega) - \frac{i\mu'}{2}\frac{d^2}{d\Omega^2}\,Y(i\Omega) \\ &= Y(i\Omega) + \frac{i\mu'}{2}\,Y^{(2)}(i\Omega). \end{aligned} \tag{16a}$$

The preceding formulas are so fundamental to variable frequency theory and the theory of frequency modulation that an alternative derivation seems worth while. We take the applied e.m.f. as

$$E \exp\left(i\omega_c t + i\theta + i\int_0^t \mu dt\right), \tag{23}$$

the phase angle θ being included for the sake of generality.

Now in any finite epoch $0 \leq t \leq T$, it is always possible to write

$$\exp\left(i\int_0^t \mu dt\right) = \int_{-\infty}^{\infty} F(i\omega)e^{i\omega t}d\omega, \tag{24}$$

thus expressing the function on the left as a Fourier integral. For present purposes it is quite unnecessary to evaluate the Fourier function $F(i\omega)$.

Substitution of (24) in (23) gives for the current

$$I = E\cdot\exp\,(i\omega_c t + i\theta)\cdot\int_{-\infty}^{\infty} F(i\omega)\,Y(i\omega_c + i\omega)e^{i\omega t}d\omega. \tag{25}$$

We suppose as before that, in the interval $0 \leq t \leq T$, $\mu(t)$ and its derivatives are continuous. We can then expand the admittance function Y in the form

$$\begin{aligned} Y(i\omega_c + i\omega) &= Y(i\omega_c) + \frac{i\omega}{1!} Y^{(1)}(i\omega_c) + \frac{(i\omega)^2}{2!} Y^{(2)}(i\omega_c) + \cdots \\ &= Y(i\omega_c) + \sum_1^\infty \frac{(i\omega)^n}{n!} Y^{(n)}(i\omega_c) \\ &= Y(i\omega_c) + \sum_1^\infty \frac{\omega^n}{n!} \frac{d^n}{d\omega_c{}^n} Y(i\omega_c). \end{aligned} \tag{26}$$

Substitution of (26) in (25) gives

$$I = E \cdot \exp\,(i\omega_c t + i\theta) \sum_0^\infty \frac{1}{n!} \frac{d^n}{d\omega_c{}^n} Y(i\omega_c) \int_{-\infty}^{\infty} \omega^n F(i\omega) e^{i\omega t} d\omega. \tag{27}$$

But by the identity (24) and repeated differentiations with respect to t, we have

$$\begin{aligned} \int_{-\infty}^{\infty} \omega F(i\omega) e^{i\omega t} d\omega &= \mu \exp\left(i \int_0^t \mu dt \right), \\ \int_{-\infty}^{\infty} \omega^2 F(i\omega) e^{i\omega t} d\omega &= \left(\mu^2 - i\frac{d\mu}{dt} \right) \exp\left(i \int_0^t \mu dt \right), \\ \cdot \quad \cdot \quad \cdot \quad &\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \int_{-\infty}^{\infty} \omega^n F(i\omega) e^{i\omega t} d\omega &= C_n \exp\left(i \int_0^t \mu dt \right) \cdot \end{aligned} \tag{28}$$

Substitution of (28) in (27) gives

$$I = E \exp\left(i \int_0^t \Omega dt + i\theta \right) \cdot \left\{ Y(i\omega_c) + \sum_1^\infty \frac{1}{n!} C_n \frac{d^n}{d\omega_c{}^n} Y(i\omega_c) \right\}, \tag{29}$$

which agrees with (16).

Formula (25), as it stands, includes the initial transients at time $t = 0$ as well as any which occur at discontinuities in $\mu(t)$. Differentiation with respect to t under the integral sign, however, in effect eliminates these transients and (29) leaves only the quasi-stationary current (plus the correction series given in (19)).

The series appearing in formula (29) may not be convergent; in any case its computation is laborious. Furthermore, in its application to the theory of frequency modulation, terms beyond the first two represent distortion. For these reasons it is often preferable to proceed as follows:

Returning to formula (25), we write

$$Y(i\omega_c + i\omega) = \left(1 + \frac{\omega}{1!} \frac{d}{d\omega_c} + \cdots + \frac{\omega^n}{n!} \frac{d^n}{d\omega_c{}^n} \right) Y(i\omega_c) + R_n(\omega_c, \omega), \tag{30}$$

thus defining the *remainder* R_n. Then (29) becomes

$$I = E \exp\left(i\int_0^t \Omega dt + i\theta\right)\cdot\left[1 + \frac{C_1}{1!}\frac{d}{d\omega_c} + \cdots + \frac{C_n}{n!}\frac{d^n}{d\omega_c{}^n}\right] Y(i\omega_c)$$
$$+ E \exp\,(i\omega_c t + i\theta)\int_{-\infty}^{\infty} R_n(\omega_c,\,\omega)F(i\omega)e^{i\omega t}d\omega. \quad (31)$$

In practice it is usually desirable to take $n = 1$.

Now the infinite integral

$$D(t) = \int_{-\infty}^{\infty} R_n(\omega_c,\,\omega)F(i\omega)e^{i\omega t}d\omega \qquad (32)$$

must be kept small if the finite series in (31) is to be an accurate representation of the current I. While it is not in general computable, we see that, in order to keep it small, $R_n(\omega_c,\,\omega)$ must be small over the essential range of frequencies of $F(i\omega)$. In cases of practical importance we shall find (see Appendix 1) this range is from $\omega = -\lambda$ to $\omega = +\lambda$.

If the transducer introduces a large phase shift, the linear part of which is predominant in the neighborhood of $\omega = \omega_c$, it is preferable to express the received current I in terms of a "retarded" time. To do this, return to (25) and write

$$\begin{aligned} Y(i\omega_c + i\omega) &= |\,Y(i\omega_c + i\omega)\,|e^{-i\phi}, \\ \phi &= \omega_c\tau + \omega\tau + \beta(\omega) + \theta_c, \\ \beta(0) &= \beta'(0) = 0, \end{aligned} \qquad (33)$$

so that

$$I = E \exp\,(i\omega_c t' + i\theta')\int_{-\infty}^{\infty} |\,Y(i\omega_c + i\omega)\,|e^{-i\beta(\omega)}F(i\omega)e^{i\omega t'}d\omega, \quad (34)$$

where $t' = t - \tau$ is the "retarded" time and $\theta' = \theta - \theta_c$. Formula (34) is identical with (25) but is expressed in the "retarded" time.

Now we can expand the function

$$|\,Y(i\omega_c + i\omega)\,|e^{-i\beta(\omega)}$$

in powers of ω; thus

$$\left(1 + \omega\frac{d}{d\omega_c}\right)|\,Y(i\omega_c)\,| + \sum_2^{\infty} r_n(\omega_c)\omega^n,$$

where

$$r_n(\omega_c) = \frac{1}{n!}\left\{\frac{\partial^n}{\partial\omega_c{}^n}\,|\,Y(i\omega_c + i\omega)\,|e^{-i\beta(\omega)}\right\}_{\omega=0};$$

and by substitution in (34) we get

$$I = E \exp\left(i \int_0^{t'} \Omega(\tau) d\tau + i\theta' \right) \times \left[\left(1 + \lambda s(t') \frac{d}{d\omega_c} \right) | Y(i\omega_c) | + \sum_2^\infty \frac{r_n}{n!} C_n(t') \right], \quad (35)$$

which corresponds precisely with (29) except that it is expressed in terms of the retarded time t'. If the transducer introduces a large phase delay, (35) may be much more rapidly convergent than (29) and should be employed in preference thereto.

Corresponding to (30) we may write

$$Y(i\omega_c + i\omega) e^{-i\beta(\omega)} = \left(1 + \omega \frac{d}{d\omega_c} \right) | Y(i\omega_c) | + R,$$

which defines the remainder. Then

$$I = E \exp\left(i \int_0^{t'} \Omega d\tau + i\theta' \right) \cdot \left[| Y(i\omega_c) | + \lambda s(t') \frac{d}{d\omega_c} | Y(i\omega_c) | \right] + E \exp\,(i\omega_c t' + i\theta') D(t'), \quad (36)$$

where

$$D(t') = \int_{-\infty}^{\infty} R(\omega_c, \omega) \cdot F(i\omega) e^{-i\omega t'} d\omega. \quad (37)$$

Formulas (36) and (37) correspond precisely with (31) and (32) and the same remarks apply.

II

The foregoing will now be applied to the Theory of Frequency Modulation. A pure frequency modulated wave may be defined as a high frequency wave of constant amplitude, the "instantaneous" frequency of which is varied in accordance with a low frequency signal wave. Thus

$$W = \exp i \left(\omega_c t + \lambda \int_0^t s(t) dt \right) \quad (38)$$

is a pure frequency modulated wave. Here ω_c is the constant carrier frequency and $s(t)$ is the low frequency signal which it is desired to transmit. λ is a real parameter which will be termed the modulation index. The "instantaneous" frequency is then defined as

$$\omega_c + \lambda s(t).$$

It is convenient to suppose that $s(t)$ varies between ± 1; in this case the instantaneous frequency varies between the limits

$$\omega_c \pm \lambda.$$

In all cases it will be postulated that $\lambda \ll \omega_c$.

With the method of producing the frequency modulated wave (38) we are not here concerned beyond stating that it may be gotten by varying the capacity or inductance of a high frequency oscillating circuit by and in accordance with the signal $s(t)$.

Corresponding to (38), the pure *amplitude* modulated wave (carrier suppressed) is of the form

$$s(t)\cdot e^{i\omega_c t}. \tag{39}$$

If the maximum essential frequency in the signal $s(t)$ is ω_a, the wave (39) occupies the frequency band lying between $\omega_c - \omega_a$ and $\omega_c + \omega_a$, so that the band width is $2\omega_a$. In the pure *frequency* modulated wave the "instantaneous" frequency band width is 2λ. In practical applications $\lambda \gg \omega_a$. We shall now examine in more detail the concept of "instantaneous" frequency and the conditions under which it has physical significance.

The instantaneous frequency is, as stated, $\omega_c + \lambda s(t)$; a steady-state analysis is of interest and importance. To this end we suppose $s(t) = \cos \omega t$ so that ω is the frequency of the signal. Then the wave (38) may be written

$$e^{i\omega_c t}\left\{ \cos\left(\frac{\lambda}{\omega}\sin \omega t\right) + i \sin\left(\frac{\lambda}{\omega}\sin \omega t\right)\right\},$$

and, from known expansions,

$$W = \sum_{n=-\infty}^{\infty} J_n(\lambda/\omega) e^{i(\omega_c+n\omega)t}, \tag{40}$$

where J_n is the Bessel function of the first kind. Thus the frequency modulated wave is made up of sinusoidal components of frequencies

$$\omega_c \pm n\omega, \qquad n = 0, 1, 2, \cdots, \infty.$$

If $\lambda/\omega \gg 1$ (the case in which we shall be interested in practice) the terms in the series (40) beyond $n = \lambda/\omega$ are negligible; this follows from known properties of the Bessel functions. In this case the frequencies lie in the range

$$\omega_c \pm n\omega = \omega_c \pm \lambda,$$

which agrees with the result arrived at from the idea of instantaneous frequency. On the other hand, suppose we make λ so small that $\lambda/\omega \ll 1$. Then (40) becomes to a first order

$$e^{i\omega_c t} + \frac{1}{2}\left(\frac{\lambda}{\omega}\right) e^{i(\omega_c+\omega)t} - \frac{1}{2}\left(\frac{\lambda}{\omega}\right) e^{i(\omega_c-\omega)t},$$

so that the frequencies ω_c, $\omega_c + \omega$, $\omega_c - \omega$ are present in the pure frequency modulated wave.

It is possible to generalize the foregoing and build up a formal steady-state theory by supposing that

$$s(t) = \sum_{m=1}^{M} A_m \cos(\omega_m t + \theta_m). \tag{41}$$

On this assumption, it can be shown that the frequency modulated wave (38) is expressible as

$$W = \exp(i\omega_c t) \prod_m \sum_{n=-\infty}^{\infty} J_n(v_m) \exp[in(\omega_m t + \theta_m)], \tag{42}$$

$$v_m = \lambda A_m/\omega_m.$$

The corresponding current is then

$$\exp(i\omega_c t) \prod_m \sum_{n=-\infty}^{\infty} J_n(v_m) Y(i\omega_c + n\omega_m) \exp[in(\omega_m t + \theta_m)]. \tag{43}$$

Formulas (42) and (43) are purely formal and far too complicated for profitable interpretation. Consequently this line of analysis will not be carried farther.[4]

If we compare the pure *frequency* modulated wave, as given by (38), with the pure *amplitude* modulated wave, as given by (39), it will be observed that, in the latter, the low frequency signal $s(t)$, which is ultimately wanted in the receiver, is *explicit* and methods for its detection and recovery are direct and simple. In the pure frequency modulated wave, on the other hand, the low frequency signal is *implicit;* indeed it may be thought of as concealed in minute phase or frequency variations in the high frequency carrier wave.

If we differentiate (38) with respect to time t, we get

$$dW/dt = [\omega_c + \lambda s(t)] \exp\left(i\omega_c t + i\lambda \int_0^t s dt\right). \tag{44}$$

[4] See Appendix 1.

The first term,

$$\omega_c \exp\left(i\omega_c t + i\lambda \int_0^t s dt \right), \tag{45}$$

is still a pure frequency modulated wave. The second term,

$$\lambda s(t)\cdot \exp\left(i\omega_c t + i\lambda \int_0^t s dt \right), \tag{46}$$

is a "hybrid" modulated wave, since it is modulated with respect to both *amplitude* and *frequency*. The important point to observe is that, by differentiation, we have "rendered explicit" the wanted low frequency signal. We infer from this that the detection of a pure frequency modulated wave involves in effect its differentiation. The process of rendering explicit the low frequency signal has been termed "frequency detection." Actually it converts the *pure frequency* modulated wave into a *hybrid* modulated wave.

Every frequency distorting transducer inherently introduces frequency detection or "hybridization" of the pure frequency-modulated wave, as may be seen from formula (16). The transmitted current is conveniently written in the form

$$I = Y(i\omega_c) \exp\left(i \int_0^t \Omega dt \right)\cdot\left\{ 1 + \frac{1}{\omega_1}\lambda s + \frac{1}{2\,!}\frac{1}{\omega_2{}^2} C_2 + \frac{1}{3\,!}\frac{1}{\omega_3{}^3} C_3 + \cdots \right\}, \tag{47}$$

where

$$\frac{1}{\omega_n{}^n} = \frac{1}{Y(i\omega_c)}\frac{d^n}{d\omega_c{}^n} Y(i\omega_c). \tag{48}$$

(Note that ω_n has the dimensions of frequency. It may be and usually is complex.)

Every term in (47) except the first, is a hybrid modulated wave.

In passing it is interesting to compare the distortion, as given by (47), undergone by the pure *frequency*-modulated wave, with that suffered by the pure *amplitude*-modulated wave (39), in passing through the same transducer. The transmitted current corresponding to the amplitude-modulated wave (39) is

$$I = Y(i\omega_c) e^{i\omega_c t}\left\{ s(t) + \frac{1}{i\omega_1}\frac{ds}{dt} + \frac{1}{2\,!(i\omega_2)^2}\frac{d^2 s}{dt^2} + \frac{1}{3\,!(i\omega_3)^3}\frac{d^3 s}{dt^3} + \cdots \right\}. \tag{49}$$

This equation corresponds to (47) for the pure frequency-modulated wave.

III

In this section we consider the recovery of the wanted low frequency signal $s(t)$ from the frequency-modulated wave. This involves two distinct processes: (1) rendering explicit the low frequency signal "implicit" in the high frequency wave; that is, "frequency detection" or "hybridization" of the high frequency wave; and (2) detection proper.

It is convenient and involves no loss of essential generality to suppose that the transducer proper is equalized in the neighborhood of the carrier frequency ω_c; that is,

$$\frac{d}{d\omega_c} Y(i\omega_c), \qquad \frac{d^2}{d\omega_c^2} Y(i\omega_c), \cdots \tag{50}$$

are negligible.

Frequency detection is then effected by a terminal network. We therefore take as the over-all transfer admittance

$$y(i\omega) \cdot Y(i\omega). \tag{51}$$

$y(i\omega)$ represents the terminal receiving network; it is under control and can be designed for the most efficient performance of its function. As we shall see, it should approximate as closely as possible a pure reactance.

Taking the over-all transfer admittance as (51), we have from (47),

$$I = y(i\omega_c) Y(i\omega_c) \cdot \exp\left(i \int_0^t \Omega dt \right) \times \left\{ 1 + \frac{1}{\omega_1} \lambda s + \frac{1}{2!\omega_2^2} C_2 + \frac{1}{3!\omega_3^3} C_3 + \cdots \right\}, \tag{52}$$

where now

$$1/\omega_n^n = \frac{1}{y(i\omega_c)} \frac{d^n}{d\omega_c^n} y(i\omega_c). \tag{53}$$

Inspection of (52) shows that the terms beyond the second simply represent distortion. The terminal network or frequency detector should be so designed as to make the series

$$1 + \frac{\lambda}{\omega_1} + \left(\frac{\lambda}{\omega_2}\right)^2 + \left(\frac{\lambda}{\omega_3}\right)^3 + \cdots$$

rapidly convergent from the start.[5] In fact the ideal frequency detector is a network whose admittance $y(i\omega)$ can be represented with

[5] See note at end of this section (p. 528) for specific example.

sufficient accuracy in the neighborhood of $\omega = \omega_c$ by the expression

$$y(i\omega) = y(i\omega_c)\left(1 + \frac{\omega - \omega_c}{\omega_1}\right). \tag{53a}$$

This approximation should be valid over the frequency range from $\omega = \omega_c - \lambda$ to $\omega = \omega_c + \lambda$.

Supposing that this condition is satisfied, the wave, after passing over the transducer and through the terminal frequency detector, is (omitting the constant $y \cdot Y$)

$$I = \left(1 + \frac{\lambda}{\omega_1} s(t)\right) \cdot \exp\left(i \int_0^t \Omega dt\right). \tag{54}$$

If y is a pure reactance, ω_1 is a pure real; due to unavoidable dissipation it will actually be complex. To take this into account we replace ω_1 in (54) by $\omega_1 e^{-i\alpha}$ where now ω_1 is real; (54) then becomes

$$I = \left\{1 + \frac{\lambda}{\omega_1} \cos\alpha \cdot s(t) + i\frac{\lambda}{\omega_1} \sin\alpha \cdot s(t)\right\} \exp\left(i \int_0^t \Omega dt\right). \tag{55}$$

The amplitude A of this wave is then

$$A = \left\{\left(1 + \frac{\lambda}{\omega_1} \cos\alpha \cdot s(t)\right)^2 + \left(\frac{\lambda}{\omega_1} \sin\alpha \cdot s(t)\right)^2\right\}^{1/2}. \tag{56}$$

Now let λ/ω_1 be *less than unity* and let the wave (55) be impressed on a straight-line rectifier. Then the rectified or detected output is

$$\left(1 + \frac{\lambda}{\omega_1} \cos\alpha \cdot s(t)\right)\left\{1 + \left(\frac{\lambda \sin\alpha \cdot s(t)}{\omega_1 + \lambda \cos\alpha \cdot s(t)}\right)^2\right\}^{1/2}, \tag{57}$$

or, to a first order,

$$1 + \frac{\lambda}{\omega_1} \cos\alpha \cdot s(t) + \frac{1}{2}\frac{\lambda^2}{\omega_1^2} \sin^2\alpha \cdot s^2(t). \tag{58}$$

The second term is the recovered signal and the third term is the first order non-linear distortion.

Inspection of the foregoing formulas shows at once that, for detection by straight rectification, the following conditions should be satisfied:

(1) λ/ω_1 *must* be less than unity.

(2) The terminal network should be as nearly as possible a pure reactance to make the phase angle α as nearly zero as possible.

(3) To minimize both linear and non-linear distortion it is necessary that the sequence

$$\frac{\lambda}{\omega_1}, \quad \left(\frac{\lambda}{\omega_2}\right)^2, \quad \left(\frac{\lambda}{\omega_3}\right)^3, \cdots$$

be rapidly convergent from the start.

The first term of (58) is simply direct current and has no significance as regards the recovered signal. When we come to consider the problem of noise in the next section, we shall find that its elimination is important. This can be effected by a scheme which may be termed *balanced rectification.* Briefly described the scheme consists in terminating the transducer in two frequency detectors y_1 and y_2 in parallel; these are so adjusted that $y_1(i\omega_c) = - y_2(i\omega_c)$ and $dy_1/d\omega_c = dy_2/d\omega_c$. ω_1 is therefore of opposite sign in the two frequency detectors. The rectified outputs of the two parallel circuits are then differentially combined in a common low frequency circuit. Corresponding to (58), the resultant detected output is

$$2\frac{\lambda}{\omega_1}\cos\alpha\cdot s(t). \tag{59}$$

This arrangement therefore eliminates first order non-linear distortion, as well as the constant term.

Rectification is the simplest and most direct mode of detection of frequency-modulated waves. However, in connection with the problem of noise reduction other methods of detection will be considered.

Note

As a specific example of the foregoing let the terminal frequency detector, specified by the admittance $y(i\omega)$, be an oscillation circuit consisting simply of an inductance L in series with a capacitance C. Then

$$y(i\omega) = i\sqrt{\frac{C}{L}}\frac{\omega/\omega_R}{1 - \omega^2/\omega_R^2},$$

where $\omega_R^2 = 1/LC$.

Then, if ω_c/ω_R is nearly equal to unity, that is, if

$$\omega_R = (1 + \delta)\omega_c,$$
$$|\delta| \ll 1,$$

we have approximately,

$$\frac{1}{\omega_n{}^n} \doteq \frac{n!}{(\omega_R - \omega_c)^n},$$

$$y(i\omega_c) \doteq \frac{i}{2}\frac{\sqrt{C/L}}{\omega_R - \omega_c}.$$

Formula (42) thus becomes

$$I = y(i\omega_c)\cdot Y(i\omega_c)\cdot \exp\left(i\int_0^t \Omega dt\right)\cdot\left\{1 + \frac{\lambda s}{\omega_R - \omega_c} + \frac{C_2}{(\omega_R - \omega_c)^2} + \frac{C_3}{(\omega_R - \omega_c)^3} + \cdots\right\}.$$

In order that the distortion shall be small it is necessary that

$$\lambda \ll |\omega_R - \omega_c|.$$

If the two networks y_1 and y_2 are oscillation circuits so adjusted that

$$\begin{aligned} C_1/L_1 &= C_2/L_2, \\ \omega_{R_1} &= (1 + \delta)\omega_c = 1/\sqrt{L_1C_1}, \\ \omega_{R_2} &= (1 - \delta)\omega_c = 1/\sqrt{L_2C_2}, \end{aligned}$$

then the combined rectified output of the two parallel circuits is proportional to

$$\frac{\lambda s}{\delta\cdot\omega_c} + \frac{C_3}{(\delta\cdot\omega_c)^3} + \frac{C_5}{(\delta\cdot\omega_c)^5} + \cdots.$$

Thus the constant term and the first order distortion are eliminated in the low frequency circuit.

IV

The most important advantage known at present of *frequency*-modulation, as compared with *amplitude*-modulation, lies in the possibility of substantial reduction in the low frequency noise-to-signal power ratio in the receiver. Such reduction requires a correspondingly large increase in the width of the high frequency transmission band. For this reason frequency-modulation appears to be inherently restricted to short wave transmission.

In the discussion of the theory of noise which follows, it is expressly assumed *that the high frequency noise is small compared with the high frequency signal wave.* Also ideal terminal networks, filters and detectors are postulated.

In view of the assumption of a low noise power level, the calculation of the low frequency noise power in the receiver proper can be made to depend on the calculation of the noise due to the typical high frequency noise element

$$A_n \exp(i\omega_c t + i\omega_n t + i\theta_n). \qquad (60)$$

Corresponding to the noise element (60), the output of the ideal frequency detector is

$$\exp\left(i\int_0^t \Omega dt\right)\cdot\left\{1+\frac{\lambda s}{\omega_1}+\left(1+\frac{\omega_n}{\omega_1}\right)A_n\exp\left(i\omega_n t+i\theta_n - i\lambda\int_0^t sdt\right)\right\}. \tag{61}$$

Since the expression

$$\exp\left(i\omega_n t+i\theta_n-i\lambda\int_0^t sdt\right)$$

occurs so frequently in the analysis which is to follow, it is convenient to adopt the notation

$$\begin{aligned}\Omega_n &= \omega_n - \lambda s(t),\\ \int_0^t \Omega_n dt &= \omega_n t - \lambda\int_0^t sdt.\end{aligned} \tag{61a}$$

With this notation and on the assumption that $A_n \ll 1$ and ω_1 real, the amplitude of the wave (61) is

$$1+\frac{\lambda s}{\omega_1}+\left(1+\frac{\omega_n}{\omega_1}\right)A_n\cos\left(\int_0^t \Omega_n dt\right). \tag{62}$$

In this formula the argument of the cosine function should be strictly

$$\int_0^t \Omega_n dt+\theta_n.$$

The phase angle θ_n is random however and does not affect the final formulas; it may therefore be omitted at the outset. Consequently, if the wave (61) is passed through a straight line rectifier, the rectified or low frequency current is proportional to

$$\lambda s(t)+(\omega_1+\omega_n)A_n\cos\left(\int_0^t \Omega_n dt\right). \tag{63}$$

The first term is the recovered signal and the second term the low frequency noise or interference corresponding to the high frequency element (60).

Now the wave (63), before reaching the receiver proper, is transmitted through a low-pass filter, which cuts off all frequencies above ω_a; ω_a is the highest essential frequency in the signal $s(t)$. Consequently, in order to find the noise actually reaching the receiver proper, it is necessary in one way or another to make a frequency analysis of the wave (63). This is done in Appendix 2, attached hereto, where however, instead of dealing with the special formula (63), a more general expression

$$\lambda s(t) + (\omega_1 + \omega_n + \mu s)A_n \cos \int_0^t \Omega_n dt, \tag{64}$$

is used for the low frequency current. This will be found to include, as special cases, several other important types of rectification, as well as amplitude limitation, which we shall wish to discuss later.[6] Then, subject to the limitation that the noise energy is uniformly distributed over the spectrum, it is shown in Appendix 2 that

$$P_S = \lambda^2 \overline{s^2}, \tag{65}$$

$$P_N = (\tfrac{1}{3}\omega_a^2 + \omega_1^2 + (1 + \nu)^2 \lambda^2 \overline{s^2})\omega_a N^2, \tag{66}$$

$$\nu = \mu/\lambda, \tag{67}$$

$$N^2 = \text{mean high frequency power level.}$$

These formulas are quite important because they make the calculation of low frequency noise-to-signal power ratio very simple for all the modes of frequency detection and demodulation which we shall discuss.

Applying them to formula (63) we find for *straight line rectification*

$$P_N = (\tfrac{1}{3}\omega_a^2 + \omega_1^2 + \lambda^2 \overline{s^2})\omega_a N^2, \tag{68}$$

$$P_S = \lambda^2 \overline{s^2}.$$

It is known that in practice $\omega_1^2 \gg \lambda^2 \overline{s^2}$ and $\lambda^2 \overline{s^2} \gg \omega_a^2$. Consequently in the factor $(\frac{1}{3}\omega_a^2 + \omega_1^2 + \lambda^2 \overline{s^2})$ the largest term is ω_1^2. Therefore it is important, if possible, to eliminate this term. This can be effected by the scheme briefly discussed at the close of section III; parallel rectification and differential recombination. For this scheme the low frequency current is found to be proportional to

$$\lambda s + \omega_n A_n \cos \left(\int_0^t \Omega_n dt \right). \tag{69}$$

Consequently, for *parallel rectification* and *differential recombination*,

$$P_N = (\tfrac{1}{3}\omega_a^2 + \lambda^2 \overline{s^2})\omega_a N^2. \tag{70}$$

[6] The formula is also general enough to include detection by a product modulator, which however is not discussed in the text as no advantage over linear rectification was found.

Here, in the factor $(\frac{1}{3}\omega_a^2 + \lambda^2\overline{s^2})$, the term $\lambda^2\overline{s^2}$ is predominant. The elimination of the term ω_1^2 has resulted in a substantial reduction in the noise power.

Returning to the general formula (66) for P_N, it is clear, that, if in addition to eliminating the term ω_1^2, the parameter $\nu = \mu/\lambda$ can be made equal to -1, the noise power will be reduced to its lowest limits:

$$P_N = \tfrac{1}{3}\omega_a^3 N^2.$$

This highly desirable result can be effected by *amplitude limitation*, the theory of which will now be discussed.

V

When amplitude limitation is employed in frequency-modulation, the incoming high frequency signal is drastically reduced in amplitude. If no interference is present this merely results in an equal reduction in the low frequency recovered signal which is *per se* undesirable. When, however, noise or interference is present, amplitude limitation prevents the interference from affecting the *amplitude* of the resultant high frequency wave; its effect then can appear only as *variations in the phase or instantaneous frequency* of the high frequency wave. To this fact is to be ascribed the potential superiority of *frequency-modulation* as regards the reduction of noise power. This superiority is only possible with wide band high frequency transmission; that is, the index of frequency-modulation λ must be large compared with the low frequency band width ω_a. Insofar as the present paper is concerned, the potential superiority of frequency-modulation with amplitude limitation is demonstrated only for the case where the high frequency noise is small compared with the high frequency signal wave.

If, to the frequency-modulated wave $\exp\left(i\int_0^t \Omega dt\right)$, there is added the typical noise element $A_n \exp(i\omega_c + i\omega_n t + \theta_n)$, the resultant wave may be written as

$$\exp\left(i\int_0^t \Omega dt\right)\cdot\left(1 + A_n \exp\left(i\int_0^t \Omega_n dt\right)\right). \qquad (71)$$

Postulating that $A_n \ll 1$ and therefore neglecting terms in A_n^2, the real part of (71) is

$$\left(1 + A_n \cos\left(\int_0^t \Omega_n dt\right)\right)\cdot\cos\left(\int_0^t \Omega dt + A_n \sin\left(\int_0^t \Omega_n dt\right)\right). \qquad (72)$$

If this wave is subjected to amplitude limitation, the amplitude variation is suppressed, leaving a pure frequency-modulated wave, *proportional* to the *real part* of

$$\exp\left[i\left(\int_0^t \Omega dt + A_n \sin\left(\int_0^t \Omega_n dt\right)\right)\right] \tag{73}$$

(but drastically reduced in amplitude).

After frequency detection the wave (73) is, within a constant,

$$\exp\left[\left(i\int_0^t \Omega dt + A_n \sin\left(\int_0^t \Omega_n dt\right)\right)\right] \\ \times\left[1+\frac{1}{\omega_1}\frac{d}{dt}\left(\lambda\int_0^t s dt + A_n \sin\left(\int_0^t \Omega_n dt\right)\right)\right]. \tag{74}$$

Consequently, since

$$\int_0^t \Omega_n dt = \omega_n t + \theta_n - \lambda\int_0^t s dt, \tag{75}$$

the amplitude of the wave (74) is

$$1+\frac{1}{\omega_1}\left\{\lambda s + (\omega_n - \lambda s)A_n \cos\left(\int_0^t \Omega_n dt\right)\right\}. \tag{76}$$

This is the amplitude of the low frequency wave after rectification; it is obviously proportional to

$$\lambda s + (\omega_n - \lambda s)A_n \cos\left(\int_0^t \Omega_n dt\right), \tag{77}$$

which is a special case of (64) and may be used in calculating the relative signal and noise power with amplitude limitation. Hence we have, by aid of (65) and (66),

$$\begin{aligned} P_S &= \lambda^2\overline{s^2}, \\ P_N &= \tfrac{1}{3}\omega_a^3 N^2. \end{aligned} \tag{78}$$

(These are, of course, relative values and take no account of the absolute reduction in power due to amplitude limitation.)

Comparing (78) with (68) it is seen that, for detection by straight line rectification, the ratio of the noise power *with* to that *without* amplitude limitation is

$$\frac{1}{1+3\omega_1^2/\omega_a^2+3\lambda^2\overline{s^2}/\omega_a^2}; \tag{79}$$

or taking $\overline{s^2} = 1/2$,

$$\frac{1}{1 + 3\omega_1^2/\omega_a^2 + 3\lambda^2/2\omega_a^2}. \tag{80}$$

Since in practice $\omega_1 \gg \omega_a$ and $\lambda \gg \omega_a$, amplitude limitation results in a very substantial reduction in low frequency noise power in the receiver proper. Reference to formula (70) shows that, as compared with parallel rectification and recombination, amplitude limitation reduces the noise power by the factor

$$\frac{1}{1 + 3\lambda^2/2\omega_a^2}. \tag{81}$$

It should be observed that *without* amplitude limitation little reduction in the noise-to-signal power ratio results from increasing the modulation index λ (and consequently the high frequency transmission band width). On the other hand, *with* amplitude limitation, the ratio ρ of noise-to-signal power is

$$\rho = P_N/P_S = \frac{2}{3}\left(\frac{\omega_a}{\lambda}\right)^2 \omega_a N^2. \tag{82}$$

The ratio ρ is then (within limits) inversely proportional to the square of the modulation index λ, so that a large value of λ is indicated. It should be noted that, within limits ($\lambda \ll \omega_c$), the power transmitted from the sending station is independent of the modulation index λ.

It might be inferred from formula (82) that the noise power ratio ρ can be reduced indefinitely by indefinitely increasing the modulation index λ. Actually there are practical limits to the size of λ. First, if λ is made sufficiently large, the variable frequency oscillator generating the frequency-modulated wave may become unstable or function imperfectly. Secondly, the frequency spread of the frequency modulated wave is 2λ (from $\omega_c - \lambda$ to $\omega_c + \lambda$) and, if this is made too large, interference with other stations will result. Finally, the stationary distortion of the recovered low frequency signal $s(t)$ increases rapidly with the size of λ.

To summarize the results of the foregoing analysis the potential advantages of frequency-modulation depend on two facts. (1) By increasing the modulation index λ it is possible to increase the recovered low frequency signal power at the receiving station without increasing the high frequency power transmitted from the sending station. (2) It is possible to employ amplitude limitation (inherently impossible with amplitude-modulation) whereby the effect of interference or noise is reduced to a phase or "instantaneous frequency" variation of the high frequency wave.

Appendix 1

Formula (40) *et sequa* establish the fact that the actual frequency of the wave (29) varies between the limits

$$\omega_c \pm \lambda$$

provided $s(t)$ is a pure sinusoid $\lambda \sin \omega t$ and $\lambda \gg \omega$. This agrees with the concept of instantaneous frequency.

When $s(t)$ is a complex function—say a Fourier series—the frequency range of W can be determined qualitatively under certain restrictions, as follows:

We write

$$W = \exp\left(i\omega_c t + i\lambda \int_0^t s dt \right) \tag{1a}$$

$$= e^{i\omega_c t} \int_{-\infty}^{\infty} F(i\omega) e^{i\omega t} d\omega. \tag{2a}$$

The Fourier formulation is supposed to be valid in the epoch $0 \leq t \leq T$ and T can be made as great as desired. Then

$$F(i\omega) = \pi \int_0^T \exp\left(i\lambda \int_0^t s dt - i\omega t \right) dt. \tag{3a}$$

We now suppose that, in the epoch $0 \leq t \leq T$,

$$\left| \lambda \int_0^T s dt \right| \tag{4a}$$

becomes very large compared with 2π. On this assumption, it follows from the Principle of Stationary Phase, that, for a fixed value of ω, the important contributions to the integral (3a) occur for those values of the integration variable t for which

$$\frac{d}{dt}\left(\lambda \int_0^t s dt - \omega t \right) = 0,$$

or

$$\omega = \lambda s(t).$$

Consequently the important part of the spectrum $F(i\omega)$ corresponds to those values of ω in the range

$$\lambda s_{\min} \leq \omega \leq \lambda s_{\max}.$$

Therefore the frequency spread of W lies in the range from $\omega_c + \lambda s_{\min}$ to $\omega_c + \lambda s_{\max}$ or $\omega_c \pm \lambda$ if $s_{\max} = - s_{\min} = 1$.

APPENDIX 2

We take the frequency modulated wave as

$$\cos\left(\omega_c t + \lambda \int_0^t s dt\right), \tag{1b}$$

where ω_c is the carrier frequency and $s = s(t)$ is the low frequency signal. λ is a real parameter, which fixes the amplitude of the frequency spread.

Correspondingly, we take the typical noise element as

$$A_n \cos\left((\omega_c + \omega_n)t + \theta_n\right). \tag{2b}$$

For reasons stated in the text, we take the more general formula for the low frequency current as proportional to

$$\lambda s + (\omega_0 + \omega_n + \mu s)A_n \cos\left(\omega_n t + \theta_n - \lambda \int_0^t s dt\right), \tag{3b}$$

where ω_0, λ, μ are real parameters. The term λs is the recovered signal and the second term is the low frequency noise corresponding to the high frequency noise element (2*b*).

We suppose that the noise is uniformly distributed over the frequency spectrum, at least in the neighborhood of $\omega = \omega_c$, so that, corresponding to the noise element

$$A_n \cos\left(\omega_n t + \theta_n\right), \tag{4b}$$

the noise is representable as the Fourier integral

$$\frac{N}{\pi}\int \cos\left(\omega_n t + \theta_n\right) d\omega_n \tag{5b}$$

and the corresponding *noise power* for the frequency interval $\omega_1 < \omega_n < \omega_2$ is, by the Fourier integral energy theorem,

$$\frac{N^2}{\pi}\int_{\omega_1}^{\omega_2} d\omega_n = \frac{1}{\pi}(\omega_2 - \omega_1)N^2. \tag{6b}$$

The Fourier integral energy theorem states that, if in the epoch $0 \leq t \leq T$, the function $f(t)$ is representable as the Fourier integral

$$f(t) = \frac{1}{\pi}\int_0^\infty F(\omega)\cdot\cos\left(\omega t + \theta(\omega)\right) d\omega, \tag{7b}$$

then

$$\int_0^T f^2 dt = \frac{1}{\pi}\int_0^\infty F^2 d\omega.^7 \tag{8b}$$

Replacing (4*b*) by (5*b*) to take care of the distributed noise, the noise term of (3*b*) becomes

$$\cos\left(\lambda\int_0^t s dt\right)\cdot\frac{N}{\pi}\int(\omega_0+\omega_n+\mu s)\cdot\cos(\omega_n t+\theta_n)d\omega_n$$
$$+\sin\left(\lambda\int_0^t s dt\right)\cdot\frac{N}{\pi}\int(\omega_0+\omega_n+\mu s)\cdot\sin(\omega_n t+\theta_n)d\omega_n. \tag{9b}$$

Now this noise in the low frequency circuit is passed through a low pass filter, which cuts off all frequencies above ω_a. ω_a is the maximum essential frequency in the signal $s(t)$.

It is therefore necessary to express (9*b*) as a frequency function before calculating the noise power. To this end we write the Fourier integrals

$$\cos\left(\lambda\int_0^t s dt\right)=\frac{1}{\pi}\int_0^\infty F_c\cos(\omega t+\theta_c)d\omega, \tag{10b}$$

$$\sin\left(\lambda\int_0^t s dt\right)=\frac{1}{\pi}\int_0^\infty F_s\sin(\omega t+\theta_s)d\omega. \tag{11b}$$

We note also that

$$\mu s\cdot\cos\left(\lambda\int_0^t s dt\right)=\frac{\mu}{\lambda}\frac{d}{dt}\sin\left(\lambda\int_0^t s dt\right)$$
$$=\frac{1}{\pi}\int_0^\infty\frac{\mu\omega}{\lambda}F_s\cos(\omega t+\theta_s)d\omega, \tag{12b}$$

$$\mu s\cdot\sin\left(\lambda\int_0^t s dt\right)=-\frac{\mu}{\lambda}\frac{d}{dt}\cos\left(\lambda\int_0^t s dt\right)$$
$$=\frac{1}{\pi}\int_0^\infty\frac{\mu\omega}{\lambda}F_c\sin(\omega t+\theta_c)d\omega. \tag{13b}$$

Substituting (10*b*), (11*b*), (12*b*) and (13*b*) in (9*b*) and carrying through straightforward operations, we find that the noise is given by

$$\frac{N}{2\pi^2}\int_0^\infty F_p d\omega\int_{\omega-\omega_a}^{\omega+\omega_a}\left(\omega_0+\omega_n+\frac{\mu}{\lambda}\omega\right)\cos((\omega-\omega_n)t+\theta_p)d\omega_n$$
$$+\frac{N}{2\pi^2}\int_0^\infty F_m d\omega\int_{-(\omega+\omega_a)}^{-(\omega-\omega_a)}\left(\omega_0+\omega_n-\frac{\mu}{\lambda}\omega\right)\cos((\omega+\omega_n)t+\theta_m)d\omega_n, \tag{14b}$$

[7] See "Transient Oscillations in Electric Wave Filters," Carson and Zobel, *B. S. T. J.*, July, 1923.

where

$$F_p^2 = F_c^2 + F_s^2 + 2F_cF_s \cos(\theta_c - \theta_s), \qquad (15b)$$

$$F_m^2 = F_c^2 + F_s^2 - 2F_cF_s \cos(\theta_c - \theta_s). \qquad (16b)$$

The limits of integration of ω_n are determined by the fact that, $\omega - \omega_n$ in the first integral of (14*b*) and $\omega + \omega_n$ in the second, must lie between $\pm\,\omega_a$; all other frequencies are eliminated by the low pass filter.

From formula (14*b*) and the Fourier integral energy theorem, the *noise power* P_N is given by

$$P_N = \frac{N^2}{4\pi^3T}\int_0^\infty F_p^2 d\omega \int_{\omega-\omega_a}^{\omega+\omega_a}\left(\omega_0 + \omega_n + \frac{\mu}{\lambda}\omega\right)^2 d\omega_n$$
$$+ \frac{N^2}{4\pi^3T}\int_0^\infty F_m^2 d\omega \int_{-(\omega+\omega_a)}^{-(\omega-\omega_a)}\left(\omega_0 + \omega_n - \frac{\mu}{\lambda}\omega\right)^2 d\omega_n. \qquad (17b)$$

Integrating with respect to ω_n, we have

$$P_N = \frac{N^2\omega_a}{2\pi^3T}\int_0^\infty d\omega\{[(\omega_0 + (1+\nu)\omega)^2 + \tfrac{1}{3}\omega_a^2]F_p^2$$
$$+ [(\omega_0 - (1+\nu)\omega)^2 + \tfrac{1}{3}\omega_a^2]F_m^2\}, \qquad (18b)$$

where $\nu = \mu/\lambda$.

Replacing F_p^2 and F_m^2 in (18*b*) by their values as given by (15*b*) and (16*b*), we get

$$P_N = \frac{\omega_aN^2}{\pi^3T}\int_0^\infty (\omega_0^2 + (1+\nu)^2\omega^2 + \tfrac{1}{3}\omega_a^2)(F_c^2 + F_s^2)d\omega$$
$$+ 4\frac{\omega_aN^2}{\pi^3T}\int_0^\infty (1+\nu)\omega_0\omega F_cF_s \cos(\theta_c - \theta_s)d\omega. \qquad (19b)$$

To evaluate (19*b*) we make use of the formulas, derived below

$$\frac{1}{\pi T}\int_0^\infty (F_c^2 + F_s^2)d\omega = 1, \qquad (20b)$$

$$\frac{1}{\pi T}\int_0^\infty \omega^2(F_c^2 + F_s^2)d\omega = \lambda^2\overline{s^2} = P_S, \qquad (21b)$$

$$\frac{1}{\pi T}\int_0^\infty \omega F_cF_s \cos(\theta_c - \theta_s)d\omega \to 0 \text{ as } T \to \infty. \qquad (22b)$$

Substitution of (20*b*), (21*b*), (22*b*) in (19*b*) gives for large values of T

$$P_N = (\tfrac{1}{3}\omega_a^2 + \omega_0^2 + (1+\nu)^2\lambda^2\overline{s^2})\omega_aN^2. \qquad (23b)$$

Here, for convenience, we have replaced N^2/π^2 of (19*b*) by N^2, so that N^2 of (23*b*) may be defined and regarded as the high frequency *noise power level.*

It remains to establish formulas (20*b*), (21*b*) and (22*b*). From the defining formulas (10*b*) and (11*b*) and the Fourier integral energy theorem, we have

$$\begin{aligned} \frac{1}{\pi T}\int_0^\infty F_c^2 d\omega &= \frac{1}{T}\int_0^T \cos^2\left(\lambda\int_0^t s dt\right)dt, \\ \frac{1}{\pi T}\int_0^\infty F_s^2 d\omega &= \frac{1}{T}\int_0^T \sin^2\left(\lambda\int_0^t s dt\right)dt. \end{aligned} \tag{24b}$$

Adding we get (20*b*).

Now differentiate (10*b*) and (11*b*) with respect to *t* and apply the Fourier integral energy theorem; we get

$$\begin{aligned} \frac{1}{\pi T}\int_0^\infty \omega^2 F_c^2 d\omega &= \frac{1}{T}\int_0^T \lambda^2 s^2 \sin^2\left(\lambda\int_0^t s dt\right)dt, \\ \frac{1}{\pi T}\int_0^\infty \omega^2 F_s^2 d\omega &= \frac{1}{T}\int_0^T \lambda^2 s^2 \cos^2\left(\lambda\int_0^t s dt\right)dt \end{aligned} \tag{25b}$$

and, by addition, we get (21*b*).

To prove (22*b*) we note that

$$\begin{aligned} &(1+\mu s)\cos\left(\lambda\int_0^t s dt\right) \\ &\quad = \cos\left(\lambda\int_0^t s dt\right) + \frac{\mu}{\lambda}\frac{d}{dt}\sin\left(\lambda\int_0^t s dt\right) \\ &\quad = \frac{1}{\pi}\int_0^\infty \left[F_c\cos(\omega t+\theta_c) + \frac{\mu}{\lambda}\omega F_s\cos(\omega t+\theta_s)\right]d\omega \\ &\quad = \frac{1}{\pi}\int_0^\infty \left[F_c^2 + \left(\frac{\mu}{\lambda}\right)^2\omega^2 F_s^2 \right. \\ &\qquad\qquad \left. + 2\frac{\mu}{\lambda}\omega F_c F_s\cos(\theta_c-\theta_s)\right]^{1/2}\cos(\omega t+\Phi)d\omega. \end{aligned} \tag{26b}$$

Consequently, by the Fourier integral energy theorem,

$$\begin{aligned} &\frac{1}{T}\int_0^T (1+\mu s)^2\cos^2\left(\lambda\int_0^t s dt\right)dt \\ &\quad = \frac{1}{\pi T}\int_0^\infty \left[F_c^2 + \left(\frac{\mu}{\lambda}\right)^2\omega^2 F_s^2 + 2\frac{\mu}{\lambda}\omega F_c F_s\cos(\theta_c-\theta_s)\right]d\omega \end{aligned} \tag{27b}$$

and

$$\begin{aligned} &\frac{1}{T}\int_0^T \mu s\cdot\cos^2\left(\lambda\int_0^t s dt\right)dt \\ &\qquad\qquad = \frac{1}{\pi T}\left(\frac{\mu}{\lambda}\right)\int_0^\infty \omega F_c F_s\cos(\theta_c-\theta_s)d\omega. \end{aligned} \tag{28b}$$

By simple transformations (28*b*) becomes

$$\begin{aligned}\frac{1}{\pi T}\int_0^\infty \omega F_c F_s \cos\,(\theta_c - \theta_s)d\omega & \\ &= \frac{1}{2T}\int_0^T \lambda s dt + \frac{1}{4T}\int_0^T \frac{d}{dt}\sin\left(2\lambda\int_0^t sdt\right)dt \\ &= \frac{1}{2}\lambda\bar{s} + \frac{1}{4T}\sin\left(2\lambda\int_0^T sdt\right) \\ &\rightarrow 0 \text{ as } T \rightarrow \infty, \qquad (29b)\end{aligned}$$

since by hypothesis $\bar{s} = 0$.

We note for reference that

$$-\frac{1}{\pi T}\int_0^\infty F_c F_s \sin\,(\theta_c - \theta_s)d\omega = \frac{1}{2T}\int_0^T \sin\left(2\lambda\int_0^t sdt\right)dt. \qquad (30b)$$

Bibliography

1. Carson, J. R., "Notes on the Theory of Modulation," *Proc. I. R. E.*, **10**, pp. 57–64, Feb., 1922.
2. Roder, H., "Über Frequenzmodulation," *Telefunken-Zeitung*, **10**, pp. 48–54, Dec., 1929.
3. Heilmann, A., "Einige Betrachtungen zum Problem der Frequenzmodulation," *Elek. Nach. Tech.*, **7**, pp. 217–225, June, 1930.
4. Van der Pol, B., "Frequency Modulation," *Proc. I. R. E.*, **18**, pp. 1194–1205, July, 1930.
5. Eckersley, T. L., "Frequency Modulation and Distortion," *Exp. Wireless and Wireless Engg.*, **7**, pp. 482–484, Sept., 1930.
6. Runge, W., "Untersuchungen an amplituden- und frequenz-modulierten Sendern," *Elek. Nach. Tech.*, **7**, pp. 488–494, Dec., 1930.
7. Roder, H., "Amplitude, Phase and Frequency-Modulation," *Proc. I. R. E.*, **19**, pp. 2145–2175, Dec., 1931.
8. Andrew, V. J., "The Reception of Frequency Modulated Radio Signals," *Proc. I. R. E.*, **20**, pp. 835–840, May, 1932.
9. Barrow, W. L., "Frequency Modulation and the Effects of a Periodic Capacity Variation in a Non-dissipative Oscillatory Circuit," *Proc. I. R. E.*, **21**, pp. 1182–1202, Aug., 1933.
10. Barrow, W. L., "On the Oscillations of a Circuit Having a Periodically Varying Capacitance; Contribution to the Theory of Nonlinear Circuits with Large Applied Voltages," *Proc. I. R. E.*, **22**, pp. 201–212, Feb., 1934, also *M. I. T. Serial* 97, Oct., 1934.
11. Chaffee, J. G., "The Detection of Frequency-Modulated Waves," *Proc. I. R. E.*, **23**, pp. 517–540, May, 1935.
12. Armstrong, E. H., "A Method of Reducing Disturbances in Radio Signaling by a System of Frequency-Modulation," *Proc. I. R. E.*, **24**, pp. 689–740, May, 1936.
13. Crosby, M. G., "Frequency Modulation Propagation Characteristics," *Proc. I. R. E.*, **24**, pp. 898–913, June, 1936.
14. Crosby, M. G., "Frequency-Modulated Noise Characteristics," *Proc. I. R. E.*, **25**, pp. 472–514, April, 1937.
15. Roder, H., "Noise in Frequency Modulation," *Electronics*, **10**, pp. 22–25, 60, 62, 64, May, 1937.

Irregularities in Broad-Band Wire Transmission Circuits

By PIERRE MERTZ and K. W. PFLEGER

The effects of inhomogeneities along the length of a wire transmission circuit are considered, affecting its use as a broad-band transmission medium. These inhomogeneities give rise to reflections of the transmitted energy which in turn cause irregularities in the measured sending or receiving end impedance of the circuit in its overall attenuation, and in its envelope delay. The irregularities comprise departures of the characteristic from the average, in an ensemble of lines, or departures from a smooth curve of the characteristic of a single line when this is plotted as a function of frequency. These irregularities are investigated quantitatively.

WIRE transmission circuits in their elementary conception are considered as perfectly uniform or homogeneous from end to end. Actually, of course, they are manufactured in comparatively short pieces and joined end to end, and there is a finite tolerance in the deviation of the characteristics of one piece from those of the next and also from one part of the same piece to another. A real transmission circuit therefore has a large number of irregularities scattered along its length which reflect wavelets back and forth when it is used for the propagation of a signal wave. When a cable pair, coaxial conductor, or similar medium is used for broad-band transmission it is important to know how these irregularities influence the transmission characteristics of the medium.

The transmission characteristics which will be studied are the impedance, the attenuation, the sinuosity of the attenuation (to be defined), and the delay distortion. The derivations for the first two characteristics parallel substantially those published by Didlaukis and Kaden (ENT, vol. 14, p. 13, Jan., 1937). They are set forth here for completeness of presentation because the steps in them illustrate the more complicated steps in the derivation of the last two characteristics.

When the characteristic impedance changes from point to point, its variation from the average characteristic impedance for the whole length of conductor forms the irregularities which produce reflections. Assume that successive discrete elementary pieces of the circuit are homogeneous throughout their length, that the lengths of these elementary pieces are equal throughout the length of the whole circuit, and that there is no correlation between the deviations from average characteristic impedance of any two elementary pieces. This represents a first approximation to the problem. It is fairly accurate for pairs in ordinary cable in which the outstanding irregularities are deviations, from the average, between whole reel lengths; and in which the lengths of the successive spliced pieces (reel lengths) are at least roughly the same.

There are irregularities in some coaxial conductors in which the impedance change is gradual rather than abrupt from one element to the next, and in which the elements can vary in length along the line. For these cases the approximation is a little over-simplified. However, although this somewhat affects the echo wavelets as computed from the impedance deviations along the line, Didlaukis and Kaden, as referred to above, have shown that it does not affect the ratio between the echo wavelets, suitably averaged, reaching the receiving end and those, similarly averaged, returning to the sending end.

With the above assumptions there will be some correlation between the reflections at the two ends of an elementary length. If, for example, this length happens to be high in characteristic impedance the reflection at one end will tend greatly to be the negative of that at the other end. For this reason we are going to break up the reflection into two parts, at a point between any two successive elementary lengths of circuit—one part from one length of the circuit to an infinitesimal length of cable of average characteristics inserted between the two elementary lengths—and the other from this infinitesimal piece to the next elementary length of circuit. There is then 100 per cent correlation between the reflections at the two ends of a given elementary length (one being exactly the negative of the other); but there is no correlation between the reflections from any one elementary length to its adjacent infinitesimal piece of average cable, and the reflections from any other elementary length to its adjacent piece. This same treatment is used in the calculation of certain types of "reflection" crosstalk.

The departure in characteristic impedance in the usual transmitting circuit in the higher frequency range, where the irregularities are most important, results essentially from deviations in the two primary constants of capacitance and inductance, each per unit length. There is a certain correlation between these, inasmuch as the capacitance deviation is contributed to both by differences in the dielectric constant of the insulation and by differences in the geometrical size, shape, and relative arrangement of the conductors; and the inductance deviation is contributed to by the latter alone. If there were no deviation in dielectric constant there would be no deviation in velocity of propagation (phase or envelope), which (at the higher frequencies) is inversely proportional to the square root of the product of the capacitance by the inductance. Consequently the portion of the fractional deviation in capacitance which is due to geometrical deviations correlates with an equal and opposite fractional deviation in inductance. Since in practice the contribution from the geometrical deviation is apt to be dominating, that due to the variation in dielectric constant will be neglected and the above correlation assumed as 100 per cent.

The standard deviation of the capacitance of the successive elementary lengths, as a fraction of the average capacitance, will be designated as δ.

The secondary constant of the line most affected by these irregularities is the sending end (or similarly receiving end) impedance. If we consider a large ensemble of lines of infinite length of similar manufacture (and equal average characteristics and δ) but in which the individual irregularities are uncorrelated, then the sending end impedances of these lines, measured at a given frequency, also form an ensemble. The standard deviation of the real parts in this latter is $\sqrt{\overline{\Delta K_r^2}}$, and that of the imaginary parts $\sqrt{\overline{\Delta K_i^2}}$.

In general, the departure in the impedance of one individual line from the average will vary with frequency; and perhaps over a moderate frequency range a sizeable sample can be collected which is fairly typical of the ensemble of the departures at a fixed frequency in the interval. If this is the case, and if at the same time the average impedance varies smoothly and slowly with frequency, and the standard deviation of the ensemble of departures also varies smoothly and slowly with frequency, then the standard deviation of the sample of departures over the moderate frequency interval is substantially equal to that of the ensemble of departures at a fixed frequency in this interval. (It is clear that this disregards exceptional lines in the ensemble, characterized by regularity in the array of their capacitance deviations, for which these conditions do not hold.) Under the circumstances where this observation is valid it makes it possible to correlate measurements on a single line, provided it is not too exceptional, with theory deduced for an ensemble.

The irregularities in the transmission line will also affect its attenuation. If again we consider an ensemble of lines and measure the attenuation of each at a given frequency these attenuations will also form an ensemble.

It will be found in this case, as will be demonstrated further below, that the average attenuation is a little higher than that of a single completely smooth line having throughout its length a characteristic impedance equal to the average of that for the irregular line. This rise varies slowly with frequency. The standard deviation of the attenuation will also include not only the effect of the reflections which we have been considering but in addition one caused by the fact that the attenuations of the successive elementary pieces are not alike, and hence their sum, aside from any reflections, will also show a distribution. This additional contribution will vary only very slowly with frequency. The standard deviation will be $\sqrt{\overline{\Delta\Lambda_1^2} + \overline{\Delta\Lambda_2^2}}$ where Λ represents the losses in the total line, the subscript 1 indicates the contribution due to the reflections, and the subscript 2 that due to the distribution of the individual attenuations.

The same observation may be made about the attenuation that was made about the terminal impedance, as regards measurements made at one frequency on an ensemble of lines and measurements over a range of frequencies on one line; except that the contribution to the deviation caused by the distribution of individual attenuations varies so slowly with frequency that on each individual line it will look like a displacement from the average attenuation, over the whole frequency range. For the purposes of the present paper only the contributions from the reflections will be computed.

When this information on irregularities is being used by a designer of equalizers he is interested in two characteristics: first, how far each attenuation curve for a number of lines will be displaced as a whole from the average; and second, how "wiggly" each individual curve is likely to be. While the observations above give the general amplitude of the latter they do not tell how closely together in frequency the individual "wiggles" are likely to come. To express this, the term "sinuosity" has been defined as the standard deviation of the difference in attenuation (for the ensemble of lines) at two frequencies separated by a given interval Δf. By the previous observations this can be extended to the attenuation differences for successive frequencies separated by the interval Δf, for a range of frequencies in a single line.

When the transmission line is used for certain types of communication, notably for telephotography or television, it is important to equalize it accurately for envelope delay as well as attenuation. The envelope delay is defined as

$$T = d\beta/d\omega \tag{1}$$

where β is the phase shift through the line and ω is 2π times the frequency. For an ensemble of lines, the envelope delay at a given frequency will also form an ensemble, the standard deviation of which will be $\sqrt{\overline{\Delta T^2}}$. By the observations which have already been made the same standard deviation also holds for the envelope delay departures over a range of frequencies on one line.

Let Fig. 1 represent a line of the type we have been discussing. The successive η's represent the reflection coefficients between successive elementary pieces of line. As mentioned before, to avoid correlation, each η is broken up as shown into two h's, representing reflections between the elementary pieces and infinitesimal lengths of average line.

The main signal transmission will flow as shown by the arrow a in Fig. 1. In addition there will be single reflections as shown by the arrow b. Following the assumptions we have set up, this really consists of two reflections from infinitesimally separated points. Further there will be double reflections, that is reflections of reflections, as shown by c. Here again each reflection point, according to our assumptions, consists of two infinitesimally separated ones. There will be a variety of double reflections according to the number of elementary lengths between reflection points. Finally there will be triple, quadruple and higher order reflections which are not shown. The wave amplitude after reflection is cut down by the reflection coefficient. Consequently, even though there are more of them, the total of any given higher order reflections can always be made smaller than that of lower order reflections by a small enough reflection coefficient. We will here study only small reflection coefficients and therefore neglect all reflections of higher order than needed to give a finite result. For effects on the impedance this means neglect of all but first-order reflections. For the other effects studied it means neglect of all but first- and second-order reflections.

The reflection coefficient between two successive impedances (one being $\overline{K}$), is, approximately

$$h = \Delta K/(2\overline{K}). \tag{2}$$

Following our earlier assumptions, namely that the principal cause of impedance departures lies in geometrical irregularities, and that these may be expressed in terms of capacitance departures,

$$\frac{\Delta K}{\overline{K}} = \frac{\Delta C}{\overline{C}}, \qquad \text{or} \qquad h = \frac{\Delta C}{2\overline{C}}, \qquad \text{or} \qquad \sqrt{\overline{h^2}} = \delta/2. \tag{3}$$

Consequently the reflection coefficients are real, namely, they introduce no phase shifts other than 0 or π in the reflections.

The irregularities in sending-end impedance have been computed in Appendix I from the single reflections of the type b in Fig. 1. The

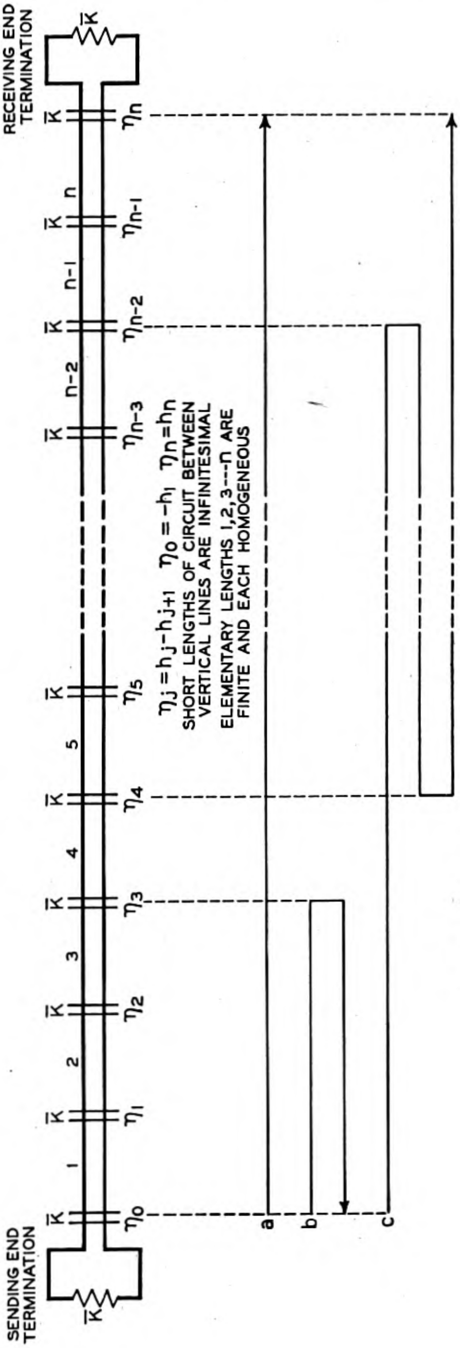


FIG. 1—Inhomogeneous line divided into elementary segments.

final simplified result is

$$\frac{\sqrt{\overline{\Delta K_r^2}}}{\overline{K}} = \frac{\sqrt{\overline{\Delta K_i^2}}}{\overline{K}} = \frac{|\phi|\delta}{2\sqrt{\epsilon}}, \tag{4}$$

where ϕ is the phase shift in radians in two elementary lengths, ϵ is the attenuation in nepers of two elementary lengths, and δ is, as mentioned before, the standard deviation in C measured as a fraction of $\overline{C}$. It will be noted that as a consequence of the single reflections, the irregularities in impedance vary as the first power of δ.

The irregularities in attenuation have been computed in Appendix II from the double reflections of the type c in Fig. 1. It is found, as mentioned before, that there is a net rise in average attenuation caused by the reflections, equal, in nepers, to

$$\left(\epsilon + \frac{\phi^2}{2}\right)\frac{n\delta^2}{4}, \tag{5}$$

where n is the number of elementary lengths in the total line. Considering the factor in parentheses in the expression above, although the term ϵ is not usually wholly negligible compared with the term $\phi^2/2$, nevertheless the latter is dominating and sets the order of magnitude of the factor. If the ϵ is disregarded, the expression can easily be put in terms of the impedance irregularities, giving

$$\left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\overline{K}}\right]^2 \Lambda, \tag{6}$$

where Λ as before represents the loss in the total line.

The standard deviation in the loss in nepers, when finally simplified, is, for the reflections,

$$\sqrt{\overline{\Delta \Lambda_1^2}} = \frac{\phi^2\delta^2\sqrt{n}}{8\sqrt{\epsilon}}. \tag{7}$$

Expressed in terms of the impedance irregularities, this amounts to

$$\sqrt{\overline{\Delta \Lambda_1^2}} = \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\overline{K}}\right]^2 \sqrt{\frac{\Lambda}{2}}. \tag{8}$$

It will be noted that these irregularities in the attenuation vary with the square of δ, or the square of the impedance irregularities. This is a consequence of the double reflections, and will continue to hold for the sinuosity and irregularities in envelope delay. It will also be noted that in this form the equation is independent of ϵ, ϕ, and n. It is in this case that Didlaukis and Kaden found that the result is independent of whether the reflection points are sharp and equally spaced or not.

The sinuosity has been computed in Appendix III. When finally simplified and measured in nepers, it amounts to

$$\sqrt{\overline{(\Delta\Lambda_1 - \overline{\Delta\Lambda_1})^2}} = \frac{\phi^2\delta^2\sqrt{n}}{8\sqrt{2}\epsilon^{3/2}}\frac{d\phi}{df}\Delta f. \tag{9}$$

Expressed in terms of the impedance irregularities this amounts to

$$\sqrt{\overline{(\Delta\Lambda_1 - \overline{\Delta\Lambda_1})^2}} = \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\overline{K}}\right]^2 \frac{\pi T}{\sqrt{\Lambda}}\Delta f, \tag{10}$$

where T is, as mentioned before, the envelope delay of the whole line, in seconds.

In computing the above it is only the components of the echoes which are in phase (or π radians out of phase) with the main transmission which affect the results. If the echo components at right angles to the main transmission are considered, they will give phase shifts in the resultant signal wave. Further, an echo component whose ratio to the main transmission is x will, when π radians out of phase with it, give a loss of x nepers; and when at right angles to it, a phase shift of x radians. Now the distribution of echo components in phase (or π radians out of phase) with the main transmission is substantially the same as that of components at right angles to it. Consequently the sinuosity is also numerically equal to the standard deviation of the difference in phase shifts at two frequencies separated by the given interval Δf. Therefore if the interval is called $\Delta\omega/2\pi$ and the resulting numerical value of the sinuosity is divided by $\Delta\omega$ it will give the standard deviation of the envelope delay. This is

$$\sqrt{\overline{(T - \overline{T})^2}} = \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\overline{K}}\right]^2 \frac{T}{2\sqrt{\Lambda}}. \tag{11}$$

The quantity which has been used in considering the suitability of a line from a delay standpoint for transmitting pictorial signals is its envelope delay distortion, or maximum departure in delay each way from a fixed average in the frequency band studied. If we make the usual assumption that the maximum departure ordinarily met (strictly speaking, except in about 3 cases out of 1000) is three times the standard deviation, then the delay distortion contributed by the irregularities is ± 3 times the expression given in equation (11).

Expressed in more usual units, the results given in equations (6), (8), (10), and (11) are repeated here.

$$\text{Rise in average attenuation (db)} = \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\bar{K}}\right]^2 \alpha L, \tag{6'}$$

$$\text{Standard deviation in attenuation (db)} = \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\bar{K}}\right]^2 \sqrt{4.343\alpha L}, \tag{8'}$$

$$\text{Sinuosity (db per kilocycle)} = 0.0256 \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\bar{K}}\right]^2 \frac{\pi\tau\sqrt{L}}{\sqrt{\alpha}}, \tag{10'}$$

$$\text{Delay distortion (microseconds)} = \pm\, 4.42 \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\bar{K}}\right]^2 \frac{\tau\,\sqrt{L}}{\sqrt{\alpha}}, \tag{11'}$$

where L = length of the line in miles,
α = attenuation of the line in db per mile,
τ = envelope delay of the line in microseconds per mile.

In order to convey a notion as to possible orders of magnitude of these effects of irregularities, and how they vary with changes in the parameters, a few calculations have been tabulated below for some hypothetical lines.

$\frac{\sqrt{\overline{\Delta K_r^2}}}{\bar{K}}$	Circuit Length, Miles	Attenuation, db per Mile	Rise in Average Loss, db	Standard Deviation in Loss, db	Sinuosity, db for Interval of 1 Kc.	Delay Distortion, Micro-Seconds
1 per cent	100	5	0.05	0.005	0.2×10^{-3}	±0.01
		10	0.10	0.007	0.15 "	±0.01
	1000	5	0.5	0.015	0.7 "	±0.04
		10	1.0	0.02	0.5 "	±0.03
2 per cent	100	5	0.2	0.02	0.9 "	±0.05
		10	0.4	0.03	0.6 "	±0.03
	1000	5	2.0	0.06	3. "	±0.15
		10	4.0	0.08	2. "	±0.1

Note: τ = 6 micro-seconds per mile.

Appendix I

Impedance

In Fig. 1 the circuit is divided into n homogeneous elementary lengths. For a current of unit value traveling down the circuit at the junction of the kth and $(k + 1)$th elementary lengths, the reflected wave is

$$h_k - h_{k+1}, \tag{1}$$

where h_k denotes the reflection coefficient (assumed to be a real number) between the impedance of the kth elementary length and the average impedance.

However, if the current starts with unit value at the sending end, then the wave has to be multiplied by the factor $e^{-kP/2}$ in reaching the point of reflection, where P is the propagation constant per two elementary lengths. In returning to the sending end the reflected wave is again multiplied by a like amount so that its value on arrival there becomes

$$(h_k - h_{k+1})e^{-kP}. \tag{2}$$

The totality of echoes returning to the sending end is

$$E_b = -h_1 + \sum_{k=1}^{n} (h_k - h_{k+1})e^{-kP} = \sum_{k=1}^{n} h_k(e^{-kP} - e^{-kP+P}). \tag{3}$$

Let

$$e^{-P} = e^{-\epsilon+i\phi} = Be^{i\phi}. \tag{4}$$

When n is large, it is permissible to use the assumption that k has ∞ for its upper limit in the above summation. The real part of E_b is accordingly

$$E_{br} = \sum_{k=1}^{\infty} h_k[B^k \cos k\phi - B^{k-1} \cos (k-1)\phi]. \tag{5}$$

By the same method as described for the more complicated case in Equation 15, Appendix II:

$$\overline{E_{br}^2} = \overline{h^2} \sum_{k=1}^{\infty} [B^{2k} \cos^2 k\phi - 2B^{2k-1} \cos k\phi \cos \{(k-1)\phi\} + B^{2k-2} \cos^2 \{(k-1)\phi\}]. \tag{6}$$

This series may next be evaluated, giving:

$$\overline{E_{br}^2} = \frac{\overline{h^2}}{2}\left(\frac{1 - 2B\cos\phi + B^2}{1 - B^2} + \frac{1 - B^2}{1 + 2B\cos\phi + B^2}\right)\cdot \tag{7}$$

In a similar manner it follows for E_{bi}, the imaginary part of E_b, that

$$\overline{E_{bi}^2} = \frac{\overline{h^2}}{2}\left(\frac{1 - 2B\cos\phi + B^2}{1 - B^2} - \frac{1 - B^2}{1 + 2B\cos\phi + B^2}\right)\cdot \tag{8}$$

Then, replacing ϵ and neglecting higher-order terms in ϕ and ϵ, which are small, and putting $\overline{h^2} = \delta^2/4$, equations (7) and (8) become

$$\overline{E_{br}{}^2} = \overline{E_{bi}{}^2} = \frac{\phi^2\delta^2}{16\epsilon}\cdot \tag{9}$$

The echo E_b affects the measured impedance. If unit voltage is impressed in series with the line, and a network having impedance $\overline{K}$, the current flowing, not counting the echoes, is $1/2\overline{K}$. The echo current is then $(E_b/1)(1/2\overline{K})$, and the total current

$$\frac{1 + E_b}{2\overline{K}}\cdot \tag{10}$$

The measured impedance is

$$\frac{2\overline{K}}{1 + E_b} \tag{11}$$

and the part due to the line is

$$K_L = \frac{2\overline{K}}{1 + E_b} - \overline{K} = \overline{K}(1 - 2E_b) \text{ approximately,} \tag{12}$$

$$K_L - \overline{K} = -\ 2E_b\overline{K}, \tag{13}$$

$$(K_{Lr} - \overline{K}_r) = -\ 2E_{br}\overline{K}, \tag{14}$$

$$(K_{Li} - \overline{K}_i) = -\ 2E_{bi}\overline{K}. \tag{15}$$

For $\overline{K}$, the real part only is to be used as it is assumed that the imaginary part is negligible in comparison with it. Where departures from $\overline{K}$ are considered, however, this imaginary part may not be negligible in comparison with the departures.

$$\overline{\Delta K_r{}^2} = 4\overline{E_{br}{}^2}(\overline{K})^2 = \frac{\phi^2\delta^2\overline{K}^2}{4\epsilon}, \tag{16}$$

$$\overline{\Delta K_i{}^2} = 4\overline{E_{bi}{}^2}(\overline{K})^2 = \frac{\phi^2\delta^2\overline{K}^2}{4\epsilon}; \tag{17}$$

$$\therefore \frac{\sqrt{\overline{\Delta K_r{}^2}}}{\overline{K}} = \frac{\sqrt{\overline{\Delta K_i{}^2}}}{\overline{K}} = \frac{|\phi|\delta}{2\sqrt{\epsilon}}\cdot \tag{18}$$

Appendix II

Attenuation

The following is a derivation of the standard deviation of the real part of the echo currents (which are received in phase with the direct transmission) over a circuit such as has been assumed in Appendix I. Accordingly, the reflected wave at the junction of the kth and $(k+1)$th homogeneous elementary lengths, for a current of unit value traveling down the circuit at this point, is:

$$h_k - h_{k+1}. \tag{1}$$

This wave returns toward the sending end and in turn suffers partial reflections. Consider this secondary reflection at the point between the jth and $(j+1)$th lengths where $j \leq k$. The wave arriving at the point in question is

$$(h_k - h_{k+1})e^{-P(k-j)/2}. \tag{2}$$

The fraction of this wave which is reflected back again is

$$-(h_j - h_{j+1}), \tag{3}$$

so that the wave which starts back from this point in the same direction as the original wave is:

$$-(h_j - h_{j+1})(h_k - h_{k+1})e^{-P(k-j)/2}. \tag{4}$$

In traveling to the junction of the kth and $(k+1)$th lengths it is again multiplied by $e^{-P(k-j)/2}$ so that the echo which is joined to the unit wave is therefore given by

$$-(h_j - h_{j+1})(h_k - h_{k+1})e^{-P(k-j)}. \tag{5}$$

If $m = k - j$, this echo is

$$-(h_j - h_{j+1})(h_{j+m} - h_{j+m+1})e^{-mP} \qquad \text{when} \qquad m > 0. \tag{6}$$

The sum of all the echoes for a given value of $m > 0$ is:

$$-e^{-mP}\sum_{j=0}^{n-m}(h_j - h_{j+1})(h_{j+m} - h_{j+m+1}) = -e^{-mP}H_m. \tag{7}$$

When $m = 0$, a slightly different treatment is necessary. Let the circuit be represented as in Fig. 1.

A unit current traveling down the circuit will suffer a reflection loss at each junction so that the current passing through the junction is $(1 - \eta_j)$ times the current entering. The ratio of the current received to the current that would be obtained without reflection loss is

$$\frac{I}{I_0} = (1 - \eta_0)(1 - \eta_1)(1 - \eta_2)(1 - \eta_3) \cdots (1 - \eta_n), \tag{8}$$

where the double reflected echoes of the previous type ($m > 0$) are omitted. The echo which is joined to the unit wave when $m = 0$ is

$$\frac{\Delta I}{I_0} = \frac{I - I_0}{I_0}. \tag{9}$$

$$\text{Log}_e \frac{I}{I_0} = \text{Log}_e \frac{I_0 + \Delta I}{I_0} = \frac{\Delta I}{I_0}, \qquad \text{when } \Delta I \text{ is small.} \tag{10}$$

Since

$$\frac{\Delta I}{I_0} = \text{Log}_e \prod_{j=0}^{n} (1 - \eta_j) = \sum_{j=0}^{n} \text{Log}_e (1 - \eta_j) \tag{11}$$

and

$$\text{Log}_e (1 - \eta) = -\eta - \eta^2/2 - \eta^3/3 \cdots, \tag{12}$$

therefore the echo is given as follows in nepers:

$$\begin{aligned} -\sum_{j=0}^{n} (\eta_j + \eta_j^2/2 + \cdots) \\ = -[-h_1 + h_1 - h_2 + h_2 - h_3 + h_3 \cdots - h_n + h_n] \\ - \frac{1}{2} \sum_{j=0}^{n} (h_j - h_{j+1})^2. \end{aligned} \tag{13}$$

The first term is zero. The sum of all the echoes is

$$\begin{aligned} -\left\{ \frac{1}{2} \sum_{j=0}^{n} (h_j - h_{j+1})^2 \right\} - \sum_{m=1}^{n} e^{-mP} H_m \\ = -\left\{ \frac{1}{2} \sum_{j=0}^{n} (h_j - h_{j+1})^2 \right\} - \sum_{m=1}^{n} H_m B^m e^{im\phi}. \end{aligned} \tag{14}$$

The in-phase component of these echoes is

$$E_{cr} = -\left\{ \frac{1}{2} \sum_{j=0}^{n} (h_j - h_{j+1})^2 \right\} - \sum_{m=1}^{n} H_m B^m \cos m\phi, \tag{15}$$

assuming h's may be taken as real and as having a symmetrical distribution curve about zero, the square of whose standard deviation may be denoted by $\overline{h^2}$.

We will consider the distribution curve of H_m, which also is real. The average value of a function $H(h)$ in a given distribution is equal to the integral of the product of the function by the frequency of occurrence for each value of it, divided by the integrated frequency of occurrence alone. The frequency of occurrence of individual values of the function is the same as that of the corresponding values of its argument, and hence can be written as $F(h)dh$ where $F(h)$ is the distribution function of the variable h. The average value of H_m is therefore

$$\begin{aligned}\overline{H}_m &= \int\int \cdots \int H_m F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n \\ &= \int\int \cdots \int \sum_{j=0}^{n-m} (h_j - h_{j+1})(h_{j+m} - h_{j+m+1}) \\ &\qquad \times F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n, \qquad (16)\end{aligned}$$

where F_k is the distribution curve of h_k, and

$$\int_{-\infty}^{\infty} F_k dh_k = 1, \qquad (17)$$

$$\int_{-\infty}^{\infty} h_k F_k dh_k = 0. \qquad (18)$$

Assuming the h's all have equal distribution curves:

$$\int_{-\infty}^{\infty} h_k^2 F_k dh_k = \overline{h^2}, \qquad (19)$$

except that since $h_0 = 0$ and $h_{n+1} = 0$, then

$$\int_{-\infty}^{\infty} h_0^2 F_0 dh_0 = 0, \qquad (20)$$

and

$$\int_{-\infty}^{\infty} h^2{}_{n+1} F_{n+1} dh_{n+1} = 0. \qquad (21)$$

Likewise

$$\int_{-\infty}^{\infty} h_k^4 F_k dh_k = \overline{h^4}, \qquad (22)$$

except that

$$\int_{-\infty}^{\infty} h_0^4 F_0 dh_0 = 0, \qquad (23)$$

$$\int_{-\infty}^{\infty} h^4{}_{n+1} dF_{n+1} dh_{n+1} = 0. \qquad (24)$$

Considering the four products $h_j h_{j+m}$, $h_j h_{j+m+1}$, $h_{j+1} h_{j+m}$ and $h_{j+1} h_{j+m+1}$, it will be seen that they all integrate to zero by virtue of symmetry unless $m = 1$ or $m = 0$. We have

$$\overline{H}_0 = \int\int \cdots \int \sum_{j=0}^{n} (h_j^2 - 2h_jh_{j+1} + h^2{}_{j+1}) \times F_1F_2 \cdots F_n dh_1dh_2 \cdots dh_n = 2n\overline{h^2}, \quad (25)$$

$$\overline{H}_1 = \int\int \cdots \int \sum_{j=0}^{n-1} (h_j - h_{j+1})(h_{j+1} - h_{j+2}) \times F_1F_2 \cdots F_n dh_1dh_2 \cdots dh_n$$
$$= \int\int \cdots \int \sum_{j=0}^{n-1} (- h^2{}_{j+1})F_1F_2 \cdots F_n dh_1dh_2 \cdots dh_n = - n\overline{h^2}, \quad (26)$$

$$\overline{H}_m = 0 \qquad \text{if} \qquad m > 1. \quad (27)$$

The average value of E_{cr} is equal to the sum of the average values of its terms. Applying the results for $\overline{H}_0$, $\overline{H}_1$, and $\overline{H}_m$, we obtain

$$\overline{E}_{cr} = - \tfrac{1}{2}\overline{H}_0 - \overline{H}_1 B \cos\phi = - [1 - B\cos\phi]n\overline{h^2}, \quad (28)$$

$$(\overline{E}_{cr})^2 = [1 - 2B\cos\phi + B^2\cos^2\phi]n^2\overline{h^2}^2. \quad (29)$$

For the mean square of E_{cr} we have:

$$\overline{E_{cr}{}^2} = \int\int \cdots \int \left(-\frac{1}{2}H_0 - \sum_{m=1}^{n} H_m B^m \cos m\phi \right)^2 \times F_1F_2 \cdots F_n dh_1dh_2 \cdots dh_n$$
$$= \int\int \cdots \int \frac{1}{4} \sum_{p=0}^{n} \sum_{q=0}^{n} (h_p - h_{p+1})^2(h_q - h_{q+1})^2 \times F_1F_2 \cdots F_n dh_1dh_2 \cdots dh_n$$
$$+ \int\int \cdots \int \sum_{m=1}^{n} B^m (\cos m\phi) \sum_{p=0}^{n} \sum_{q=0}^{n-m} (h_p - h_{p+1})^2 \times (h_q - h_{q+1})(h_{q+m} - h_{q+m+1})F_1F_2 \cdots F_n dh_1dh_2 \cdots dh_n$$
$$+ \int\int \cdots \int \sum_{r=1}^{n} \sum_{s=1}^{n} B^{r+s} (\cos r\phi)(\cos s\phi) \times \sum_{p=0}^{n-r} \sum_{q=0}^{n-s} (h_p - h_{p+1})(h_{p+r} - h_{p+r+1})(h_q - h_{q+1}) \times (h_{q+s} - h_{q+s+1})F_1F_2 \cdots F_n dh_1dh_2 \cdots dh_n. \quad (30)$$

Multiplying the factors containing the h's as indicated in (30) gives terms containing $h_ah_bh_ch_d$ where the subscripts denote some integer such as the value for p, $p + 1$, $p + r$, q, $q + 1$, $q + s$, etc. When there is equality among subscripts so that the terms become $h_u{}^2 h_v{}^2$ or $h_w{}^4$ the integration gives $(\overline{h^2})^2$ or $\overline{h^4}$, respectively. However, if such equality does not exist, or if one of the subscripts is zero or $n + 1$, the integration gives zero. By integrating term by term in the manner above indicated, adding the results, and finally thereafter putting $r = m$ and $s = m$, the following result is obtained:

$$\begin{aligned}\overline{E_{cr}{}^2} = (1 - B\cos\phi)^2 n\overline{h^4} + \Big[& n^2 - 1 - 2(n^2 + n - 2)B\cos\phi \\ & + 2(n-1)B^2\cos 2\phi + (n^2 + 4n - 6)B^2\cos^2\phi \\ & + \left\{ \sum_{m=2}^{n} (6n - 6m)B^{2m}\cos^2 m\phi \right\} \\ & - 8\left\{ \sum_{m=1}^{n-1} (n - m - \tfrac{1}{2})B^{2m+1}\cos\{(m+1)\phi\}\cos m\phi \right\} \\ & + 2\left\{ \sum_{m=1}^{n-2} (n - m - 1)B^{2m+2}\cos\{(m+2)\phi\}\cos m\phi \right\} \Big] \overline{h^2}^2. \quad (31)\end{aligned}$$

If the distribution of the h's is assumed to be a normal distribution, then:

$$\overline{h^4} = 3(\overline{h^2})^2. \quad (32)$$

Making this substitution and subtracting $(\overline{E}_{cr})^2$ gives:

$$\begin{aligned}\overline{E_{cr}{}^2} - (\overline{E}_{cr})^2 = \Big[& 3n - 1 - 8(n - \tfrac{1}{2})B\cos\phi + 2(n-1)B^2\cos 2\phi \\ & + nB^2\cos^2\phi + \left\{ \sum_{m=1}^{n} (6n - 6m)B^{2m}\cos^2 m\phi \right\} \\ & - 8\left\{ \sum_{m=1}^{n-1} (n - m - \tfrac{1}{2})B^{2m+1}\left(\frac{\cos\{(2m+1)\phi\}}{2} + \frac{\cos\phi}{2} \right) \right\} \\ & + 2\left\{ \sum_{m=1}^{n-2} (n - m - 1)B^{2m+2}\left(\frac{\cos\{(2m+2)\phi\}}{2} + \frac{\cos 2\phi}{2} \right) \right\} \Big] \overline{h^2}^2. \quad (33)\end{aligned}$$

When n is large, it is permissible to use the assumption that m has ∞ for its upper limit in the above summations. It is likewise permissible to neglect terms in the result which do not contain the factor n. Accordingly,

$$\begin{aligned}\overline{E_{cr}{}^2} - (\overline{E}_{cr})^2 = \Big[& -3 + B^2\cos^2\phi + 2\frac{(1 - B\cos\phi)^2}{1 - B^2} \\ & + 4\frac{(1 + B\cos\phi)}{1 + B^2 + 2B\cos\phi} \Big] n\overline{h^2}^2. \quad (34)\end{aligned}$$

The echo current which is joined to the unit received wave affects the final resultant and therefore the effective loss of the line. From equation (28), neglecting higher-order terms, the attenuation of the whole line is increased (in nepers) by

$$\left(\epsilon + \frac{\phi^2}{2}\right)\frac{n\delta^2}{4}\,. \tag{35}$$

The standard deviation of the attenuation (Λ, in nepers), from equation (34) and neglecting higher order terms, is

$$\sqrt{\overline{(\Lambda - \bar{\Lambda})^2}} = \frac{\phi^2\delta^2\sqrt{n}}{8\sqrt{\epsilon}}\,. \tag{36}$$

Appendix III

Sinuosity

The following is a derivation of the sinuosity of the attenuation, defined as the standard deviation of the difference $\Lambda(f + \Delta f) - \Lambda(f)$. Here $\Lambda(f)$ is the loss in the circuit at the frequency, f.

For practical purposes, the difference of the expression $E_{cr} - \bar{E}_{cr}$ at two discrete frequencies is

$$\lambda = \frac{d(E_{cr} - \bar{E}_{cr})}{df}\Delta f, \tag{1}$$

whose standard deviation will be derived below. From values of E_{cr} and $\bar{E}_{cr}$ given in Appendix II we obtain

$$E_{cr} - \bar{E}_{cr} = -\frac{1}{2}\left[\sum_{j=0}^{n}(h_j - h_{j+1})^2\right] - \left[\sum_{m=1}^{n} H_m B^m \cos m\phi\right] + [1 - B\cos\phi]n\overline{h^2}, \tag{2}$$

$$\begin{aligned}\lambda &= -\left[n\overline{h^2}\frac{d(B\cos\phi)}{df} + \sum_{m=1}^{n} H_m \frac{d(B^m\cos m\phi)}{df}\right]\Delta f \\ &= -\left[n\overline{h^2}\left\{B\frac{d\cos\phi}{df} + (\cos\phi)\frac{dB}{df}\right\}\right. \\ &\qquad \left. + \sum_{m=1}^{n} H_m\left\{B^m\frac{d\cos m\phi}{df} + (\cos m\phi)\frac{dB^m}{df}\right\}\right]\Delta f \\ &= \left[n\overline{h^2}(BQ\sin\phi - D\cos\phi) + \sum_{m=1}^{n} mH_m \right. \\ &\qquad \left. \times (B^m Q\sin m\phi - B^{m-1}D\cos m\phi)\right]\Delta f, \end{aligned} \tag{3}$$

where $Q = d\phi/df$ and $D = dB/df$.

$$
\begin{aligned}
\overline{\lambda^2} &= \int\int \cdots \int \lambda^2 F_1 F_2 F_3 \cdots F_n dh_1 dh_2 dh_3 \cdots dh_n \\
&= \int\int \cdots \int n^2 \overline{h^2}^2 (BQ \sin \phi - D \cos \phi)^2 (\Delta f)^2 \\
&\qquad\qquad \times F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n \\
&\quad + \int\int \cdots \int 2n\overline{h^2}(BQ \sin \phi - D \cos \phi) \\
&\qquad\qquad \times \left[\sum_{m=1}^{n} mH_m (B^m Q \sin m\phi - B^{m-1} D \cos m\phi) \right] \\
&\qquad\qquad \times (\Delta f)^2 F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n \\
&\quad + \int\int \cdots \int \sum_{r=1}^{n} \sum_{s=1}^{n} rs (B^r Q \sin r\phi - B^{r-1} D \cos r\phi) \\
&\quad \times (B^s Q \sin s\phi - B^{s-1} D \cos s\phi) \left[\sum_{p=0}^{n-r} \sum_{q=0}^{n-s} (h_p - h_{p+1}) \right. \\
&\quad \left. \times (h_{p+r} - h_{p+r+1})(h_q - h_{q+1})(h_{q\ +s} - h_{q+s+1}) \right] (\Delta f)^2 \\
&\qquad\qquad \times F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n. \qquad (4)
\end{aligned}
$$

By methods similar to those employed in Appendix II it follows that

$$
\begin{aligned}
\overline{\lambda^2} = \Bigg[& (BQ \sin \phi - D \cos \phi)^2 \overline{h^4} + \Bigg[- 2(BQ \sin \phi - D \cos \phi)^2 \\
& + \frac{(B^2Q^2 + D^2)(3 - 8B \cos \phi + \{6B^2 - 2B^4\} \cos^2 \phi + B^4)}{(1 - B^2)^3} \\
& - \left(Q^2 - \frac{D^2}{B^2} \right) \left(1 - \frac{1 + 6B^2 - 3B^4}{(1 + 2B \cos \phi + B^2)^3} \right. \\
& \left. - \frac{6B(1 + B^2) \cos \phi + 6B^2(1 + B^2) \cos^2 \phi + 4B^3 \cos^3 \phi}{(1 + 2B \cos \phi + B^2)^3} \right) \\
& - 2BQD \frac{\{6(1 + B^2) \cos \phi + 4B \cos^2 \phi + 8B\} \sin \phi}{(1 + 2B \cos \phi + B^2)^3} \Bigg] \\
& \qquad\qquad \times (\overline{h^2})^2 \Bigg] n(\Delta f)^2. \qquad (5)
\end{aligned}
$$

When the distribution of the h's is normal, this expression can be simplified by noting that

$$\overline{h^4} = 3\overline{h^2}^2. \tag{6}$$

The sinuosity may be obtained from $\overline{\lambda^2}$ as follows:

$$\Delta\Lambda - \overline{\Delta\Lambda} = \Lambda(f + \Delta f) - \Lambda(f) - \{\overline{\Lambda(f + \Delta f) - \Lambda(f)}\} \tag{7}$$

$$= \Lambda(f + \Delta f) - \overline{\Lambda(f + \Delta f)} - \Lambda(f) + \overline{\Lambda(f)} \tag{8}$$

$$= E_{cr}(f + \Delta f) - \overline{E_{cr}(f + \Delta f)} - E_{cr}(f) + \overline{E_{cr}(f)}. \tag{9}$$

Consequently,

$$\sqrt{\overline{(\Delta\Lambda - \overline{\Delta\Lambda})^2}} = \sqrt{\overline{\lambda^2}}. \tag{10}$$

Therefore the sinuosity, expressed in nepers, is

$$\sqrt{\overline{(\Delta\Lambda - \overline{\Delta\Lambda})^2}} = S\delta^2\sqrt{n}, \tag{11}$$

where, in accordance with equations (5) and (6):

$$\begin{aligned} S = \frac{1}{4}\Bigg[& (BQ \sin\phi - D\cos\phi)^2 \\ & + \frac{(B^2Q^2 + D^2)(3 - 8B\cos\phi + \{6B^2 - 2B^4\}\cos^2\phi + B^4)}{(1 - B^2)^3} \\ & - \left(Q^2 - \frac{D^2}{B^2}\right)\left(1 - \frac{1 + 6B^2 - 3B^4}{(1 + 2B\cos\phi + B^2)^3}\right. \\ & \left. - \frac{6B(1 + B^2)\cos\phi + 6B^2(1 + B^2)\cos^2\phi + 4B^3\cos^3\phi}{(1 + 2B\cos\phi + B^2)^3}\right) \\ & - 2BQD\,\frac{\{6(1 + B^2)\cos\phi + 4B\cos^2\phi + 8B\}\sin\phi}{(1 + 2B\cos\phi + B^2)^3}\Bigg]^{\frac{1}{2}}(\Delta f) \end{aligned} \tag{12}$$

and

$$\delta^2 = 4\overline{h^2}. \tag{13}$$

By expanding S in powers of ϵ and ϕ, and neglecting those higher than needed to give a finite result, it is found that

$$S = \frac{\phi^2\sqrt{Q^2 + D^2}}{8\epsilon\sqrt{2\epsilon}}(\Delta f). \tag{14}$$

In general, D is negligible compared to Q and the sinuosity is

$$\sqrt{\overline{(\Delta\Lambda - \overline{\Delta\Lambda})^2}} = \frac{Q\phi^2\delta^2\sqrt{n}}{8\epsilon\sqrt{2\epsilon}}(\Delta f). \tag{15}$$

Transoceanic Radio Telephone Development *

By RALPH BOWN

TEN years have elapsed since the opening to public use on January 7, 1927, of the first long distance radio telephone circuit. This form of intercontinental communication has now come into practical business and social use. A network of radio circuits interconnects nearly all the land wire telephone systems of the world. The art has passed through the pioneering stage and is well into a period of growth.

The technical side of this development, which the present paper reviews, divides naturally into four categories. The first covers those factors which made possible the beginning of commercial radio telephony.[1] In the second are the things without which its rapid growth and wide expansion could not have occurred. In the third, are a few incidental but interesting or valuable technical features. The fourth considers future improvements now in view.

Essential Initial Developments

Radio telephony presents difficulties in addition to those existing in radio telegraphy because: (1) The communication is two-way, and the radio system must be linked in with the wire telephone systems and available to any telephone instrument; (2) The subscriber cannot deliver himself of his message until the connection is actually established, and on this account delay due to unfavorable transmission conditions is less tolerable; (3) The grade of transmission required to satisfy the average telephone user is higher than that tolerable in aural tone telegraph reception by an experienced operator.

These requirements emphasized the need for accurate and quantitative knowledge of radio transmission performance as a basis for engineering radio telephone systems. There was at the same time a similar need for transmission data in the engineering of early radio broadcast installations. The effort brought to bear on these twin problems resulted in the development of practical field methods of measuring

* Digest of a paper presented at the Spring Convention of the Institute of Radio Engineers, New York, May 10, 1937, and published in full in *Proc. I. R. E.*, September, 1937.

[1] A description of the early years of radio telephone development preceding extensive commercial application, together with a discussion of the origins of the whole art, will be found in companion paper "The Origin and Development of Radio Telephony," by Lloyd Espenschied, published in *Proc. I. R. E.*, September, 1937.

radio signal strength and radio noise. The employment of long distance radio telephony in commercial use was preceded by experimental operation and tests which gave a considerable fund of statistical information covering the cyclical changes characteristic of overseas radio transmission.

The realization that a relatively high degree of reliability was essential to success discouraged any attempt at commercial service until high-power transmission on a practical basis was assured by the invention of a method of making water-cooled tubes.

In searching for the most efficient way of applying the power made available by water-cooled tubes telephone engineers were led to the employment of a method which had already been successfully used in high-frequency wire telephony. This method, now well known to radio engineers, is called single-sideband suppressed-carrier transmission. As compared with the ordinary modulated carrier transmission, it increases the effectiveness of a radio telephone system by about 10 to 1 in power. This accrues partly because none of the power capacity of the transmitter is used up in sending the non-communication bearing carrier frequency and partly because the narrower band width permits greater selectivity and noise exclusion at the receiver.

A very important final element was also necessary to prevent voice-frequency singing through residual unbalances and around the entire radio link when wire circuits and radio channels are connected together.

Recourse was again had to a device newly worked out for wire telephone transmission. By associating together and electrically interlocking several of the voice current operated switching devices which had been developed for suppressing echoes on long wire lines, an arrangement now commonly known as a "vodas"[2] was developed. When the subscriber talks, his own speech currents, acting on the vodas, cause it to connect the radio transmitter to the wire line and at the same time to disconnect the radio receiver. When the same subscriber listens the connection automatically switches back to the receiver. No singing path ever exists. The amplification in the two oppositely directed paths can be adjusted substantially independently of each other, and constant full load output from the radio transmitters is secured. With this device it became possible to connect almost any telephone line to a radio system and to adjust amplification so that a weak talker over a long wire line could operate the radio transmitter as effectively as a strong local talker.

[2] This word, "vodas," is synthesized from the initial letters of the words "voice-operated device, anti-singing."

Developments Essential to Growth

The first long distance radio telephone circuit operated (and it still operates) between the United States and England with long-wave transmission at about 5000 meters. We did not then, and we do not today, know how any considerable amount of intercontinental radio telephony could have been accomplished with circuits of this kind. The frequency space available in the long-wave range would accommodate comparatively few channels. The high attenuation to overland transmission and the high noise level at these wave-lengths preclude their satisfactory use for very great distances or in or through tropical regions. The discovery that short waves could be transmitted to the greatest terrestrial distances and could be satisfactorily received in the tropics came at a most opportune time.

Short-wave transmission not only released the limitations on distance and location inherent to long waves but also opened up such a wide range of frequency space as to give opportunity for an extensive growth in numbers of both radio telegraph and radio telephone circuits. Short waves further encouraged the growth of radio telephony by making it cheaper. Thus, it became possible to make directive antenna structures of moderate size which increased the effectiveness of transmission many times, thereby reducing the transmitter power required for a given reliability of communication. Short waves were the indispensable element without which material growth could not have occurred, but there were other significant things.

An important desideratum in telephony is privacy. Commercial radio telephony would have been severely hampered if privacy systems had not been developed to convert speech into apparently meaningless sounds during its radio transit.

Another item of great aid in promoting growth was the development of methods of accurate stabilization of transmitted frequencies. The first effect of this was to eliminate the extreme distortion which characterized early short-wave telephone transmission and which was found to be due to parasitic phase or frequency modulation effects in the transmitters. As the number of radio communication facilities, both telegraph and telephone, grew, accurate stabilization of frequency became a necessity in order to permit effective utilization of the available frequency space without mutual interference between stations.

Later Technical Advances

The "rhombic" antenna is mechanically simple and electrically nearly aperiodic, covering a wide wave-length range efficiently. It has radically changed the character of the physical plant and investment necessary to the employment of directivity in short-wave transmitting and receiving.

In Hawaii and the Philippines on circuits to the United States the "diversity" method of reception is used wherein three individual separated antennas and receivers with interlocked automatic gain controls are combined to produce a common output having less distortion and noise than a single receiver.

The effects of distortion in short-wave circuits are avoided to some extent by an arrangement called a "spread sideband system," which has been used on circuits between Europe and South America. By raising the speech in frequency before modulation the speech sidebands are displaced two or three kilocycles from the carrier and many of the product frequencies resulting from intermodulation fall into the gap rather than into the sidebands.

On the Holland-Java route a system is being used whereby more than one sideband is associated with a single carrier or pilot frequency, each such sideband representing a different communication.

An improved signal-to-noise ratio is given by a device called a "compandor" [3] employed on the New York-London long wave circuit. It raises the amplitude of the weaker parts of the speech previous to transmission. In depressing these raised parts to their proper relative amplitude, after reception, the compandor also depresses the accumulated radio noise.

Present Outlook

The foregoing makes it evident that many fundamental engineering problems have been solved and that the pioneering stage of the service, when its possibility of continued existence might reasonably have been in doubt, has definitely been passed. In looking toward the future we find that the greatest needs are for improvement in reliability and in grade of service, accompanied by reduced costs.

Improving the reliability struggles against the fact that short wave transmission varies through such a wide range of effectiveness, and seems to be so much influenced by the sun. We have not only a daily cycle in the transmission of a given frequency but also an annual cycle and beyond this an eleven-year cycle associated with the change in sunspot activity. Superimposed upon these are erratic and occasionally large variations associated with magnetic storms.

[3] The synthetic word "compandor" is a contraction of the compound word "compressor-expander," which describes the effects the device has on the volume range of speech.

A statistical study of the data secured from operation of transoceanic radio telephone circuits over the past several years has given valuable help in engineering circuits to meet a given standard of reliability. This study has shown that the percentage of lost time suffered on a circuit appears to follow a probability law and that its relation to the transmission effectiveness of the circuit in decibels is given by a straight line when plotted to an arithmetic probability scale. Such a relation tells us, for example, that if a circuit as it stands suffers 15 per cent lost time, the lost time can be reduced to a selected lower value, say 5 per cent, by improving the circuit a definite amount, in the assumed case 10 decibels. It then becomes possible, by making engineering cost studies of the various available ways of securing the necessary number of decibels improvement in performance, to choose the most economical one. This approach is being applied to study of the radio telephone circuits extending outward from the United States. Some of the technical possibilities which are being considered for improving these circuits are discussed below.

The performance of a radio telephone circuit may be changed by dynamically modifying the amplification or other characteristics of the circuit in accordance with the speech transmitted. The compandor already mentioned is an example of this kind of improvement on long waves. Further developments particularly suited to the vagaries of short-wave transmission are possible.

The operation of the vodas, or voice-operated switching device linking the wire and radio circuits, is adversely affected by noise. Methods are being investigated for using single frequencies, called "control tones," transmitted alongside the speech band and under the control of speech currents, to give more positive operation of the switching devices and reduce the adjustment required.

The transmission improvement of about 9 decibels (about 10 : 1 in power) offered by single-sideband suppressed-carrier transmission has been delayed in its application to short-wave transmission partly because of the high degree of precision in frequency control and selectivity necessary to its accomplishment. In recent years successful apparatus has been developed and proved satisfactory in trials. The introduction of single sideband into commercial usage is already in progress.

Turning now from the transmitting to the receiving end, one fundamental way to reduce noise in radio telephony is to employ sharper directivity. It has been found by observation that there is a limit to which directivity, as ordinarily practiced, can be carried to advantage. It is easy to design antennas so sharp that at times very large improvements in signal-to-noise ratio are secured. But it is found that at other times these antennas are actually poorer than are much less sharply directive systems. Such observations also indicate a wide variation in the performance of antennas as regards selective fading, and the signal distortion accompanying it.

The result of all this work has been the development of a system based on an entirely new approach to the problem of sharp directivity and of telephone receiving. This system is called a MUSA System, the word MUSA being synthesized from the initial letters of the descriptive words Multiple Unit Steerable Antenna. An outline of the principles and methods is given below.

By sending short spurts or pulses of short-wave radiation from one side of the Atlantic, and receiving on the other side, it has been observed that each spurt may be received several times in quick succession. But these echoes do not arrive like successive bullets from the same gun, all following the same path. They come slanting down to the receiver from different angles of elevation, these vertical angular directions remaining comparatively stable. While the signal received at each of the individual directions may be subject to fading, the fading is somewhat slower and is not very selective as to frequency. The signal component coming in at a low angle takes less time in its trip from the transmitter than a high angle component. Evidently the low-angle paths are shorter. All these facts fit in well, on the average, with the ideal geometrical picture of waves bouncing back and forth between the ionosphere and the ground and reaching the receiver as several distinct components which started out at different angles, have been reflected at different angles, and have suffered different numbers of bounces.

The ordinary directive antenna is blunt enough in its vertical receiving characteristic to receive all or nearly all of these signal components at once. Because of their different times of transit the various components do not mix well but clash and interfere with one another at the receiver. This shows up as the selective fading and distortion which characterize short-wave reception much of the time. The MUSA method remedies this trouble.

The MUSA provides extremely sharp directivity in the vertical plane. By its use a vertical angular component can be selected individually. It consists of a number of rhombic antennas stretched out in a line toward the transmitter and connected by individual coaxial lines to the receiving apparatus. The apparatus is adjustable so that the vertical angle of reception can be aimed or "steered" to select any desired component as a telescope is elevated to pick out a star. The antennas remain mechanically fixed. The steering is done electrically with phase shifters in the receiving set. By taking several branch circuits in parallel from the antennas to different sets of adjusting and receiving apparatus the vertical signal components may be separated from each other.

Nature breaks the wave into several components and jumbles them together. The first function of the MUSA system, as just described, is to sort the components out again. Its second function is to correct their differences so that they may be combined smoothly into a replica of the original signal. To do this the received wave components are separately detected and passed through individual delay circuits to equalize their differences in transit time. They are then combined to give a single output. As compared with a simple receiver the MUSA receiving system gives (1) improvement in signal-to-noise ratio, as a result of the sharp directive selectivity of the antenna; (2) improvement against selective fading distortion, by virtue of the equalization of the time differences between the components before they are allowed to mix; and (3) improvement against noise and distortion, because of the diversity effect of combining the several components.

Fortunately, it is found that the directive selection and the delay compensation adjustments correct for one frequency are satisfactory for a considerable band of frequencies adjacent thereto. Thus there is offered the possibility of receiving a number of grouped channels through one system and the prospect appears not only of improved transmission but also of reduced cost per channel.

The possibility of grouping channels at the transmitting station may be conceived on the basis of either "multiple" or "multiplex" transmission. In the multiple arrangement each channel has its own antenna and its individual transmitter whose frequency is closely spaced from and coordinated with the adjacent channels of the group. In "multiplex" transmission, the channels are aggregated into a group at low power and handled *en bloc* through a common high-power amplifier and radiating system. Particularly in the multiplex case, there are possibilities of important economies if the technical problems are satisfactorily solved. Passing a multiplicity of channels simultaneously through a common-power amplifier involves interchannel interference due to modulation products which is not met with when only one channel is present. Severe requirements are thereby placed on the distortion characteristics of the power amplifier.

It seems a fair conclusion that the tendency in the engineering solution of the problems of economy and growth in radio telephone development (and perhaps also radio telegraph development) will be toward channel grouping methods, especially for backbone routes between important centers where large traffic may develop. This will be a considerable departure from past practice which has resulted in the existing system of scattered frequency assignments. It is to be hoped that the obvious difficulties in rearranging frequency assignments will not prove so unyielding as to preclude putting new engineering developments into service.

A Negative-Grid Triode Oscillator and Amplifier for Ultra-High Frequencies *

By A. L. SAMUEL

THE author describes three negative-grid triodes of unusual design which operate both as oscillators and as amplifiers at ultra-high frequencies. The power output of the smallest tube as an oscillator at 1500 megacycles is 2 watts, and is still capable of an output of 1 watt at 1700 megacycles with an oscillation limit of 1870 megacycles corresponding to a wave-length of 16 centimeters. This tube also offers possibilities as an amplifier at frequencies as high as 1000 megacycles. Such capabilities of the negative-grid triode are notable since this device has appeared to lag behind the magnetron as an *oscillator* at frequencies of above roughly 500 megacycles, while the only successful power *amplifiers* which have been described for frequencies of the order of 300 megacycles are multi-element tubes.

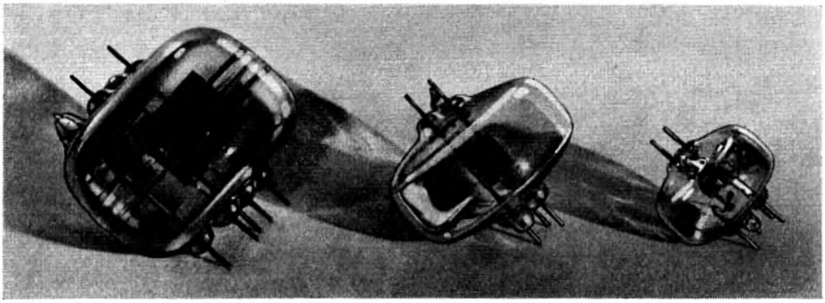


Fig. 1—Experimental double-lead tubes.

The triode as used at radio frequencies differs from the multi-element tube chiefly in the manner in which interaction is prevented between the input and output circuits. This is obviously a circuit limitation, as contrasted with the electron transit time limitation which has received so much attention. The greatest opportunity for improvement seems to be in the direction of improved circuit design. The tubes described in this paper were developed from this point of view.

Sample tubes are shown in Fig. 1. They differ from previous designs

* Digest of a paper presented before International Scientific Radio Union April 30, 1937 at Washington, D. C. Published in *Proc. I. R. E.*, October, 1937.

primarily in the lead arrangement. From the section view of one of these tubes, shown in Fig. 2, it will be observed that the grid and plate elements are supported by wires which in effect go straight through the tube envelope providing two independent leads to each of these elements. The filament leads are at one end only and one of these leads is extremely short. This unusual lead arrangement possesses a number of unique advantages.

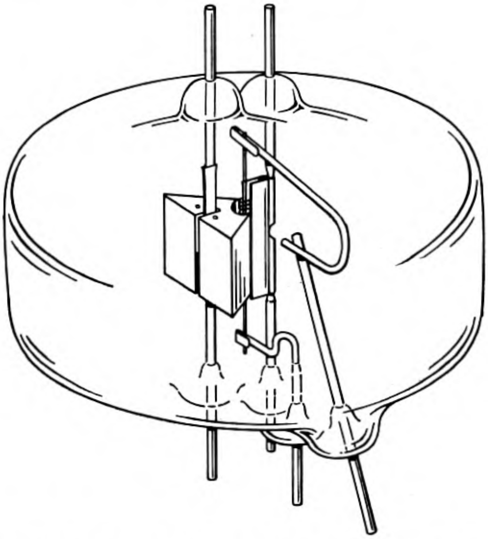


Fig. 2—Section view of one of the double-lead tubes.

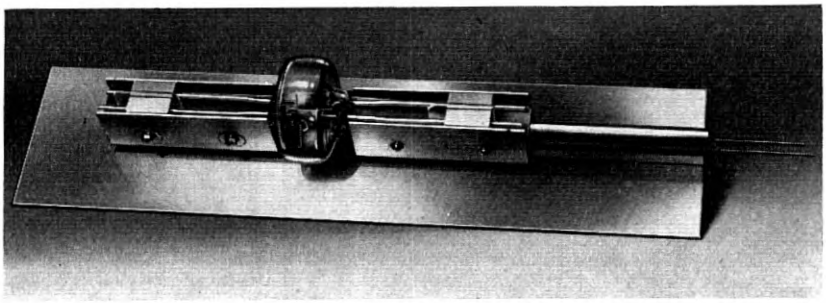


Fig. 3—Typical oscillator circuit.

A typical oscillator circuit is shown in Fig. 3. Here the tube is mounted at the center of a half-wave Lecher system. This arrangement provides a higher natural frequency circuit than that of the quarter-wave Lecher system formed by removing one set of leads. Since only half of the total charging current to the inter-electrode capacitances flows through each set of leads, the losses due to the lead resistances are also reduced. In the tubes under discussion the electron transit time limitation has been met by the use of extremely small inter-electrode spacings so that full advantage may be taken of the increased frequency range.

For the purpose of confirming the above conclusion, efficiency curves have been obtained on the large size tube, as shown in Fig. 1, when operated both single- and double-ended. The results are shown in Fig. 4. It will be observed that the efficiencies for double-ended operation are always higher than for the single-ended case over the range covered by the experimental data. In fact, usable outputs are obtained at frequencies well beyond the point where the single-ended tube fails to operate. The ratio of the cut-off frequencies for the two tubes happens to be 1.23 for the particular conditions under which these data were obtained.

Output and efficiency curves for the large size tube are shown in Fig. 5. The values of 60 watts at 300 megacycles and 40 watts at 400 megacycles compare quite favorably with outputs reported from radiation-cooled magnetrons. When the problems of modulation and the complications of the magnetron's magnetic field are considered, the advantages of the negative-grid triode become more apparent. The improvement in power output made possible by this departure in design is illustrated by the comparison plot shown in Fig. 6.

The double-lead arrangement is also responsible for an increase in the upper frequency limit at which stable operation as an *amplifier* may be secured.

The primary cause for instability of the triode amplifier is the interaction between the input and output circuits which results from the admittance coupling between these circuits provided by the grid-plate capacitance. A second source of coupling is that caused by common impedances in the two circuits in the nature of the self and mutual inductance of the tube leads. At moderately high frequencies this latter coupling is usually of negligible importance. Stable operation is thus possible when suitable means are provided to compensate or "neutralize" the admittance coupling. At ultra-high frequencies lead-impedance coupling can no longer be neglected. It may, of course, be minimized by the use of short leads. The ultimate solution is to provide independent leads for the input, output and admittance neutralizing circuits. The double-lead tube is an attempt to fulfill these conditions. It will be observed that the only common impedance remaining is that caused by one filament lead and that this lead is extremely short.

In the present investigation the method of neutralizing admittance coupling has been that disclosed by H. W. Nichols in U. S. Patent 1,325,879 and involves the resonating of the offending admittances at the desired operating frequency so that the resulting parallel admittance is reduced to a very low value. This takes the form of an inductance connected between the grid and plate of the tube and adjusted

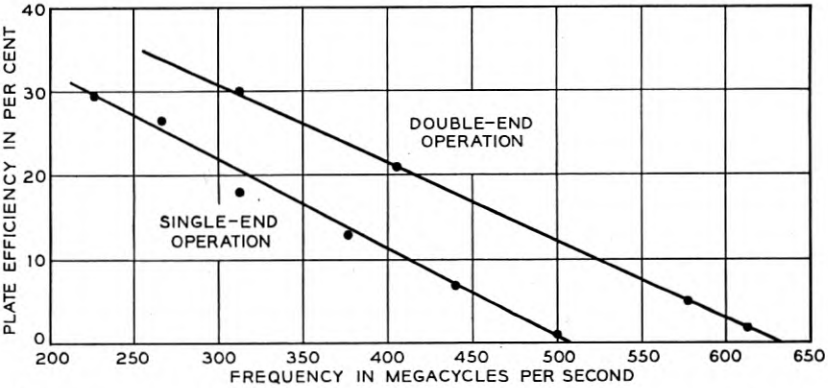


Fig. 4—Comparison plot of output efficiency for the large tube when operated single-ended and double-ended.

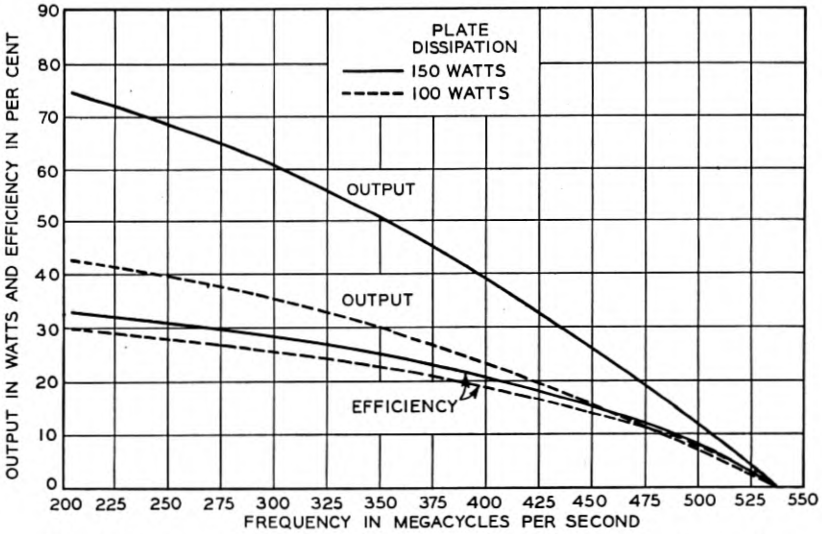


Fig. 5—Output and efficiency as a function of frequency for the large tube.

to resonate with the grid-plate capacitance. For ease of adjustment a somewhat lower fixed inductance may be used and tuned by the adjustment of a small variable condenser in parallel. This form of neutralization is commonly referred to as "coil" neutralization. At ultra-high frequencies where unavoidable inductances are already present in the form of lead inductances, this "coil" scheme possesses outstanding advantages over the more usual "capacitance" schemes. These advantages become even more pronounced with the availability of the double-lead tube.

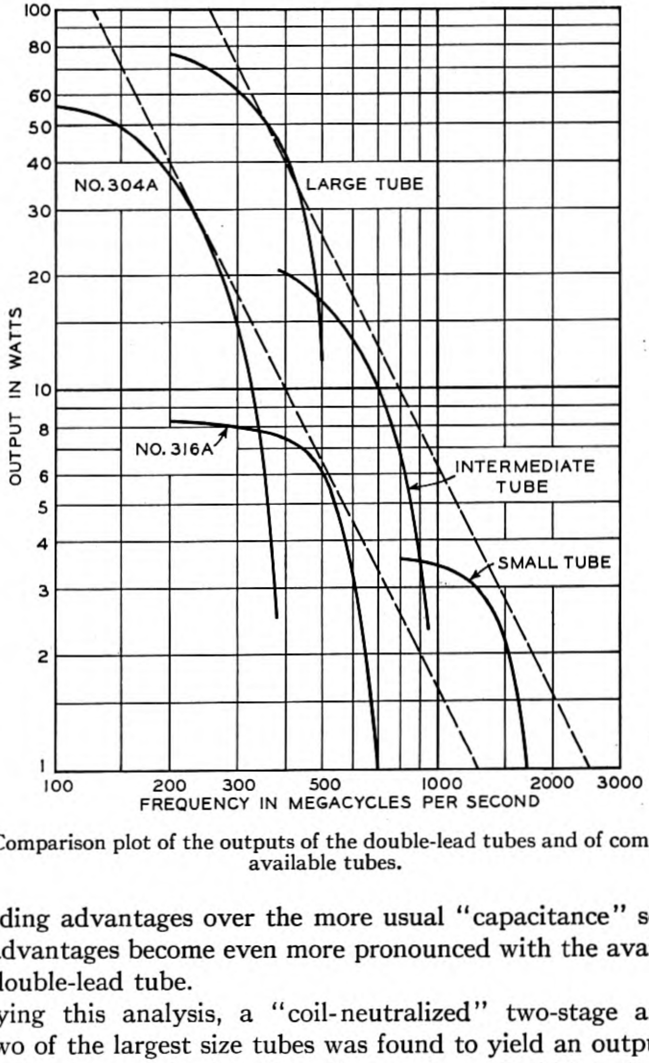


Fig. 6—Comparison plot of the outputs of the double-lead tubes and of commercially available tubes.

Verifying this analysis, a "coil-neutralized" two-stage amplifier using two of the largest size tubes was found to yield an output of 60 watts at 144 megacycles for Class B operation. Stability, distortion, and band width were quite comparable to the results obtained on a pentode of similar rating. A four-stage amplifier employing the intermediate tube gave comparable results and although experimental data are not yet available, it seems reasonable to assume that the small size tube will permit of stable operation as an amplifier at frequencies as high as 1000 megacycles.

The double-lead tube is therefore seen to possess a number of distinct advantages, both as oscillator and as amplifier, in the frequency range from 100 megacycles to 1000 megacycles. While the ultimate limit to which such developments may be pushed is still a matter of conjure it seems safe to predict that the triode will be able to meet the demands of the circuit designer at least for some time to come.

Addendum to "Radio Propagation Over Plane Earth—Field Strength Curves"

By CHAS. R. BURROWS

IN the paper of the above title in the January 1937 issue of the *Bell System Technical Journal*, an approximation which was not explicitly pointed out was made in deriving equation (17). A note from Mr. K. A. Norton* of the Federal Communications Commission points out that equation (17) does not give a reasonable result when $\tau = 1$. The explanation is that two terms which are unimportant except near the transmitter when the ground is a perfect dielectric were deleted. The complete equation is

$$\frac{E}{2E} = \frac{W}{1+\tau^2} + \frac{1}{1-\tau^4}\left[\frac{1-\tau e^{(2\pi id/\lambda)(1-1/\tau)}}{2\pi id/\lambda} + \frac{1-\tau^2 e^{(2\pi id/\lambda)(1-1/\tau)}}{(2\pi id/\lambda)^2}\right]. \qquad (17)$$

When $\tau = 1$ by virtue of equation (13) W must equal 1/2 and accordingly the first term on the right of equation (17) is 1/4. The second term gives $1/4 + \dfrac{1/4}{2\pi id/\lambda}$ and the last term gives $\dfrac{1/4}{2\pi id/\lambda} + \dfrac{1/2}{(2\pi id/\lambda)^2}$. Hence when $\tau = 1$ equation (17) gives the following relation for the field strength in free space,

$$\frac{E}{2E_0} = \frac{1}{2} + \frac{1/2}{2\pi id/\lambda} + \frac{1/2}{(2\pi id/\lambda)^2},$$

as it should.

The terms added to equation (17) produce oscillations in the curves of Fig. 3 as shown on the following page. For any physical dielectric the conductivity is not zero and the oscillations disappear at the greater distances giving curves like those of the original Fig. 3.

Equation (19) should read

$$E \rightarrow \left[\frac{1}{1-\tau^4}\frac{1+\tau^2}{2\pi\tau^2 id/\lambda}\right]\left[1 - \tau^3 e^{\frac{2\pi id}{\lambda}\left(1-\frac{1}{\tau}\right)}\right] 2E_0. \qquad (19)$$

This increases the deviation of the second factor on the right from unity but if the ground is not a perfect dielectric the exponential reduces the second factor to unity at the greater distances irrespective of the value of ϵ.

* See the note at the end of " The Propagation of Radio Waves over the Surface of the Earth and in the Upper Atmosphere," *Proc. I.R.E.*, **25**, 1203–1236, September, 1937.

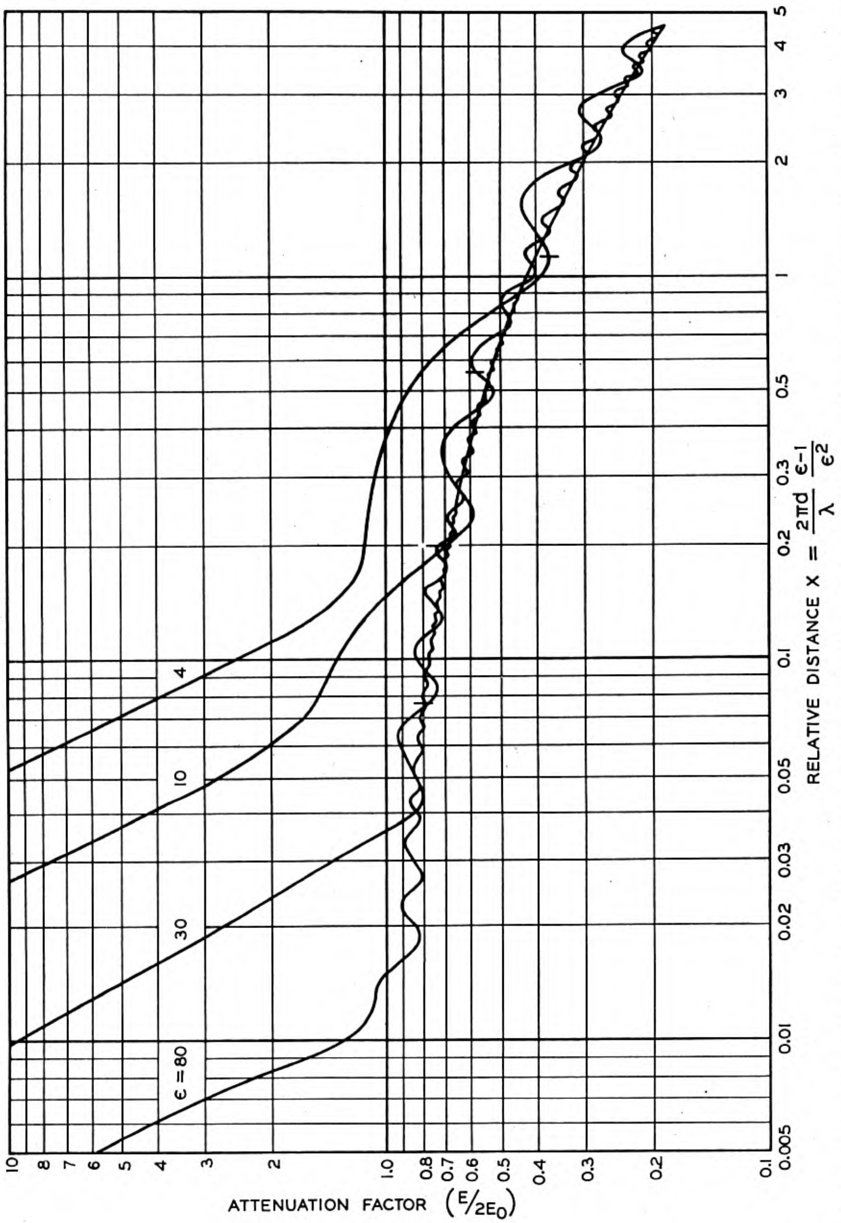


Fig. 3—Attenuation factor for radio propagation over a dielectric plane. The number on each curve gives the value of the dielectric constant to which it applies.

The situation in the immediate vicinity of the antenna is more clearly represented in Fig. 3A in which the attenuation factor is plotted against distance in wave-lengths. This allows inclusion of curves for $\epsilon = 1$ (i.e. for the earth replaced by air) and $\epsilon = \infty$ (which is equivalent to perfectly conducting earth). Comparison of these curves with the broken lines which are replotted from Fig. 2 shows that for distances greater than a wave-length the main effect of using the curves of Fig. 2 is to ignore the presence of the oscillations in the curves. For a perfect dielectric the amplitudes of these oscillations do not decrease below $\pm 1/\epsilon^{3/2}$ even at great distances as can be seen from equation (19). The presence of some conductivity causes these oscillations to be damped out. For example, a Q of 5 reduces the amplitudes of these oscillations within the first four wave-lengths to a value too small to show on the figure.

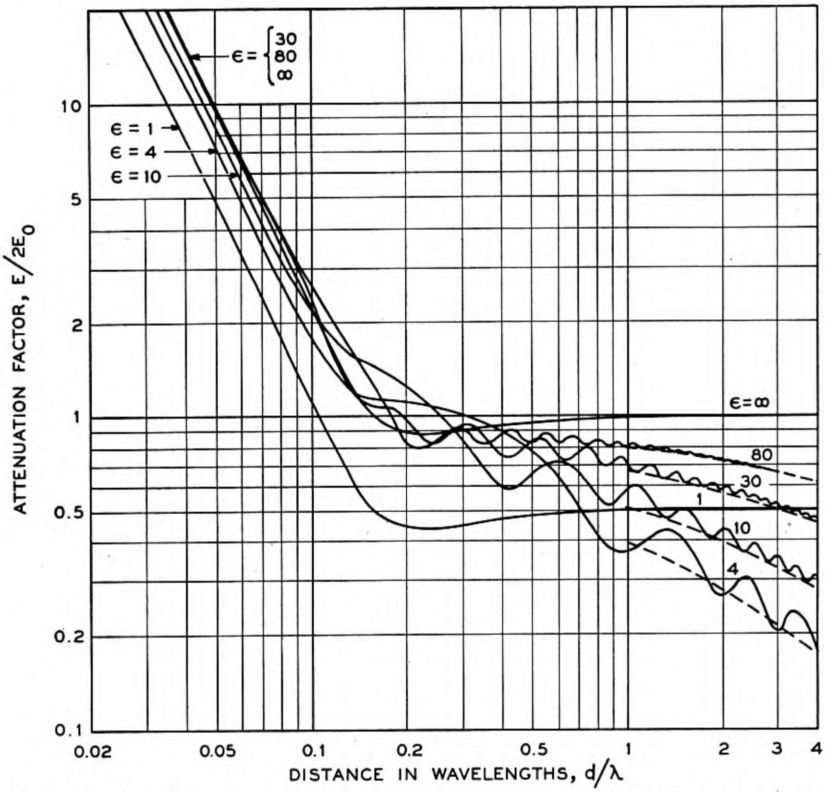


Fig. 3A—Variation of attenuation factor with distance in wave-lengths for transmission over a dielectric plane. For d/λ small,

$$E/2E_0 = \left(\frac{\epsilon}{\epsilon + 1}\right) \Big/ \left(\frac{2\pi d}{\lambda}\right)^2.$$

The broken curves are replots of the curve for $Q = \infty$ from Fig. 2.

The second paragraph of the footnote referring to equation (17) should read:

The differential equation given by Wise for $\boldsymbol{A\Pi}_0$ becomes

$$-\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial d^2}+\frac{1}{d}\frac{\partial}{\partial d}\right)\boldsymbol{A\Pi}_0 = \left(\frac{A}{1+\tau^2}+\frac{1}{1-\tau^4}\left[\frac{1}{2\pi i d/\lambda}+\frac{1}{(2\pi i d/\lambda)^2}\right]\right)\boldsymbol{\Pi}_0,$$

when the value of $y = (1+\tau^2)\boldsymbol{A\Pi}_0$ is substituted in his equation (7) and the result divided by $1+\tau^2$. The i of this paper is equal to $-i$ in Wise's paper. By interchanging k_1 and k_2 in Wise's equation (7) and proceeding along parallel lines the corresponding equation of $\boldsymbol{D\Pi}_0 = y/(1+\tau^2)$ is found to be

$$-\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial d^2}+\frac{1}{d}\frac{\partial}{\partial d}\right)\boldsymbol{D\Pi}_0 = \left(\frac{D}{1+\tau^2}-\frac{1}{1-\tau^4}\left[\frac{\tau}{2\pi i d/\lambda}+\frac{\tau^2}{(2\pi i d/\lambda)^2}\right]\right)\boldsymbol{\Pi}_0.$$

Adding these two relations gives an expression for $\left(\frac{\partial^2}{\partial d^2}+\frac{1}{d}\frac{\partial}{\partial d}\right)\boldsymbol{\Pi}$, where

$$\boldsymbol{\Pi} = 2(A+D)\boldsymbol{\Pi}_0,$$

which when substituted in the above equation for $\boldsymbol{E}$ and the result divided by $2\boldsymbol{E}_0$, where $\boldsymbol{E}_0 = -240 i\pi^2\boldsymbol{\Pi}_0/\lambda$, gives equation (17). Since $\boldsymbol{E}_0$ is the inverse distance component of the free space field, this relation for $\boldsymbol{E}_0$ follows from equation (11).

In the last line of the footnotes on page 51 read $2/(1+\tau^2)$ for $2/(1-\tau^4)$.

Abstracts of Technical Articles from Bell System Sources

What Electrons Can Tell Us about Metals.[1] C. J. Davisson. Some general statements are made about electron waves and electron diffraction, three typical investigations in which electron diffraction has been employed are described, and the technique of a new type of electron crystal analysis is discussed.

Relation between Loudness and Masking.[2] Harvey Fletcher and W. A. Munson. A functional relationship between the loudness of a sound and the degree to which it masks single frequency tones, that is, the masking audiogram of the sound, is developed. A loudness-masking function is determined experimentally. From this loudness-masking relationship the loudness of a sound can be computed by simply integrating the area under the masking audiogram plotted on a special chart. Comparisons of computed and observed loudness levels are shown for a number of sounds and serve to illustrate the precision to be expected from the method. Finally, the results of a large number of masking tests are given in the form of masking contours, which enable one to predict the masking audiogram of a sound from measurements of its intensity spectrum.

Coupling between Parallel Earth-Return Circuits under D.-C. Transient Conditions.[3] K. E. Gould. In tests conducted in connection with several d.-c. railway electrifications, the induced voltages recorded in paralleling communication circuits at times of short circuit on the railway have shown marked divergences from values computed on the basis of uniform earth resistivity and a rate of change of earth current determined from measurements in trolley and rail circuits. Due to the numerous factors which might contribute to these divergences, such as non-uniform division of transient current along the tracks and associated return conductors, the presence of shielding conductors along or near the right-of-way, etc., it was felt that a better understanding of the problem of induction under d.-c. transient conditions could be obtained by experimental studies of the transient coupling between parallel earth-return circuits, free from the effects of shielding conductors, and with concentrated, rather than distributed, grounds. The study described in this paper was undertaken for this purpose.

[1] *Jour. of Applied Physics*, June 1937.
[2] *Jour. Acous. Soc. Amer.*, July 1937.
[3] *Electrical Engineering*, September 1937.

The locations for the tests were selected to provide a reasonably large range of earth resistivity; also, at one location it was known that the earth structure departed substantially from uniformity. At each of these locations d.-c. transient coupling tests were performed in which transient currents, approximately of the form encountered during faults on d.-c. railway electrifications, were produced in an earth-return circuit, herein referred to as the primary, and measurements were made of the resultant voltages in earth-return circuits, herein called secondary circuits, parallel to and at separations from the primary circuit of from 50 or 60 to 2,000 feet. In addition to the d.-c. transient tests, measurements were made at each location of the steady state a.-c. coupling, in magnitude and phase angle, over a range of frequencies from 20 or 30 cycles to 3,200 cycles. From these a.-c. measurements the transient voltages were computed for a number of cases by evaluating the Fourier integral. The results of the a.-c. coupling tests were useful also in helping to explain, in a general way, the departures of the measured transient voltages from the voltages computed for uniform earth resistivity.

The measured transient voltages and voltages computed (1) from the a.-c. coupling measurements and (2) on the basis of a uniform earth resistivity, are shown for several representative cases in figures accompanying the paper.

The Shunt-Excited Antenna.[4] J. F. MORRISON and P. H. SMITH. The paper describes an arrangement for exciting a vertical broadcast antenna with the base grounded. Construction economy results through the elimination of the base insulator, the tower lighting chokes, and the usual lightning protective devices. The coupling apparatus at the antenna end of the transmission line is reduced to an extent which may make unnecessary a separate building for its protection. Greater freedom from interruptions resulting from static discharges is expected. The performance of the design is substantially the same as that obtained from the antennas now in general use.

The paper describes experimental work done, results obtained, and inferences to be drawn from them. A mathematical appendix is attached.

Some Fundamental Experiments with Wave Guides.[5] G. C. SOUTHWORTH. This paper describes in considerable detail the early apparatus and methods used to verify some of the fundamental properties of wave guides. Cylinders of water about ten inches in diameter and

[4] *Proc. I. R. E.*, June 1937.
[5] *Proc. I. R. E.*, July 1937.

four feet long were used as the experimental guides. At one end of these guides were launched waves having frequencies of roughly 150 megacycles. The lengths of the standing waves so produced gave the velocity of propagation. Other experiments utilizing a probe made up of short pickup wires attached to a crystal detector and meter enabled the configuration of the lines of force in the wave front to be determined. This was done for each of four types of waves. For certain types the properties had already been predicted mathematically. For others the properties were determined experimentally in advance of analysis. In both cases analysis and experiment proved to be in good agreement.

The Dependence of Hearing Impairment on Sound Intensity.[6] JOHN C. STEINBERG and MARK B. GARDNER. This paper discusses the measurement of hearing loss for levels of sound that were well above the deafened threshold and hence were audible to the deafened person. In the tests, observers having unilateral deafness, i.e., one impaired and one normal ear, balanced a tone heard with the deafened ear against the tone heard with the normal ear. For some persons, the impaired ear heard less well than the normal ear for all sound levels. For others, tones which were well above the deafened threshold were heard about equally well with either ear. In other words, such deafened ears tended to hear loud sounds with normal loudness. It was found that this type of deafness could be represented quantitatively on the assumption that it was due to nerve atrophy. Loudness judgments for a normal ear in the presence of noise were found to be similar to judgments by a nerve deafened ear. Relations, based on the loudness properties of normal ears, have been extended to represent the loudness heard by deafened ears.

[6] *Jour. Acous. Soc. Amer.*, July 1937.

Contributors to this Issue

H. A. Affel, S. B. in Electrical Engineering, Massachusetts Institute of Technology, 1914; Research Assistant in Electrical Engineering, 1914–16. American Telephone and Telegraph Company, Engineering Department and the Department of Development and Research, 1916–34; Bell Telephone Laboratories, 1934–. As Assistant Director of Transmission Development, Mr. Affel is engaged in development work connected with carrier telephone and telegraph systems.

Ralph Bown, M.E., 1913; M.M.E., 1915; Ph.D., 1917, Cornell University. Captain, Signal Corps, U. S. Army, 1917–19. American Telephone and Telegraph Company, Department of Development and Research, 1919–34; Bell Telephone Laboratories, 1934–. As Radio Research Director, Dr. Bown is concerned with radio development problems. He is a Past President of the Institute of Radio Engineers.

Charles R. Burrows, B.S. in Electrical Engineering, University of Michigan, 1924; A.M., Columbia University, 1927; E.E., University of Michigan, 1935. Research Assistant, University of Michigan, 1922–23. Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, Research Department, 1925–. Mr. Burrows has been associated continuously with radio research and is now in charge of a group investigating the propagation of ultra-short waves.

John R. Carson, B.S., Princeton, 1907; E.E., 1909; M.S., 1912. American Telephone and Telegraph Company, 1914–34; Bell Telephone Laboratories, 1934–. As Transmission Theory Engineer for the American Telephone and Telegraph Company and later for the Laboratories, Mr. Carson has made substantial contributions to electric circuit and transmission theory and has published extensively on these subjects. He is now a research mathematician.

Thornton C. Fry, A.B., Findlay College, 1912; A.M., University of Wisconsin, 1913; Ph.D., 1920; Instructor in mathematics, University of Wisconsin, 1912–16. Mathematician, Western Electric Company, 1916–24; Bell Telephone Laboratories, since 1924. Lecturer electrical engineering, M.I.T., 1927; Lecturer mathematics, Princeton, 1929–30. Dr. Fry's work in the Laboratories has been of a mathematical character.

J. M. Manley, B.S. in Electrical Engineering, University of Missouri, 1930; Bell Telephone Laboratories, 1930–. Mr. Manley has been engaged principally in theoretical studies of non-linear electrical circuits.

W. P. Mason, B.S. in Electrical Engineering, University of Kansas, 1921; M.A., Columbia University, 1924; Ph.D., 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged in investigations on carrier transmission systems and more recently in work on wave transmission networks, both electrical and mechanical.

Pierre Mertz, A.B., Cornell University, 1918; Ph.D., 1926. American Telephone and Telegraph Company, Department of Development and Research, 1919–23, 1926–34; Bell Telephone Laboratories, 1934–. Dr. Mertz has been engaged in special problems in toll transmission, chiefly in telephotography and television.

S. O. Morgan, B.S. in Chemistry, Union College, 1922; M.A., Princeton University, 1925; Ph.D., 1928. Western Electric Company, Engineering Department, 1922–24; Bell Telephone Laboratories, 1927–. Dr. Morgan's work has been on the relation between dielectric properties and chemical composition.

E. J. Murphy, B.S., University of Saskatchewan, Canada, 1918; McGill University, Montreal, 1919–20; Harvard University, 1922–23. Western Electric Company, Engineering Department, 1923–25; Bell Telephone Laboratories, 1925–. Mr. Murphy's work is largely confined to the study of the electrical properties of dielectrics.

E. Peterson, Cornell University, 1911–14; Brooklyn Polytechnic, E.E., 1917; Columbia, A.M., 1923; Ph.D., 1926; Electrical Testing Laboratories, 1915–17; Signal Corps, U. S. Army, 1917–19. Bell Telephone Laboratories, 1919–. Dr. Peterson's work has been largely in theoretical studies of carrier current apparatus.

K. W. Pfleger, A.B., Cornell University, 1921; E.E., 1923. American Telephone and Telegraph Company, Department of Development and Research, 1923–34; Bell Telephone Laboratories, 1934–. Mr. Pfleger has been engaged in transmission development work, chiefly on problems pertaining to delay equalization, delay measuring, temperature effects in loaded-cable circuits, and telegraph theory.

A. L. Samuel, A.B., College of Emporia (Kansas), 1923; S.B. and S.M. in Electrical Engineering, Massachusetts Institute of Technology, 1926. Instructor in Electrical Engineering, Massachusetts

Institute of Technology, 1926–28. Bell Telephone Laboratories, 1928–. Mr. Samuel has been engaged in research and development work on vacuum tubes.

C. C. Taylor, B.S. in Electrical Engineering, Colorado College, 1917. American Telephone and Telegraph Company, Long Lines Department, 1920–28; Department of Development and Research, 1929–34. Bell Telephone Laboratories, 1934–. Mr. Taylor's work has been concerned with radio-wire systems.

L. R. Wrathall, B.S., University of Utah, 1927; Graduate School, 1927–28. Bell Telephone Laboratories, 1929–. Mr. Wrathall has been engaged in the study of non-linear reactances.

S. B. Wright, M.E. in Electrical Engineering, Cornell University, 1919. Engineering Department and Department of Development and Research, American Telephone and Telegraph Company, 1919–34; Bell Telephone Laboratories, 1934–. Mr. Wright is engaged in transmission development of radio systems.

x