THE BELL SYSTEM TECHNICAL JOURNAL

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

TABLE OF CONTENTS AND INDEX

VOLUME XVI

1937

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PRINTED IN U. S. A.

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XVI, 1937

Table of Contents

JANUARY, 1937

APRIL, 1937

JULY, 1937

OCTOBER, 1937

Index to Volume XVI

S

T

V

W

Z

VOLUME XVI JANUARY, 1937 NUMBER 1

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XVI *January, 1937* *No. 1*

A Million-Cycle Telephone System *

By M. E. STRIEBY

ABOUT two years ago a new wide-band system for multi-channel telephone transmission over coaxial cables was described.[1] An experimental system has now been installed between New York and Philadelphia. The various tests and trials which are planned for this system have not been carried far enough to justify a formal technical paper. Meanwhile, the considerable interest that has been aroused in the system has led to this brief statement of its principal features and its general technical performance as so far measured.

The coaxial cable itself has been installed between the long distance telephone buildings in New York and Philadelphia, a distance of 94.5 miles. It has been equipped with repeaters, at intervals of about 10 miles, capable of handling a frequency band of about 1,000,000 cycles.

This million-cycle system is designed to handle 240 simultaneous two-way telephone conversations. Only a part of the terminal apparatus has been installed, sufficient in this case to enable adequate tests to be made of the performance of the entire system. A general view of the New York terminal is shown in Fig. 2. Some preliminary test conversations have been held over the system, both in its normal arrangement for providing New York-Philadelphia circuits, and with certain special arrangements whereby the circuit is looped back and forth many times to provide an approximate equivalent of a very long cable circuit. The performance has been up to expectations, and no important technical difficulties have arisen to cast doubt upon the future usefulness of such systems. Much work remains to be done, however, before coaxial systems suitable for general commercial service can be produced.

The Coaxial Cable

Figure 1 shows a photograph of the particular cable used in this installation. It contains two coaxial units, each having a 0.265-inch inside diameter, together with four pairs of 19-gauge paper insulated wires, the whole enclosed in a lead sheath of 7/8-inch outside diameter.

* Published in *Electrical Engineering* for January, 1937.

[1] "Systems for Wide-Band Transmission Over Coaxial Lines" by L. Espenschied and M. E. Strieby, *Bell Sys. Tech. Jour.*, October, 1934; *Elec. Engg.* (*A. I. E. E. Transactions*), Vol. 53, 1934, pages 1371–80.

Fig. 1—View showing structure of coaxial cable.

Fig. 2—The New York terminal of the coaxial system.

The central conductor of the coaxial units is a 13-gauge copper wire insulated with hard rubber discs at intervals of 3/4 inch. The outer conductor is made up of nine overlapping copper tapes which form a tube 0.02-inch thick; this is held together with a double wrapping of iron tape.

The transmission losses of this coaxial conductor at various frequencies are shown in Fig. 3. This attenuation is about 4 per cent higher than is calculated for a solid tube of the same dimensions and material. Another matter of importance is the shielding obtained from one conductor to the other or to outside interference. Inasmuch as the most severe requirement is that of crosstalk from one coaxial unit to another, this has been used as a criterion of design. Figure 4

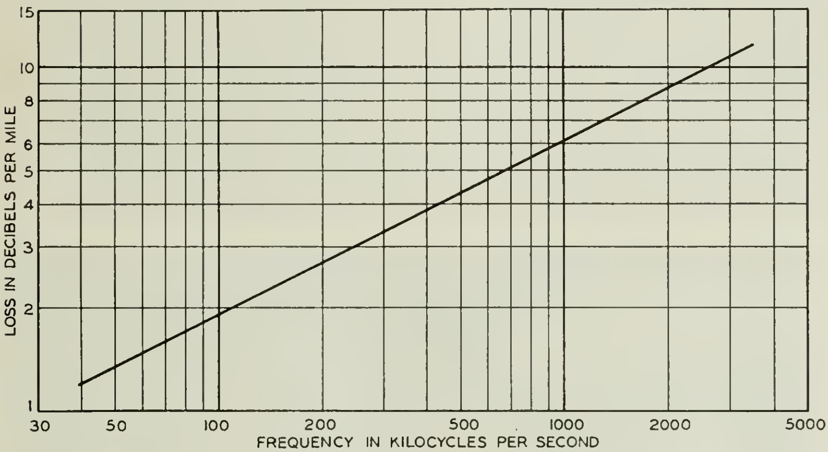


Fig. 3—Attenuation of the coaxial conductor.

shows the average measured high-frequency crosstalk in this particular cable on a 10-mile length without repeaters, both near-end and far-end.

Repeaters

The amplifiers used in this system were designed for a 10.5-mile spacing and a frequency range of 60 to 1024 kc. A total of 10 complete two-way repeaters has been provided including those at the terminals. Two of the intermediate repeaters are at existing repeater stations along the route, the other six being at unattended locations along the line. Four of these are in existing manholes, while the other two are placed above ground for a test of such operation. Figure 5 shows a manhole repeater with the cover removed for routine replacement of vacuum tubes. Figure 6 shows one of the installations above ground.

The measured gain of a typical repeater is shown by the points on the curve of Fig. 7. The curve itself is the line loss that the repeater

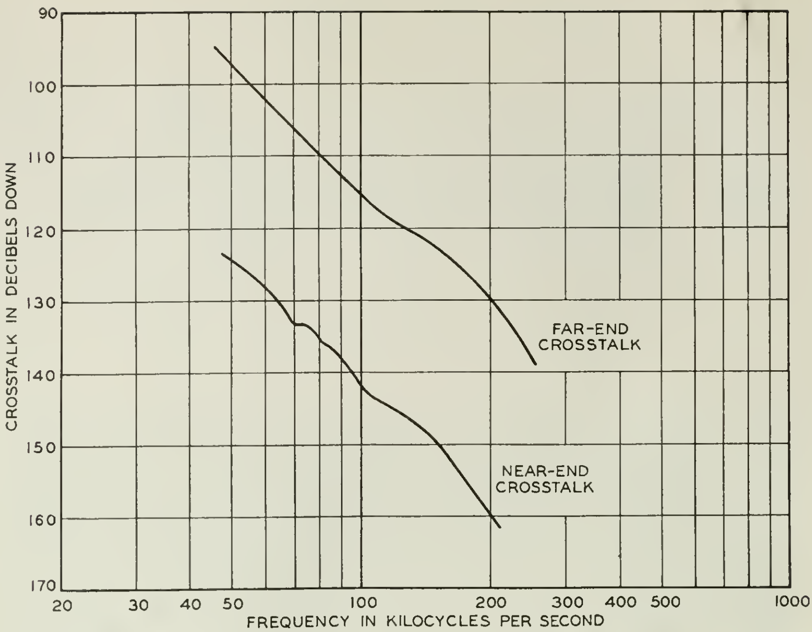


Fig. 4—Crosstalk between the two coaxial conductors in the new cable.

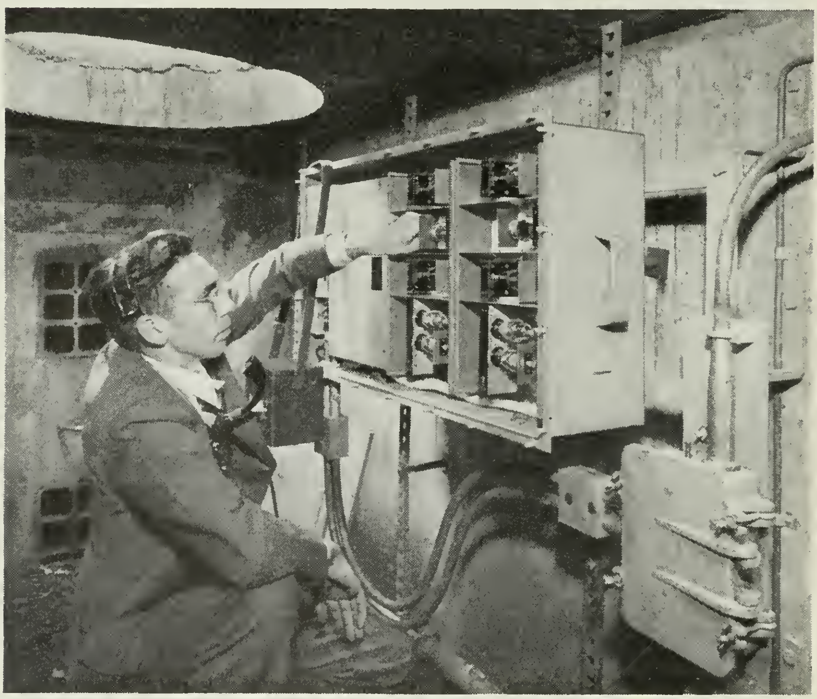


Fig. 5—Million-cycle repeater mounted in a manhole.

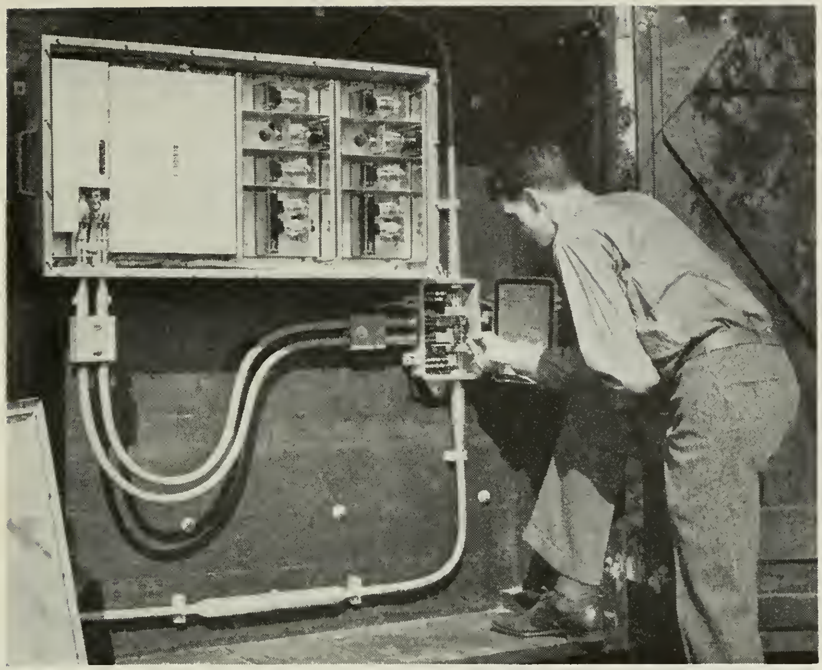


Fig. 6—Installation of coaxial repeater above ground.

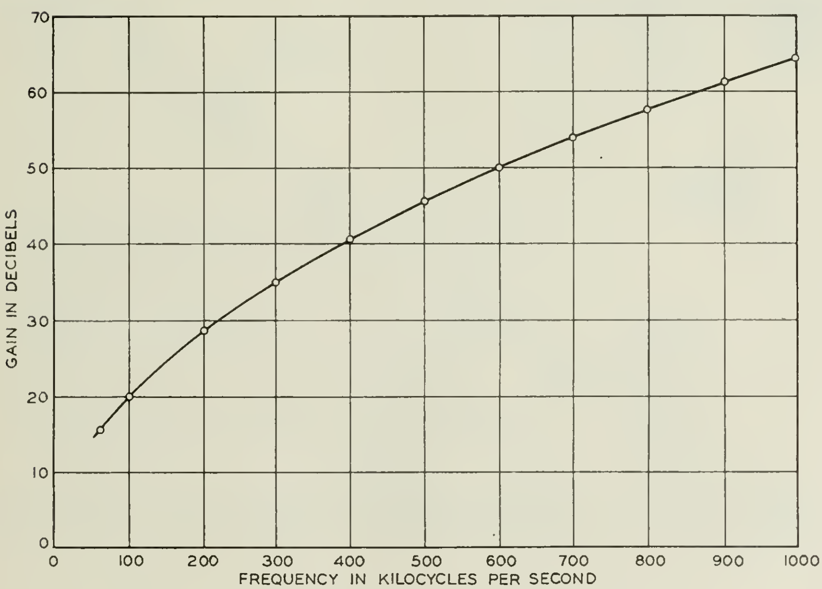


Fig. 7—Gain-frequency characteristic of coaxial repeater.

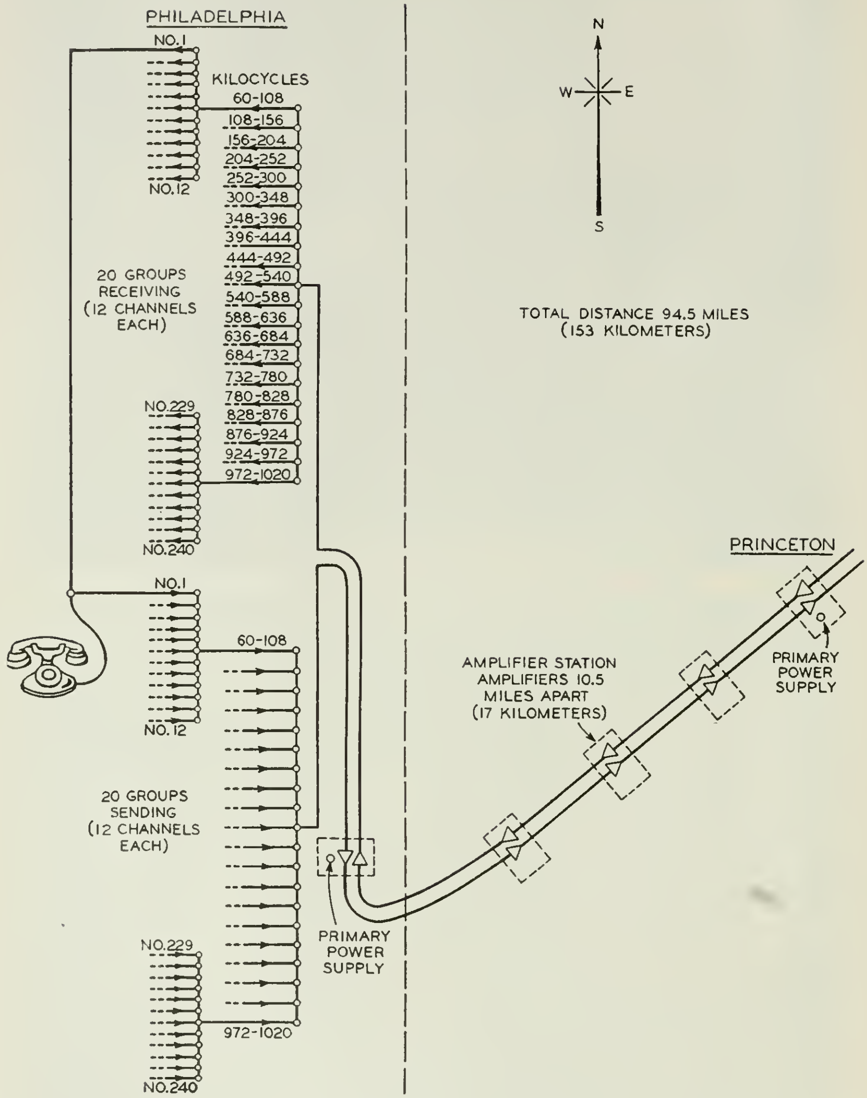


Fig. 8—Frequency characteristic allocation assignments of a typical speech channel. Broad-band system over coaxial cables (240 telephone circuits).—*Continued on page 7.*

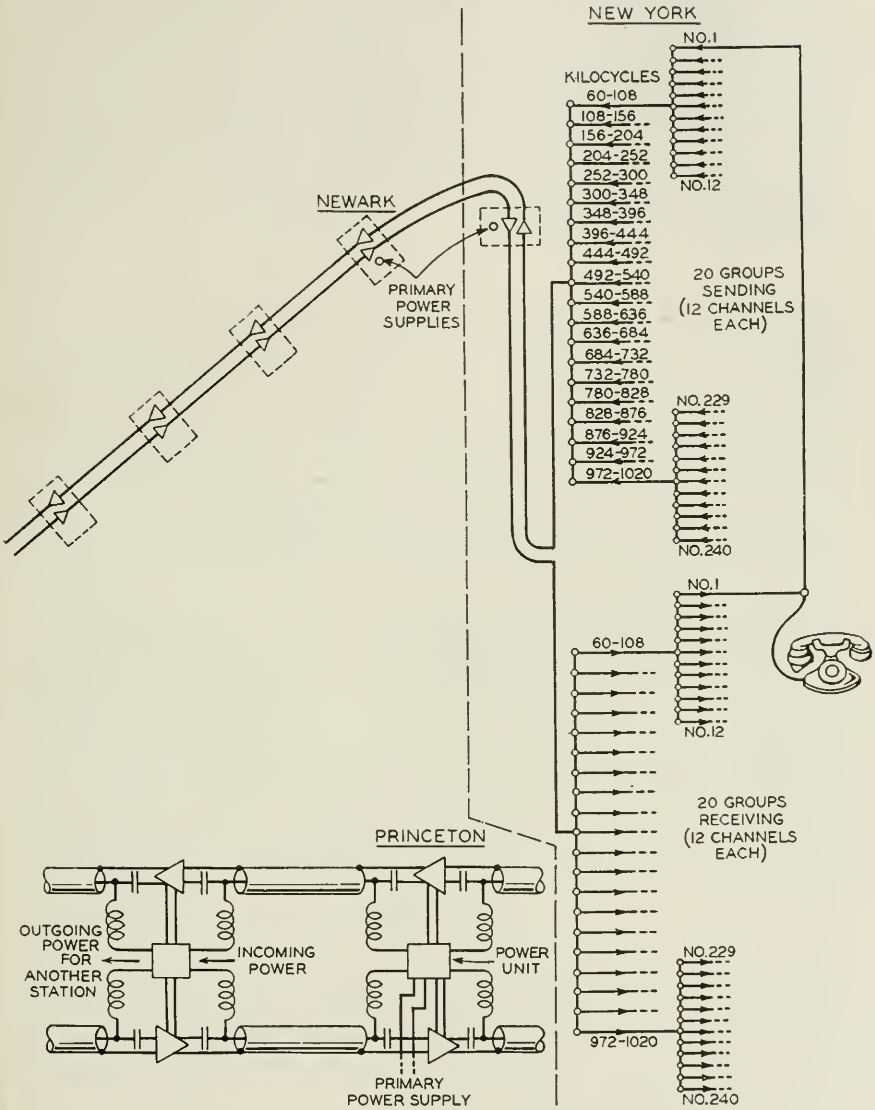


Fig. 8—*Continued from page 6.*

is designed to compensate. Three stages of pentodes are used with negative feedback [2] around the last two stages. Attenuation changes due to temperature of the line are compensated automatically by a pilot channel device which has been installed at every second or third repeater. The regulating mechanism uses four small tubes and is added to the normal repeater when desired. The amplifiers shown in Figs. 5 and 6 are regulating. As the cable is underground, the temperature changes are very slow and but meagre data on the accuracy of compensation are yet available.

Terminals

A schematic diagram of the terminal arrangements for a 240-channel million-cycle system is shown on Fig. 8. In this installation the New York and Philadelphia terminals have each been equipped to handle only 36 two-way telephone conversations. As has been pointed out, the scheme employed involves two steps of modulation, the first of which is used to set up a 12-channel group in the frequency range from 60 to 108 kc. Three such groups have been provided in this installation. In order to transmit at the higher frequencies, a second step of modulation is used in which an entire 12-channel group is moved to the desired frequency location by a "group" modulator. Six such group modulators have been provided at various frequencies throughout the range, including both the top and bottom. Patching facilities have been provided so that any 12-channel group may be transmitted over any one of the high-frequency paths. A typical frequency characteristic of one of the channels is shown in Fig. 9. It may be observed that relatively high quality has been obtained, due largely to the use of quartz crystal electric wave filters, even though the channels are spaced throughout the frequency range at 4000-cycle intervals.

Preliminary Tests

As already noted, various long circuits have been built up by looping back and forth through the coaxial system. One setup over which conversations were successfully carried out consisted of five voice-frequency links in tandem, each link being 760 miles long, giving a total circuit length of 3800 miles. This setup included, in each direction, seventy stages of modulation and the equivalent of 400 line amplifiers, the transmission passing twenty times through each one of the twenty one-way line amplifiers constituting the ten two-way repeaters.

[2] "Stabilized Feedback Amplifiers" by H. S. Black, *Bell Sys. Tech. Jour.*, January, 1934.

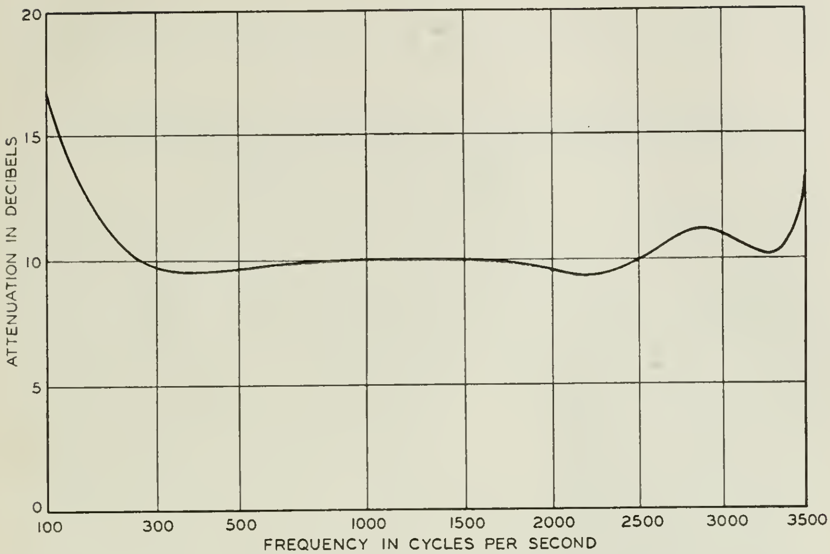


Fig. 9—Schematic diagram of a coaxial million-cycle system showing frequencies assigned to the different channels.

This demonstrated that the complete assemblage of parts, including filters which divide the frequency range into the required bands, modulators which produce the necessary frequency transformations, and amplifiers which counteract the line attenuation, introduced very little distortion. Many problems require further consideration, however, before these systems will be ready for design and production for general use. The final systems must have such refinement that they are suitable for transcontinental distances; the tremendous amplifications needed for such distances must have very precisely designed regulation systems, particularly where aerial construction is involved; noise and crosstalk must not accumulate over the long distances; the repeaters must have such stability and reliability that continuity of service will be assured with hundreds of repeaters operating in series and each repeater handling several hundred different communications simultaneously.

A Power Amplifier for Ultra-High Frequencies *

By A. L. SAMUEL and N. E. SOWERS

A consideration of the special problems encountered at ultra-high frequencies has led to the design of a push-pull power pentode, useful as an amplifier, frequency multiplier, and modulator at frequencies of 300 megacycles per second and below. Unusual construction features include the mounting of two pentodes in the same envelope with interconnected screen and suppressor grids, complete shielding between the input and output circuits with no common leads, and provision for cooling all grids while maintaining extremely small inter-electrode spacings. The electrical characteristics depart from the conventional mainly in the low value of lead inductances and the high value of the grid input resistance at ultra-high frequencies.

The second part of the paper describes a single stage amplifier unit built for testing the tube at frequencies between eighty and 300 megacycles, and the associated apparatus for measuring input impedance, gain, and harmonic distortion. The results given indicate that by using this new tube it is possible to construct stable amplifiers at ultra-high frequencies up to 300 megacycles, having gains of twelve to twenty-five decibels per stage and delivering several watts of useful power. Stability and distortion compare favorably with those obtained from conventional tubes at much lower frequencies.

PART I—THE VACUUM TUBE

By A. L. SAMUEL

WE ARE witnessing a rapid expansion and extension in the use of radio communication. A corresponding extension in the usable portion of the radio-frequency spectrum is highly desirable. With this in mind, special forms of vacuum tubes have already been developed for use as oscillators at frequencies above 100 megacycles.[1, 2] Except at low power levels,[3] amplifier tubes have not been available.

* Presented at Institute of Radio Engineers meeting, New York City, October 7, 1936. Published in *Proceedings I.R.E.*, November, 1936.

[1] M. J. Kelly and A. L. Samuel, "Vacuum Tubes as High-Frequency Oscillators," *Elec. Eng.*, vol. 53, pp. 1504–1517, November, 1934; *Bell Sys. Tech. Jour.*, vol. 14, pp. 97–134, January, 1935.

[2] C. E. Fay and A. L. Samuel, "Vacuum Tubes for Generating Frequencies Above One Hundred Megacycles," *Proc. I.R.E.*, vol. 23, pp. 199–212, March, 1935.

[3] B. J. Thompson and G. M. Rose, "Vacuum Tubes of Small Dimensions for Use at Extremely High Frequencies," *Proc. I.R.E.*, vol. 21, pp. 1707–1721, December, 1933.

It is the purpose of this paper to discuss the problem of amplification at ultra-high frequencies and to describe one form of amplifier tube designed for moderate power in that frequency range.

The Triode as an Amplifier at Ultra-High Frequencies

A simple triode amplifier as used at low frequencies becomes unstable as the operating frequency is increased, exhibiting a tendency to oscillate or "sing" because of the interaction between the input and output circuits. This interaction or "feedback" is, in the main, produced by the grid-plate capacitance of the tube. It may be overcome either by the introduction of a compensating capacitance somewhere in the circuit or by the introduction of an electrostatic shield or screen within the tube envelope. The first expedient, known as neutralization, is employed in the case of a triode. The second expedient results in the screen-grid tetrode. At moderately high frequencies either arrangement may be used.

The conventional triode is unsatisfactory at very high frequencies. The usual capacitance neutralization scheme fails, partly because of the inductance of the tube leads which makes difficult the correct location of the neutralizing capacitance. The appreciable time required for the electrons to traverse the interelectrode spaces within the tube structure makes neutralization more difficult by introducing a shift in the phase of the necessary compensation.

A more serious effect of electron transit time is the marked increase at high frequencies in the input conductance of a tube over the value observed at low frequencies. This effect has been the subject of considerable study.[4, 5, 6, 7] Theory and experiment both agree in relating the input conductance loss to the tube geometry and the applied electrode potentials. The conductance depends upon the electron transit time and increases rapidly with increasing frequency. The transit time may be reduced either by decreasing the electron paths or by increasing the electron velocities. Decreasing the path calls for smaller interelectrode spacings, and increasing the velocity calls for higher electrode potentials. On the other hand, practical considerations limit both the dimensions and the potentials. An optimum design may utilize special mechanical arrangements to combine both expedients.

[4] J. G. Chaffee, "The Determination of Dielectric Properties at Very High Frequencies," *Proc. I.R.E.*, vol. 22, pp. 1009–1020, August, 1934.

[5] F. B. Llewellyn, "Operation of Ultra-High-Frequency Vacuum Tubes," *Bell Sys. Tech. Jour.*, vol. 14, pp. 632–665, October, 1935.

[6] W. R. Ferris, "Input Resistance of Vacuum Tubes as Ultra-High-Frequency Amplifiers," *Proc. I.R.E.*, vol. 24, pp. 82–104, January, 1936.

[7] D. O. North, "Analysis of the Effects of Space Charge on Grid Impedance," *Proc. I.R.E.*, vol. 24, pp. 108–136, January, 1936.

The electron transit time limitation becomes of particular importance at frequencies above one hundred megacycles and sets an upper frequency limit on the useful operation of the usual triode as an amplifier just as it sets the limit at which the tube will operate as an oscillator. Because of the similarity in the special high-frequency requirements, negative grid tubes designed for use primarily as ultra-high-frequency oscillators are good amplifiers at somewhat lower frequencies. The necessity for very careful circuit design and for critical adjustment of the neutralization becomes particularly pronounced when triodes are used as ultra-high-frequency amplifiers.

The Multi-element Tube as an Amplifier at Ultra-High Frequencies

Conventional screen-grid tetrodes and pentodes are also unsatisfactory at very high frequencies. Two factors are again primarily responsible, the one set by the circuit requirements, the other set by the electron transit time. These limitations will be considered in detail.

In the usual radio-frequency amplifiers using tetrodes or pentodes the input and output circuits are tuned to the desired frequency. For most practical purposes the upper limit to the frequency for which these circuits may be tuned is set by the natural period of the circuits formed by the corresponding lead inductances and interelectrode capacitances. Even before this limit is reached the major portions of the tuned circuits are within the tube envelope. Their inaccessibility makes it difficult to obtain effective coupling between amplifier stages.

Interaction between the input and output circuits if excessive may cause "singing." Such interaction is usually due to the residual value of the grid-plate capacitance. Not only must this capacitance be made very low by the proper design of the screen and suppressor grids, but its effective value must remain low at the operating frequency. This latter is realizable only if the screen and suppressor grids can be coupled to the cathode by leads having extremely small inductances. A further desirable feature is that there be no appreciable circuit impedance in the form of lead inductance common to both input and output circuits. The use of short leads is thus seen to be just as important in the design and use of the multi-element tube as it is in the design of the triode.

As in the case of the triode, the electron transit time is effective in limiting the useful frequency range of the multi-element tube. The increase in the input conductance which it introduces is again primarily responsible.

In considering the design of an amplifier tube for ultra-high frequencies, it appeared desirable to select frequency and power levels such that a break from conventional design was inevitable, leaving for future work the satisfactory coverage of the transition region. Since triodes had already been studied as oscillators it was decided to design and construct a pentode. A tentative rating of fifteen watts anode dissipation (per tube) with an operating range up to 300 megacycles was chosen. It was further thought desirable to limit the sum of the grid-to-ground and plate-to-ground capacitances to a value less than eight micromicrofarads in order to facilitate the design of the accompanying circuits.

Preliminary considerations led to the conclusion that the desired results could be best obtained by push-pull operation. In view of the required shortness of leads it seemed logical, if not essential, to inclose both sets of tube elements within one envelope and to provide an internal by-pass condenser between the screen and suppressor grids. It also appeared desirable to design the structure so that a simple extension of the screen-grid element would form a partition separating the input portion of the tube from the output portion. By mounting the tube so that the internal partition forms a continuation of the external partition separating the input and output circuits, quite adequate shielding should be possible. From previous experience, it was concluded that the special frequency requirements for a 300-megacycle amplifying tube would be satisfied by a design patterned after a 600-megacycle oscillator tube.[2]

To summarize, the following construction features were considered desirable:

(1) The mounting of two sets of elements in the same envelope.
(2) A method of interconnecting the two screen grids by a low impedance conductor.
(3) A method of grounding the screen and suppressor grids inside the tube envelope.
(4) Complete shielding between input and output sides of the tube.
(5) The use of extremely short leads.
(6) Means for maintaining very small spacings between the elements.
(7) Provision for adequate cooling of all grids.
(8) Adequate insulation paths to permit a high anode potential.
(9) The absence of any leads common to both input and output circuits.

The first of the experimental tubes designed to have a fifteen-watt dissipation per anode is shown in Fig. 1. It will be noted that a partition divides the envelope into two parts. This partition is in reality double, being made up of two sheets, one being connected to the suppressor grids and the mid-point of the filament circuit and the other being connected to the screen grids. Slots in these sheets provide space to mount the tube elements. The capacitance between the two closely

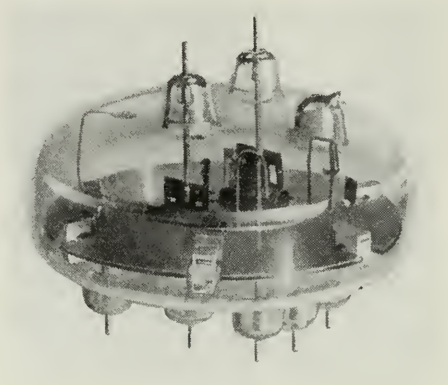


Fig. 1—An early experimental type tube.

spaced sheets forms an effective radio-frequency by-pass condenser between the screen grids and the filaments. Fig. 2 is a section view

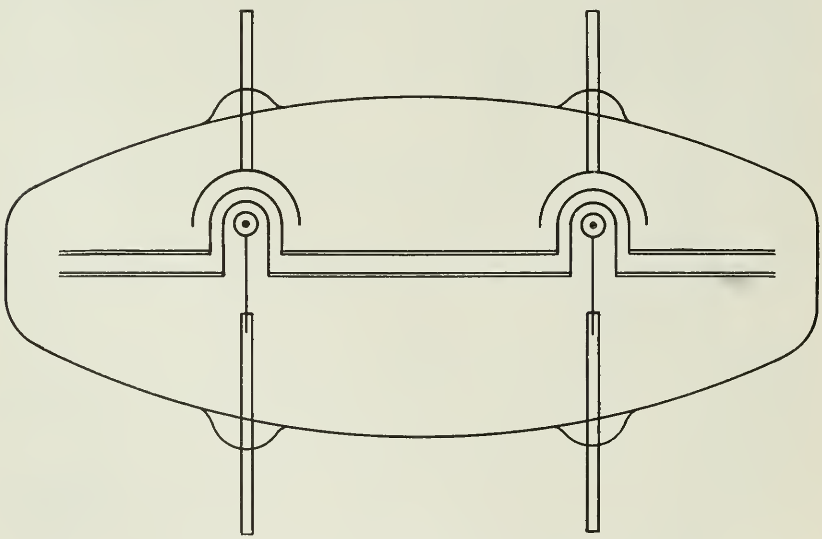


Fig. 2—Section view of the tube shown in Fig. 1.

through the middle of the tube showing the disposition of the elements. While entirely satisfactory from an operating viewpoint, this design proved to be rather difficult to fabricate.

The Ultra-High-Frequency Double Pentode Tube

The successful operation of the experimental models described above indicated the desirability of continuing this line of attack. A complete mechanical redesign to facilitate the fabrication and pumping was undertaken. Fig. 3 is a photograph of this design. Section views are shown in Fig. 4. The large capacitance between the screen and suppressor which characterized previous models was retained in the form of concentric cylinders instead of parallel plates. These cylinders and the flange at one end effectively shield the input and output sides of the tube. The low impedance connection between the two screens provided by these cylinders is an important feature of the design. Adequate cooling of the screen grid is provided by mounting it directly

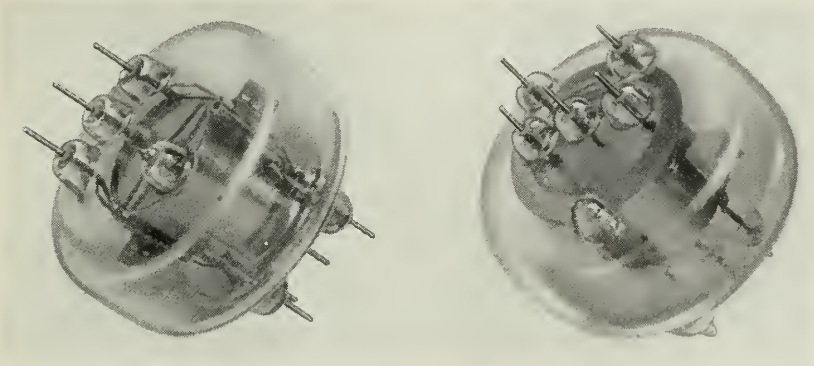


Fig. 3—The ultra-high-frequency double pentode vacuum tube.

in a slot in one of the cylinders. The control grids are of the so-called fin type of construction already employed with considerable success in triode oscillators. They consist of a series of tungsten loops attached to a common cooling fin. This construction makes feasible the use of extremely small dimensions, so that the grid-filament spacing is comparable with the filament diameter. One of these grids is illustrated in Fig. 5. The length of leads has been kept as small as is consistent with mechanical requirements. The longest lead, measured from the mid-point of its attached element to the outside of the envelope, is about three centimeters. Other details of the design are evident from the photograph and the diagram.

Operating characteristics and constants are listed in Table I.

Special attention is directed to the values of interelectrode capacitances and lead inductances. It will be observed that while the inter-

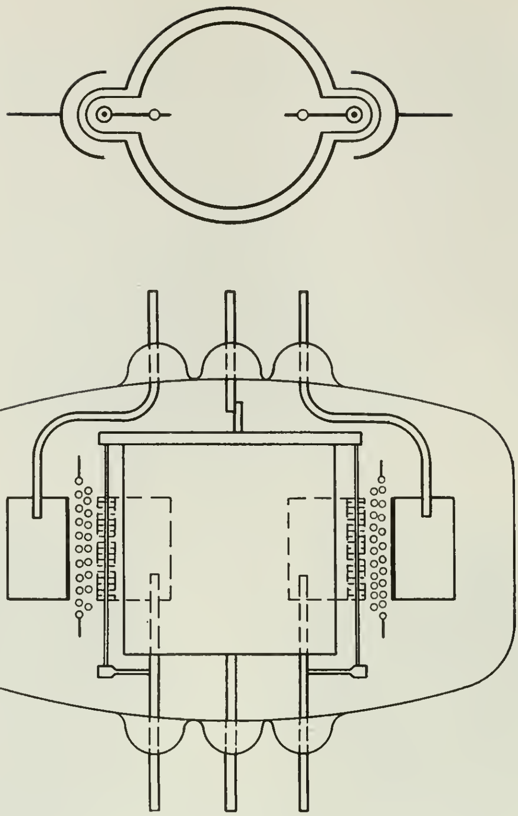


Fig. 4—Section view of the double pentode tube.

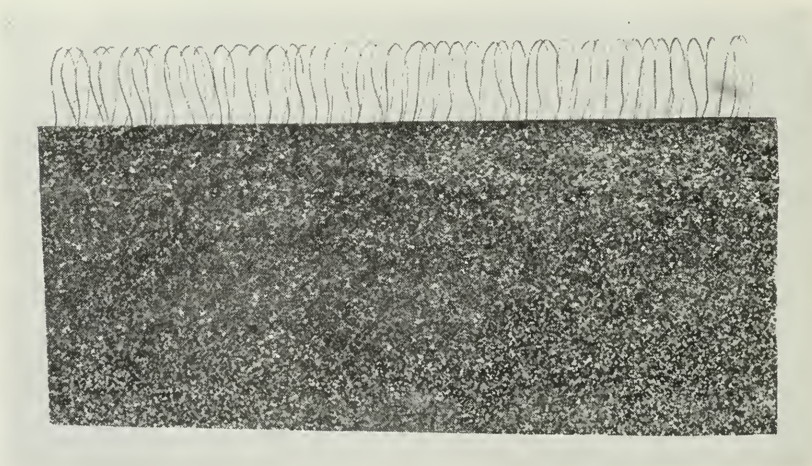


Fig. 5—One of the control grids used in the double pentode tube.

TABLE I

Operating Characteristics and Constants of the Double Pentode Tube

Filament current (each side)	5.0 amperes
Filament potential (each side)	1.5 volts
Rated anode dissipation (each anode)	15 watts
Rated screen dissipation (each side)	5 watts
At anode and screen potentials of 500 volts and anode current of 0.030 ampere—characteristics of each side	
Transconductance	1250 micromhos
Anode resistance	200,000 ohms
Normal control grid potential	−45 volts
Interelectrode capacitances (when properly mounted)	
Direct control grid to control grid	0.02 micromicrofarad
Direct plate to plate	0.06 micromicrofarad
Total control grid to ground (each side)	3.8 micromicrofarads
Total plate to ground (each side)	3.0 micromicrofarads
Control grid to plate (each side)	0.01 micromicrofarad
Lead inductances	
Total grid to grid	0.07 microhenry
Total plate to plate	0.08 microhenry
Rating as class A amplifier	
Maximum direct plate potential	500 volts
Maximum direct screen potential	500 volts
Maximum continuous plate dissipation (each)	15 watts
Maximum continuous screen dissipation (total)	10 watts
Maximum output at 150 megacycles with distortion down 40 decibels	1 watt
Nominal stage gain at 150 megacycles	20 decibels
Nominal control grid potential	−45 volts
Rating as class B amplifier	
Maximum direct plate potential	500 volts
Maximum direct screen potential	500 volts
Maximum space current (total)	150 milliamperes
Maximum continuous plate dissipation (each)	15 watts
Maximum continuous screen dissipation (total)	10 watts
Maximum output at 150 megacycles	10 watts

electrode capacitances are low they have not been reduced in proportion to the reduction in operating wave length. The more important feature is the reduction of the lead inductances. Tabulation of the value of these inductances represents a departure from the conventional practice and is made desirable by their relative importance.

A feature of the design not directly measurable under actual operating conditions but nevertheless responsible for some of the improvement over the more conventional designs is the reduction of an auxiliary dielectric material and the attending dielectric losses that occur at ultra-high frequencies.

The usual static characteristics given in Figs. 6 and 7 are seen to resemble those of the conventional pentode. For a tube which is to be used at ultra-high frequencies, certain other characteristics have a much greater significance. One of the most important of these is the active grid loss which as already mentioned comes about because of the appreciable electron transit time. Fig. 8 gives a plot of the push-pull input shunting resistance of this tube as a function of frequency. The

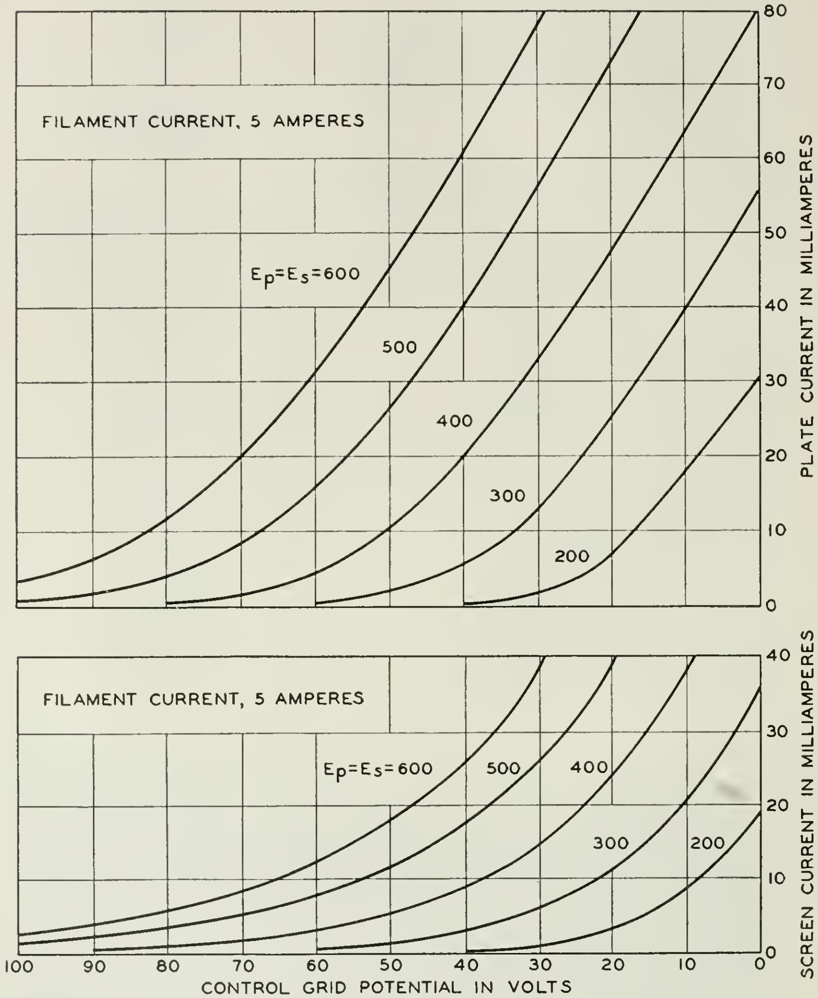


Fig. 6—Mutual characteristics of the double pentode tube.

value of 30,000 ohms at 150 megacycles is to be compared with 2000 ohms, a typical value for two conventional tubes in push-pull. At 300 megacycles the input resistance of the twin pentode is still above 5000 ohms while for conventional tubes it is so low as to make them completely inoperative. The variation in the input resistance with the operating conditions of the tube for a constant frequency of 150 megacycles is shown in Fig. 9. It is evident that if a high value of input resistance is to be realized, high anode potentials with low space cur-

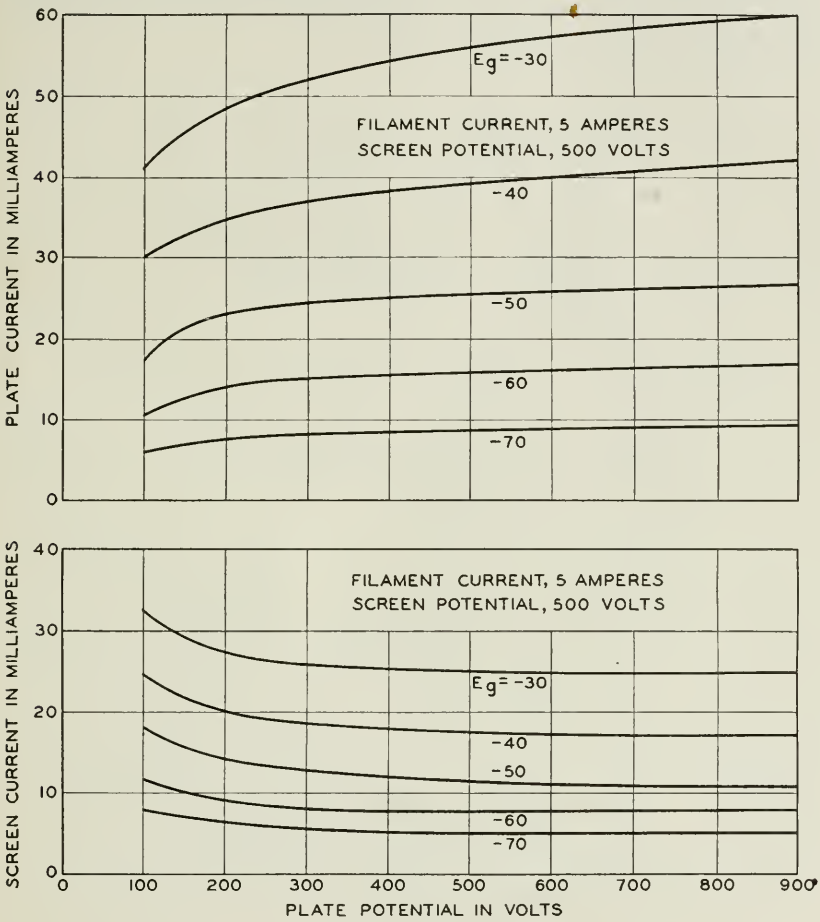


Fig. 7—Anode characteristics of the double pentode tube.

rents must be used. The reduction in the filament grid spacing made possible by the unusual construction is in a large measure responsible for the improvement in the input resistance just noted.

A characteristic measurable only at the operating frequency is the interaction between the input and output circuits which results from the residual value of the grid-plate capacitance. This reaction differs from that predicted from the low-frequency capacitance measured on a cold tube because of the inductance of the screen-grid lead and because of the electron space charge. The reaction can be measured by observing the variation in the input impedance resulting from a variation in the tuning and loading of the output circuit. Experimentally determined values are given in Fig. 10.

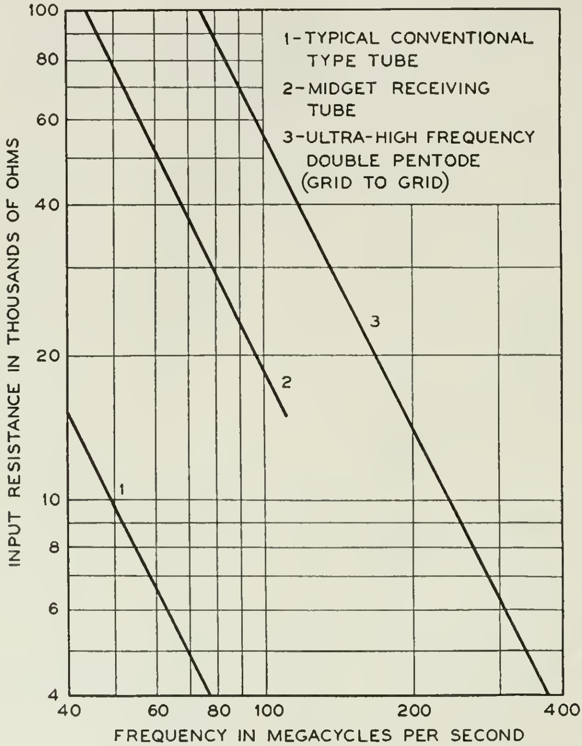


Fig. 8—The input resistance as a function of frequency.

The double pentode tube has been found useful as a high quality class A amplifier, a class B amplifier, a frequency multiplier, and as a modulator at frequencies of 300 megacycles per second and below. Its performance in these various modes of operation is quite comparable to the performance of conventional pentodes of similar ratings at much lower frequencies. Stable operation with some gain has been obtained at frequencies as high as 500 megacycles. Because of the increased importance at ultra-high frequencies of circuit design in the over-all performance of an amplifier or modulator, such tests cannot be considered as a definite measure of the capabilities and limitations of the tube but they indicate what has already been accomplished.

When operating as a class A amplifier at 150 megacycles an output of one watt is obtained with the distortion forty decibels below the fundamental.

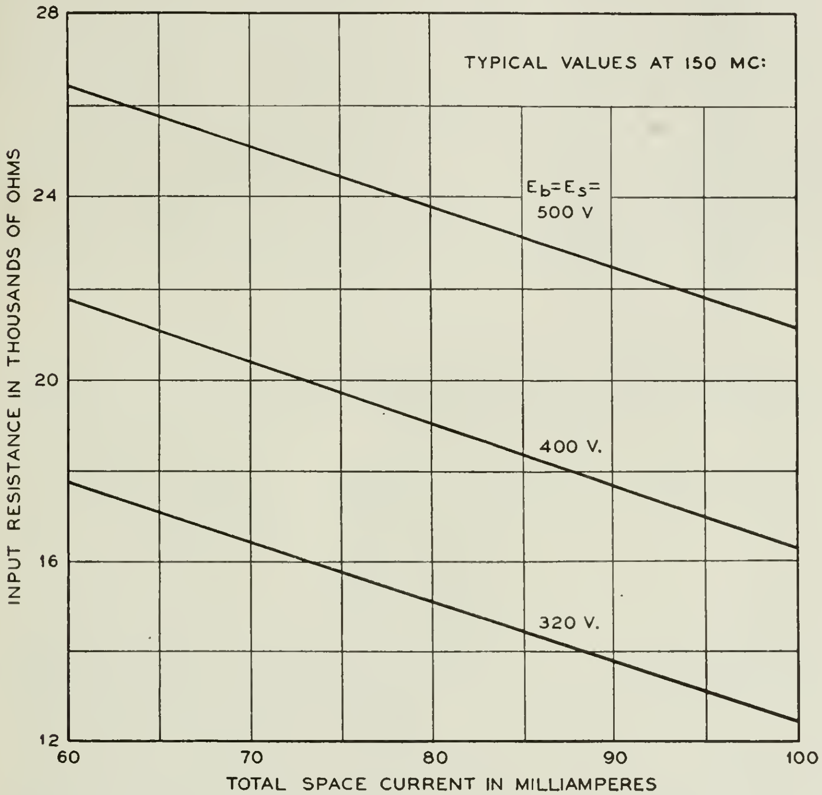


Fig. 9—The variation in input resistance with operating conditions at 150 megacycles.

Under these conditions the stage gain is twenty decibels. Outputs of ten watts with a plate efficiency of sixty to seventy per cent and a gain of ten decibels are secured with class B operation. Experimental results confirming these statements together with a discussion of the principles of circuit design and the technique of measurements are given in the accompanying paper by N. E. Sowers.

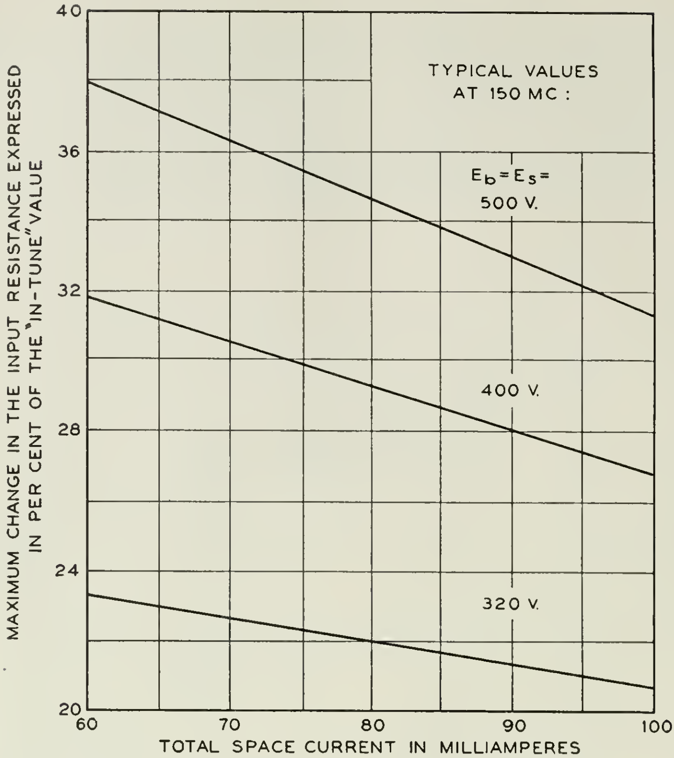


Fig. 10—The input-output reaction at 150 megacycles.

Conclusion

The development of this ultra-high-frequency pentode demonstrates that amplifier tubes of the negative grid type are usable at higher power levels and frequencies than have been reported previously. The extension of the principles underlying the design of this tube to the design of a tube with approximately ten times the output is now being considered. This type of development removes a practical barrier which has, up to the present, prevented the successful utilization of this frequency range.

PART II—THE CIRCUIT

By N. E. SOWERS

Introduction

In the first section of this paper A. L. Samuel has described the development of a push-pull pentode designed to function as a stable amplifier at frequencies up to at least 300 megacycles. It is the purpose of the present section to describe the methods and apparatus used in testing this tube and to set forth the results of some of the tests.

An attempt to study the operating characteristics of an amplifier tube at ultra-high frequencies brings up many new problems. Such fundamental properties of the tube as amplification factor, transconductance, and plate impedance do not convey as much information about the behavior of the tube at these frequencies as they do at lower frequencies. The presence of unavoidable stray inductances and capacities makes it much more difficult to separate tube problems from circuit problems. Consequently, at ultra-high frequencies we are virtually forced to consider the tube and its associated circuits as comprising a single piece of apparatus. If the circuit design is carefully made the stray inductances and capacitances can be greatly reduced in magnitude and so localized that their effects upon the over-all performance of such a piece of apparatus can, to a certain extent, be computed.

Circuit Design

Some idea of the extreme attention to detail required in designing amplifier circuits for use at ultra-high frequencies may be gained from the following considerations. Computations indicate that even with the tuned plate and grid circuits placed as close as physically possible to one of these push-pull pentodes, at 300 megacycles, the radio-frequency voltage actually applied to the grids of the tube may be as much as twenty-five per cent greater than the voltage developed across the tuned grid circuit. At the same time the load presented to the tube plates may be as much as twice the load actually present across the tuned plate circuit. These discrepancies are a direct result of the inductance of grid and plate leads which, in the case of this new tube, have already been reduced well nigh to the minimum possible.

In studying the performance of these tubes we wished to be able to check experimental results against theory at every possible point. Consequently the simplest auxiliary circuits were chosen, namely, shunt-tuned antiresonant circuits from grid to grid and from plate to plate, with screens and filaments by-passed as directly as possible to ground. In their mechanical design these circuits embody a number of features intended to reduce and localize stray inductances and capacities, into the details of which it is not possible to go at present. A simple arrangement was evolved to provide a maximum of convenience and flexibility for experimental work. The single stage amplifier unit consists of three sections, an input circuit section, a tube housing section, and an output circuit section. This arrangement permits tubes to be changed with a minimum of disturbance to the circuits. During experimental work it is almost inevitable that circumstances will arise calling for major changes in the nature of the circuits, or the size, shape, and lead arrangement of the tubes. This sectional construction provides the necessary flexibility to take care of such needs, as the construction and substitution of appropriate new sections would permit the experimental work to proceed with a minimum of delay. To facilitate the operation of several units in tandem for tests on a multistage amplifier, each section is provided with its own power supply jacks so that the only longitudinal connections required

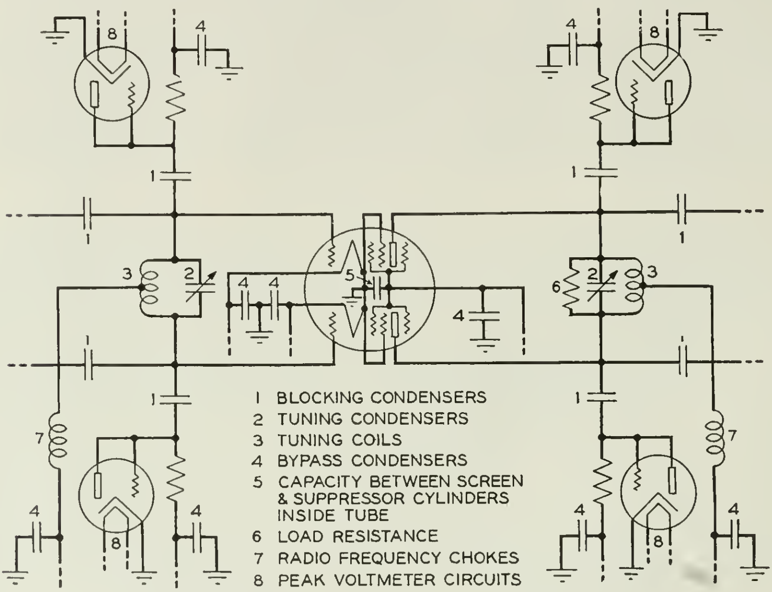


Fig. 11—Circuit diagram of single stage test amplifier.

within the sections are those between tube leads and the circuits. These connections are so arranged as to be very easily broken when sections are to be separated. Each circuit section has built into it a pair of peak voltmeters for indicating the radio-frequency voltage developed across the tuned circuit. These voltmeters consist of RCA type 955 tubes used as diode rectifiers in the familiar self-biased peak voltmeter circuit. Fig. 11 shows the circuit in schematic form. Fig. 12 shows an experimental two-stage amplifier constructed in substantially the same fashion as the test circuit, but without the sectionalizing feature.

The desire to reduce the length of all leads to a minimum has naturally resulted in bringing the tuned circuits rather close to the sides of the circuit housings. Nevertheless care and attention to

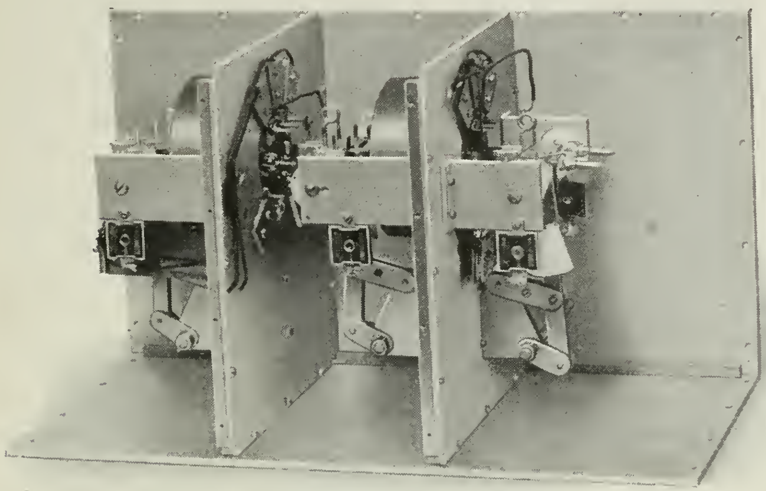


Fig. 12—An experimental two-stage one-meter amplifier using two of the earlier type push-pull pentode tubes.

detail in the circuit design have enabled the stray capacities to be kept down to satisfactory values. Fig. 13 shows in schematic form one of

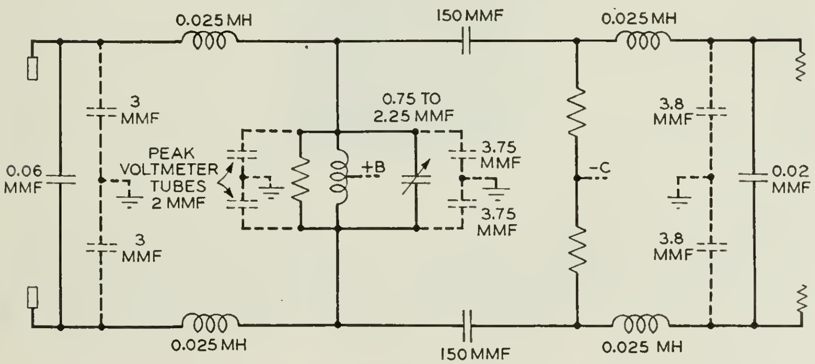


Fig. 13—Diagram of interstage circuit.

these circuits employed as the interstage circuit between two of these push-pull pentodes, all of the important inductances and capacities being included.

Input Impedance Measurement

One of the factors which effectually limits the performance of a vacuum tube at ultra-high frequencies is the internal grid resistance or active grid loss. Consequently, this factor is of extreme interest in the development of amplifier tubes for use in the ultra-high-frequency range and much of this work has centered around the development of apparatus and technique for rapidly and accurately measuring these input resistances. The method employed has been the simple resistance substitution method used by Crawford.[8]

An adjustable quarter-wave Lecher frame is provided with suitable means for inducing a radio-frequency voltage across it and a suitable detector for indicating the current flowing at the short-circuited end. A calibration is made by noting the detector indication corresponding to various known resistances connected across the open end of the frame, with the input voltage held constant. The input circuit of the tube under test is then connected to the end of the Lecher frame in place of the calibrating resistors and the detector indications corresponding to various voltages and loads applied to the tube are noted. Since the Lecher frame is initially tuned to the operating frequency, and when the tube input circuit is attached the circuit itself is retuned for resonance, it follows that the quantity actually measured is the effective resistance across the tuned circuit, including both the circuit losses and the active grid loss of the tube. It is of course possible to determine the circuit losses separately and to compute the contribution to the total resistance offered by the tube losses, and also to compute the active grid loss existing directly at the grids of the tube, taking into account the impedance transformation existing between the tube grids and the tuned circuit, brought about by the lead inductances. Practically, however, the total effective shunt resistance across the tuned circuit as actually measured is a more significant quantity, as this quantity determines more or less directly the gain which can be obtained from a multistage amplifier. It frequently happens that changes in the voltages applied to the tube produce small changes in the reactive component of the input impedance. These may be taken into account by noting the changes in grid circuit tuning required to maintain resonance. These changes are usually so small as to be of only minor interest.

The Lecher frame used in these measurements is shown in Fig. 14. The plate bridging the frame nearest the open end carries the detector, an RCA type 955 tube set into the plate. The grid of this tube is

[8] A. B. Crawford, "Input Impedance of Vacuum Tube Detectors at Ultra-Short Waves" (Abstract), *Proc. I.R.E.*, vol. 22, pp. 684–685, June, 1934.

coupled to the frame by means of a small rectangular single turn loop mounted just beneath and quite close to the bars at the short-circuited end. The second plate bridging the bars, in conjunction with the electrostatic screen between the bars and the input coupling coil, aids in

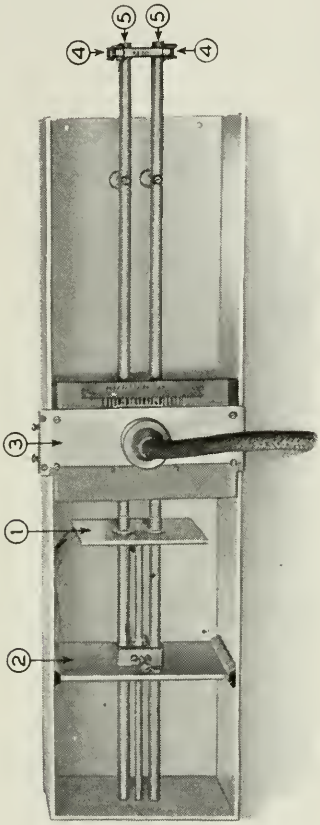


Fig. 14—Photograph of impedance measuring Lecher frame.

1. Short-circuiting bridge carrying detector tube and detector coupling coil.
2. Auxiliary bridge for breaking up unbalance currents flowing on the frame.
3. Input circuit. Note electrostatic screen between frame and input coil mounted on end of flexible transmission line leading to driving oscillator.
4. Clips carrying calibrating resistors.
5. Jacks into which plugs on amplifier input circuit fit.

eliminating any unbalance of the currents flowing in the two sides of the frame. The aluminum trough surrounding the frame provides sufficient shielding to render the apparatus virtually immune to the operator's body capacity effects. The whole resistance measuring setup is remarkably stable and satisfactory to operate. Resistance measurements on a given tube at specified operating points can be repeated with a precision of two or three per cent even when weeks elapse between measurements.

In addition to being a function of frequency, the input resistance of one of these tubes is also a function of all of the operating conditions, that is, applied voltages, plate circuit tuning, and load. In Table II are shown values of this input resistance for a typical tube at several frequencies and over a considerable range of operating conditions. Because of the large number of variables which affect this input resistance it is difficult to devise any way of plotting up these data so as to give a comprehensive picture of tube performance.

The variation of input resistance with plate circuit tuning has, for this design, consistently been of the form illustrated in Fig. 15. How-

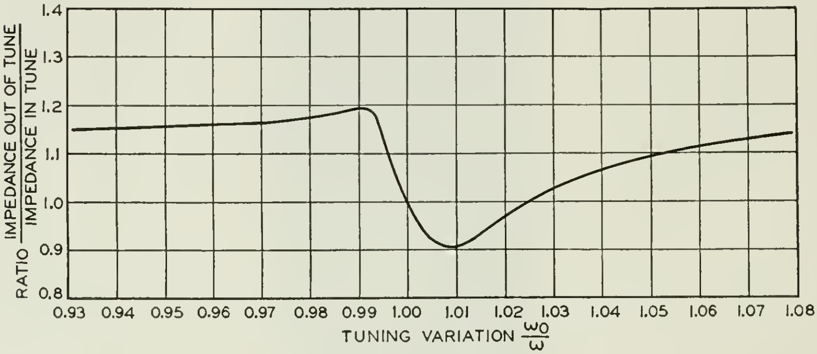


Fig. 15—Reaction curve.

ever, the relations between maximum, minimum, and "in-tune" values vary somewhat with frequency, operating conditions, and plate load. Also, as may be expected, they vary somewhat in different tubes which have been made up with various grid and screen spacing, etc. A convenient numerical measure of the magnitude of this reaction is obtained by taking the difference between the maximum and minimum values at any specified operating point and dividing this difference by the "in-tune" value. This reaction ratio will also be found listed in Table II for various operating conditions.

Gain Measurements

The measurement of the voltage gain of an amplifier stage containing one of these tubes is a relatively simple matter. As stated in the description of the circuit, provision is made for connecting a peak voltmeter directly to each tuning condenser plate in both plate and

TABLE II

INPUT RESISTANCE AND REACTION RATIO AS A FUNCTION OF FREQUENCY AND APPLIED VOLTAGES AND CURRENTS
PLATE CIRCUIT LOAD 15,000 OHMS

	$f=150$ megacycles			$f=200$ megacycles			$f=250$ megacycles			$f=300$ megacycles		
$I_P + I_S$ mils	64	80	100	64	80	100	64	80	100	64	80	100
$E_P = E_S = 320$ volts, resistance	17200	15050	13500	9900	8650	7750	6600	5800	5200	4900	4300	3850
$E_P = E_S = 320$ volts, reaction ratio	0.230	0.220	0.207	0.197	0.185	0.174	0.159	0.147	0.144	0.133	0.128	0.177
$E_P = E_S = 400$ volts, resistance	21100	19050	16250	12150	10950	9350	8150	7350	6250	6000	5450	4600
$E_P = E_S = 400$ volts, reaction ratio	0.311	0.289	0.271	0.267	0.251	0.241	0.227	0.211	0.192	0.175	0.174	0.163
$E_P = E_S = 500$ volts, resistance	26000	23700	21300	14950	13600	12250	10000	9150	8250	7400	6750	6050
$E_P = E_S = 500$ volts, reaction ratio	0.373	0.346	0.314	0.324	0.298	0.273	0.270	0.257	0.232	0.216	0.207	0.190

$$\text{Reaction Ratio} = \frac{(\text{maximum resistance} - \text{minimum resistance}) \text{ as plate circuit is tuned}}{\text{resistance with plate circuit in tune}}$$

grid circuits so that the applied grid drive and developed plate voltages may be read directly. Of course, the gain figure arrived at in this manner is an over-all factor, a function both of tube conditions and circuit construction and loading. Nevertheless, it is a satisfactory figure of merit for the stage. In Table III are shown these gain figures for a typical tube under various conditions.

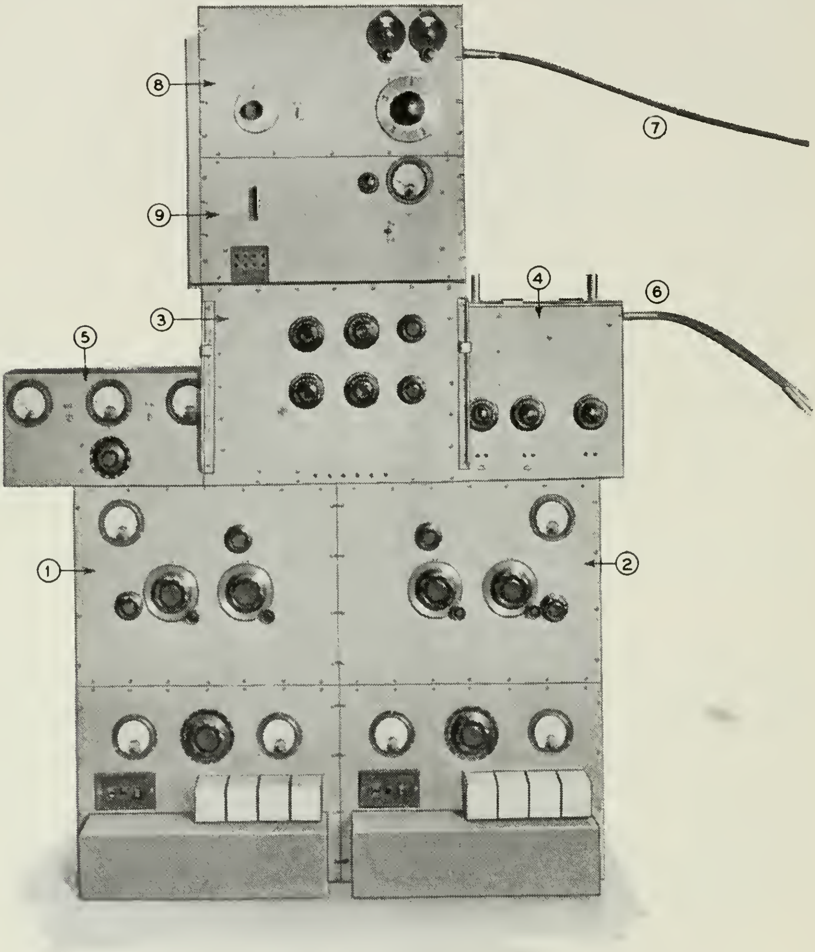


Fig. 16—Distortion measuring equipment.
Nos. 1 and 2—Signal oscillators.
3—Capacitance bridge.
4—Auxiliary amplifier.
5—Power supply unit for auxiliary amplifier.
6—Transmission lines to tube under test.
7—Transmission line from tube under test to radio receiver.
8—Beating oscillator, first detector and attenuator.
9—Intermediate amplifier and second detector of receiver.

TABLE III

STAGE GAIN IN DECIBELS AS A FUNCTION OF CURRENT, VOLTAGE, PLATE CIRCUIT LOAD, AND FREQUENCY

		$f=150$ megacycles			$f=200$ megacycles			$f=250$ megacycles			$f=300$ megacycles		
$I_P + I_S$	Mils	64	80	100	64	80	100	64	80	100	64	80	100
	Plate Cct. Load. Ohms												
$E_P = E_S$ = 320 volts	unloaded*	25.2	26.1	26.5	23.6	24.5	24.9	23.0	23.9	24.3	23.1	24.0	24.3
	15000†	18.2	19.1	19.5	17.7	18.6	19.0	18.0	18.9	19.3	18.7	19.6	20.0
	5000†	11.8	12.7	13.1	11.6	12.5	12.9	12.5	13.4	13.8	13.6	14.5	14.9
$E_P = E_S$ = 400 volts	unloaded*	24.6	25.5	25.6	23.0	23.9	24.0	22.4	23.3	23.4	22.5	23.4	23.5
	15000†	17.6	18.5	18.6	17.1	18.0	18.1	17.4	18.3	18.4	18.1	19.0	19.1
	5000†	11.2	12.1	12.2	11.0	11.9	12.0	11.9	12.8	12.9	13.0	13.9	14.0
$E_P = E_S$ = 500 volts	unloaded*	23.9	24.7	25.2	22.3	23.1	23.6	21.7	22.5	23.0	21.8	22.6	23.1
	15000†	16.9	17.7	18.2	16.4	17.2	17.7	16.7	17.5	18.0	17.4	18.2	18.7
	5000†	10.5	11.2	11.8	10.3	11.0	11.6	11.2	11.9	12.5	12.3	13.0	13.6

* Except for peak voltmeters.
† In addition to peak voltmeters.

Distortion Measurements

One of the quantities of fundamental interest in studying class A amplifiers is the amount of distortion to the applied signal generated in the tube. The technique of making distortion measurements at audio and carrier frequencies is well understood and presents no outstanding problems. However we would not expect distortion measurements made at low frequencies to have any significant application to ultra-high-frequency operation. Since the input resistance of a tube at these frequencies is obviously a function of the various voltages and currents we should expect this input resistance to vary throughout the radio-frequency cycle, that is, to be essentially nonlinear. The question of whether or not this nonlinearity is of sufficient magnitude to cause trouble can best be answered by making direct distortion measurements at the ultra-high frequencies. After some consideration of the various methods of measuring distortion we have chosen the two-tone method as being the most promising. In this method two independent frequencies suitably chosen in the transmission band of the amplifier are fed into the amplifier and the amplitudes of these two tones and such of their modulation products as are of interest are measured in the output of the amplifier by means of a suitable voltage analyzer. In the present case the "tones" are actually a pair of ultra-high-frequency signals. The principal precaution which must be taken in this method is to prevent the oscillators which supply the driving frequencies from reacting on each other and producing distortion products ahead of the amplifier under test. In the present case we have taken care of this requirement by using relatively high powered driving oscillators, very well shielded, from which only very small amounts of power are taken by means of very loosely coupled and electrostatically screened coupling coils. The outputs of the two oscillators are still further isolated from each other by connecting each across opposite diagonals of a balanced capacity bridge and taking off the voltage to drive the circuit under test across one arm of the bridge. A small amount of the voltage developed in the output circuit of the amplifier under test is picked up by a small coupling coil and fed into a voltage analyzer by means of which the relative amplitudes of the testing frequencies and their modulation products may be measured. This voltage analyzer consists of a high gain superheterodyne receiver having a rather sharply tuned, intermediate-frequency amplifier and an extremely precise tuning arrangement on the beating oscillator. The intermediate-frequency amplifier contains an attenuator which, in conjunction with the second detector current meter, permits the relative amplitude of signals to be measured.

The oscillators are push-pull tuned-plate—tuned-grid oscillators employing Western Electric type 304-A tubes with about 900 volts on their plates. These oscillators each deliver about twenty-five watts of radio-frequency power, nearly all of which is dissipated in a resistance load inside the shielding compartments. The receiver (voltage analyzer) has approximately one hundred decibels gain and a ninety-three-decibel attenuator adjustable in one-decibel steps so that measurements over a very wide range of amplitudes are possible. It was found desirable to interpose an additional amplifier (also using these push-pull pentodes) between the output of the bridge and the tube and circuits under test. Of course this amplifier introduces a possible source of distortion ahead of the circuit under test and care must be taken to operate it under such conditions that an adequate margin exists between distortion level measured at its output and distortion level existing at the output of the tube under test.

In Table IV are shown the results of distortion measurements made under several typical sets of operating conditions.

TABLE IV

RATIO OF AMPLITUDE OF THIRD ORDER MODULATION PRODUCTS TO AMPLITUDE OF ONE OF TWO EQUAL TEST FREQUENCIES

Frequency = 80 megacycles

E_P, E_S Volts	E_G Volts	I_P Mils	I_S Mils	Distortion ratio, decibels at 0.33 watt * output	Distortion ratio, decibels at 0.75 watt * output
320	−27.4	43.5	19.5	−52	−44
320	−23.8	54.0	26.0	−54	−46
320	−19.0	66.5	33.5	−56	−48
400	−38.3	44.0	22.0	−53	−44
400	−34.5	55.0	25.0	−54	−45
400	−29.5	68.5	31.5	−57	−49
500	−53.5	45.5	19.5	−57	−50
500	−49.0	56.0	24.0	−58	−50
500	−44.2	70.0	30.0	−56	−48

* For single frequency whose amplitude is the sum of the amplitudes of the two test frequencies.

OTHER APPLICATIONS

A study of the performance of these tubes as class B amplifiers, as harmonic generators, and as modulators apparently presents no serious additional problems and requires very little in the way of additional new technique. Tests indicate that in the neighborhood of 150 megacycles the performance of these tubes in such modes of operation is comparable to that of conventional pentodes in the ordinary short-wave range. In a two-stage amplifier using these tubes, with the first tube working as a class A amplifier and the second tube under class *B* conditions an output of over ten watts has been obtained with a second stage plate efficiency of around seventy per cent and with an over-all voltage gain for the two stages of twenty-four decibels. Using the first tube as a harmonic generator, driven at fifty megacycles, and the second tube as a class B amplifier, over six watts of 150-megacycle power have been obtained with an over-all voltage gain from fifty-megacycle input to 150-megacycle output of about four decibels.

Conclusions

It is often little realized how completely our present highly developed technique of making communications measurements depends upon our ability to set up stable and reliable amplifiers at the frequencies we wish to use. We are now in a position to set up such amplifiers in the ultra-short-wave range; amplifiers of sufficient gain, stability, and most important, of sufficient power handling capacity to enable us to make many of the measurements we may wish, at low enough impedance levels to minimize some of the effects of unavoidable stray inductances and capacitances in our circuits and at high enough power levels to make practicable the use of simple and reliable, and almost necessarily rather insensitive measuring apparatus. Furthermore, our experience in this work indicates that it is not necessary to modify drastically our experimental procedures when we move into the ultra-short wave field. Much more care in circuit design is required, but with more attention to details formerly unimportant, much of the background of electrical measuring technique becomes, with the advent of this new tool, available in the ultra-short-wave range.

The Physical Reality of Zenneck's Surface Wave

By W. HOWARD WISE

The first part of the paper shows that a vertical dipole does not generate a surface wave which at great distances behaves like Zenneck's plane surface wave. In Parts Two and Three it is shown that it is not necessary to call upon the Zenneck wave to explain the success of the wave antennas.

IN 1907 [1] Zenneck showed that a plane interface between two semi-infinite media could support, or guide, an electromagnetic wave which is exponentially attenuated in the direction of propagation along the interface and vertically upwards and downwards from the interface. Zenneck did not show that an antenna could generate such a wave but, because this "surface wave" seemed to be a plausible explanation of the propagation of radio waves to great distances, it was commonly accepted as one of the components of the radiation from an antenna.

After Sommerfeld [2] formulated the wave function for a vertical infinitesimal dipole as an infinite integral and noted that the integral around the pole of the integrand is the wave function for a surface wave, which at great distances is identical with the Zenneck wave, no one questioned the reality of Zenneck's surface wave.

There has been recently pointed out by C. R. Burrows [10] the lack of agreement between various formulas and curves of radio attenuation over land when the dielectric constant of the ground must be taken into account. The values of Sommerfeld [2] and Rolf [5] are stated to differ from those of Weyl [7] and Norton [9] by an amount just equal to the surface wave of Zenneck. Burrows [10] presents experimental data supporting the correctness of the Weyl-Norton values and raises a question as to whether a surface wave really is set up by a radio antenna. A vertical current dipole does not generate a surface wave which at great distances behaves like Zenneck's plane surface wave. Theoretical and numerical evidence leading to this conclusion is presented in Part One of this paper. A contemporary theoretical investigation by S. O. Rice * leads to the same conclusion.

The reader familiar with wave antennas will at once ask why the wave antennas seem to justify the Zenneck surface wave theory by means of which they were conceived and designed if there is no surface

* "Series for the Wave Function of a Radiating Dipole at the Earth's Surface," this issue of the *Bell Sys. Tech. Jour.*

wave. In Part Two of this paper it is shown that a plane electromagnetic wave, polarized with the electric vector in the plane of incidence and in the wave front, impinging on a plane solid at nearly grazing incidence produces a total field in which the horizontal electric field near the solid has very nearly the same ratio to the vertical electric field as in the Zenneck surface wave. In Part Three of this paper it is shown that the wave tilt near the ground at a great distance from a vertical dipole is almost the same as that found for the plane wave at nearly grazing incidence.

Part One—The Evidence Against the Surface Wave

The following discussion centers around the surface wave wave-function P and the series (5), (6), (8) and (9) of paper 3 in the bibliography.* These series and P follow

$$P = -\frac{\pi s\tau}{1-\tau^2} H_0^{(2)}(sr)e^{i\tau s z},\dagger \tag{12}$$

$$Q_1 + P/2 = \frac{1}{r}\frac{e^x}{1-\tau^2}\sum_{n=0}^{\infty} A_n(-x)^n, \tag{5}$$

$$Q_2 + P/2 = \frac{1}{r}\frac{1-\tau^2 e^{x_2}}{1-\tau^2}\sum_{n=0}^{\infty} B_n(-x_2)^n, \tag{6}$$

$$Q_0 = Q_1 + P \sim \frac{1}{r}\frac{e^x}{1-\tau^2}\sum_{n=1} C_n x^{-n}, \tag{8}$$

$$Q_1 \sim \frac{1}{r}\frac{1-\tau^2 e^{x_2}}{1-\tau^2}\sum_{n=1} D_n x_2^{-n}, \tag{9}$$

where r = horizontal distance, $x = -ik_1 r$, $x_2 = -ik_2 r$, $\tau = k_1/k_2$, $s = k_1/\sqrt{1+\tau^2}$, $k^2 = \epsilon\mu\omega^2 - 4\pi\sigma\mu i\omega$, $k_2^2 = k_1^2(\epsilon - i2c\lambda\sigma)$, $k_1 \approx 2\pi/\lambda$ in air, $a = \tau^2/(1+\tau^2)$, $a_2 = 1/(1+\tau^2)$,

$$\begin{aligned}
A_0 &= 1, \quad A_1 = \sqrt{a}\tanh^{-1}\sqrt{a}, \quad A_2 = A_1 - a,\\
A_n &= [(2n-3)A_{n-1} - aA_{n-2}]/(n-1)^2,\\
B_0 &= 1, \quad B_1 = \sqrt{a_2}\tanh^{-1}\sqrt{a_2}, \quad B_2 = B_1 - a_2,\\
B_n &= [(2n-3)B_{n-1} - a_2 B_{n-2}]/(n-1)^2,\\
C_1 &= -1/a, \quad C_2 = -3/a^2 + 1/a,\\
C_n &= [(2n-1)C_{n-1} - (n-1)^2 C_{n-2}]/a,\\
D_1 &= -1/a_2, \quad D_2 = -3/a_2^2 + 1/a_2,\\
D_n &= [(2n-1)D_{n-1} - (n-1)^2 D_{n-2}]/a_2.
\end{aligned}$$

* Sommerfeld's time factor $e^{-i\omega t}$ which was used in paper 3 has been replaced by $e^{i\omega t}$.

† z, the height above ground, is zero in paper 3.

The left hand side of (8) has been altered to correspond with the facts as now known.

P is the wave-function for a surface wave which at great distances behaves like Zenneck's plane surface wave.

The series (5) and (6) constitute the complete wave-function for a unit vertical dipole centered on the interface between air and ground.

The series (8) and (9) are the asymptotic expansions of (5) $+ P/2$ and (6) $- P/2$.

The series (5), (6), (8) and (9) are exact and it is from them that the attenuation charts in a paper by C. R. Burrows in this issue of the *Bell System Technical Journal* were computed.

Since interchanging k_1 and k_2 in (5) gives (6) and interchanging k_1 and k_2 in (8) gives (9) but interchanging k_1 and k_2 in P changes its sign it follows that if (6) $\sim$ (9) $+ P/2$ then (5) $\sim$ (8) $- P/2$. Hence the complete wave-function $\Pi_z =$ (5) + (6) $\sim$ [(8) $- P/2$] + [(9) $+ P/2$] = (8) + (9) and P does not appear in the asymptotic expansion of the wave-function.

The series (5) and (6) have been computed and found to be respectively equal to (8) $- P/2$ and (9) $+ P/2$.* These computations show again that $\Pi_z =$ (5) + (6) $\sim$ (8) + (9) or putting it in words, that there is no surface wave wave-function P in the asymptotic expansion of the complete wave-function.

As a further check S. O. Rice has derived the series (5) and (6) in an entirely different manner and verified that their asymptotic expansions are indeed $Q_0 - P/2$ and $Q_2 + P/2$.

In order to get a direct numerical check on the series the wave-function integral was computed by mechanical quadrature for two cases. Van der Pol's transformation of the wave-function integral with the path of integration deformed upward along the lines $Im(ihru)$ constant was used.[6]

1. With $r/\lambda = 1/4\pi$ and $\epsilon - i2c\lambda\sigma = 12.5 - i\,12.5$ mechanical quadrature gave $\Pi_z = (.800 - i\,.578)/r$ while the series (5) and (6) gave $(.9247 - i\,.4334)/r$ and $(-\,.1242 - i\,.1438)/r$ respectively which add up to $(.8005 - i\,.5772)/r$. This is a good check on the series (5) and (6).

2. With $r/\lambda = 50$ and $\epsilon - i2c\lambda\sigma = 80 - i\,.7512$ mechanical quadrature gave $\Pi_z = (.094 - i\,.178)/r$ while the series (8) and (9) gave $Q_0 \approx (.086 - i\,.187)/r$ and $Q_2 \approx 1.2 \times 10^{-11} \lceil 13^0/r$. Since $P = (4.47 - i\,1.92)/r$ there can be no doubt that it must be omitted in computing Π_z asymptotically. This is a good check on the above stated relation $\Pi_z =$ (5) + (6) $\sim$ (8) + (9) or $\Pi_z \sim Q_0 + Q_2$. Because the asymptotic series Q_0 here starts to diverge at the third term

* Eq. (1) in paper 4 says that (5) $\sim$ (8) $- P/2$.

it is not possible to determine Q_0 with an accuracy better than the discrepancy between the values given for Q_0 and Π_z.

The above cited facts prove that on the ground the wave-function for a vertical dipole centered on the interface between air and ground is

$$\Pi_z = (5) + (6) \sim [(8) - P/2] + [(9) + P/2] = (8) + (9)$$

or

$$\Pi_z = (Q_1 + P/2) + (Q_2 + P/2) \sim Q_1 + Q_2 + P = Q_0 + Q_2.$$

The function P can only be thought of as follows. The convergent series (5) and (6) comprising the wave function can not be directly expressed as inverse power series; but if the function $P/2$ is respectively added and subtracted the resulting sum and difference do have the asymptotic inverse power series expansions (8) and (9).

Part Two—Superseding the Surface Wave

It has now been shown by theory, by numerical studies and by crucial experiment that Zenneck's surface wave is not a component in the asymptotic expansion of the wave-function for a vertical dipole.

Since the wave antennas were designed to utilize the horizontal component of the Zenneck wave electric field and do pick up radio signals it is desirable that we explain the success of the wave antennas in some other way at the same time that we throw away the Zenneck wave.

The object of this part of the paper is to show that the success of the wave antennas can be well accounted for by means of a plane wave theory. It will be shown that if a plane electromagnetic wave polarized with the electric vector in the plane of incidence and in the wave front impinges on a plane solid at a large angle with the normal to the surface then near the surface the ratio of the horizontal to the vertical component of the total electric field is very nearly the same as though the total field were that of a Zenneck surface wave.

Since the electric and magnetic fields of an antenna ultimately lie in the wave front and since the wave front at any considerable distance is effectively plane for a structure the size of a wave antenna and since the radiation coming down from the ionosphere consists chiefly of that which has been subjected to the minimum number of reflections and the angle at which the radiation arrives at the receiving wave antenna is usually rather low this plane wave theory easily accounts for the success of the wave antennas.

A plane electromagnetic wave polarized with its electric vector in the plane of incidence falls upon a plane semi-conducting surface. We are interested in the total field.

Let the incident electric field be

$$E_{xi} = e^{i\omega t - ik_1(x \sin\theta - z \cos)} \cos\theta,$$
$$E_{zi} = e^{i\omega t - ik_1(\sin\theta - z\cos\theta)} \sin\theta.$$

The reflected field is

$$E_{xr} = -\, e^{i\omega t - ik_1(x\sin\theta + z\cos\theta)} \cos\theta \cdot R,$$
$$E_{zr} = e^{i\omega t - ik_1(x\sin\theta + z\cos\theta)} \sin\theta \cdot R,$$

where

$$R = \frac{\cos\theta - \tau\sqrt{1 - \tau^2 \sin^2\theta}}{\cos\theta + \tau\sqrt{1 - \tau^2\sin^2\theta}},$$
$$\tau = k_1/k_2 = 1/(\epsilon - i2c\lambda\sigma)^{1/2}.$$

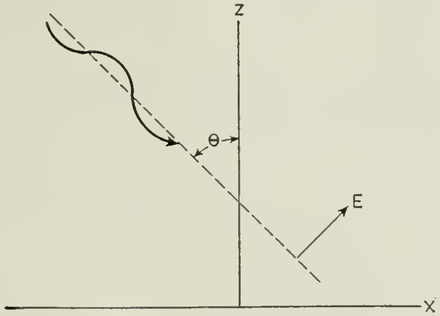


Fig. 1

Then the total field is

$$E_x = e^{i\omega t - ik_1 x \sin\theta} \cos\theta(e^{i\eta} - Re^{-i\eta}),$$
$$E_z = e^{i\omega t - ik_1 x\sin\theta} \sin\theta(e^{i\eta} + Re^{-i\eta}),$$

where $\eta = k_1 z \cos\theta$.

$$E_x = e^{i\omega t - ik_1 x \sin\theta}\, 2\cos\theta \frac{i\cos\theta\sin\eta + \tau\sqrt{1 - \tau^2\sin^2\theta}\cos\eta}{\cos\theta + \tau\sqrt{1 - \tau^2\sin^2\theta}},$$

$$E_z = e^{i\omega t - ik_1 x\sin\theta}\, 2\sin\theta \frac{\cos\theta\cos\eta + i\tau\sqrt{1 - \tau^2\sin^2\theta}\sin\eta}{\cos\theta + \tau\sqrt{1 - \tau^2\sin^2\theta}}.$$

In order to see better the significance of these formulas it is necessary to write $\theta = \pi/2 - \delta$ where δ is small, say less than 20°, and suppose that $k_1 z \cos\theta$ is small. Then we may use the expansions

$$\cos\theta = \sin\delta = \delta - \delta^3/3! + - \cdots,$$
$$\sin\theta = \cos\delta = 1 - \delta^2/2! + - \cdots,$$
$$\tan\eta = \cos\theta \cdot k_1 z(1 + k_1^2 z^2 \delta^2/3 + \cdots),$$
$$\cos\eta = 1 - k_1^2 z^2 \delta^2/2 + - \cdots.$$

If terms of third and higher order in δ are dropped we have left

$$E_x = e^{i\omega t - ik_1 x \sin\theta}\, 2\delta\left(1 - \frac{\delta^2}{6}\right)\left(1 - k_1^2 z^2 \frac{\delta^2}{2}\right)$$

$$\left(\tau\sqrt{1-\tau^2} + \frac{\tau^3\delta^2}{2\sqrt{1-\tau^2}} + ik_1 z\delta^2\right)\Big/\left(\tau\sqrt{1-\tau^2} + \delta + \frac{\tau^3\delta^2}{2\sqrt{1-\tau^2}}\right),$$

$$= e^{i\omega t - ik_1 x \sin\theta}\frac{\tau\sqrt{1-\tau^2}}{\delta + \tau\sqrt{1-\tau^2} + \tau^3\delta^2/2\sqrt{1-\tau^2}}\, 2\delta$$

$$[1 - \delta^2(\tfrac{1}{6} + k_1^2 z^2/2 - \tau^2/2(1-\tau^2) - ik_1 z/\tau\sqrt{1-\tau^2})],$$

$$E_z = e^{i\omega t - ik_1 x \sin\theta}\, 2\left(1 - \frac{\delta^2}{2}\right)\delta\left(1 - \frac{\delta^2}{6}\right)\left(1 - k_1^2 z^2\frac{\delta^2}{2}\right)$$

$$\left[1 + \left(\tau\sqrt{1-\tau^2} + \frac{\tau^3\delta^2}{2\sqrt{1-\tau^2}}\right) ik_1 z(1 + k_1^2 z^2 \delta^2/3)\right]\Big/$$

$$\left(\tau\sqrt{1-\tau^2} + \delta + \frac{\tau^3\delta^2}{2\sqrt{1-\tau^2}}\right),$$

$$= e^{i\omega t - ik_1 x \sin\theta}\frac{1 + \tau\sqrt{1-\tau^2}k_1 zi}{\delta + \tau\sqrt{1-\tau^2} + \tau^3\delta^2/2\sqrt{1-\tau^2}}\, 2\delta$$

$$\left[1 - \delta^2\left(\frac{2}{3} + k_1^2\frac{z^2}{2} - i\frac{\tau^3 k_1 z/2 + \tau(1-\tau^2)k_1^3 z^3/3}{(1 + \tau\sqrt{1-\tau^2}k_1 z)\sqrt{1-\tau^2}}\right)\right].$$

The wave tilt is then

$$\frac{E_x}{E_z} = \frac{\tau\sqrt{1-\tau^2}}{1 + \tau\sqrt{1-\tau^2}k_1 zi}\left\{1 + \delta^2\left[\frac{1}{2} + \frac{ik_1 z}{\tau\sqrt{1-\tau^2}} + \frac{\tau^2}{2(1-\tau^2)}\right.\right.$$

$$\left.\left. - i\frac{\tau^3 k_1 z/2 + \tau(1-\tau^2)k_1^3 z^3/3}{(1 + \tau\sqrt{1-\tau^2}ik_1 z)\sqrt{1-\tau^2}}\right]\right\}.$$

The wave tilt in the Zenneck wave is just τ.

As a particular and probably typical case we may take $\epsilon = 9$, $\sigma = 2 \times 10^{-14}$ and $f = 60{,}000$ and then $\tau = k_1/k_2 = 1/\sqrt{\epsilon - i2c\lambda\sigma}$ $= 1/\sqrt{9 - i600} = .04082\,\underline{|44.570°}$. If $z = 30$ ft. then

$$k_1 z = 2\pi\; 30 \times 30.48/5 \times 10^5 = .01149.$$

If $\delta = 10° = .1745$ radian then $\delta^2 = .03045$. The coefficient of τ in E_x/E_z then turns out to differ from unity by only about 1 per cent.

These figures show that if we retain only the principal terms in our formulæ we have

$$E_x = e^{i\omega t - ik_1x\sin\theta}\frac{\tau\sqrt{1-\tau^2}}{\delta+\tau\sqrt{1-\tau^2}}2\delta\left[1-\delta^2\left(\frac{1}{6}-\frac{ik_1z}{\tau}\right)\right],$$

$$E_z = e^{i\omega t - ik_1x\sin\theta}\frac{1+\tau\sqrt{1-\tau^2}ik_1z}{\delta+\tau\sqrt{1-\tau^2}}2\delta[1-\tfrac{2}{3}\delta^2],$$

$$\frac{E_x}{E_z} = \frac{\tau\sqrt{1-\tau^2}}{1+\tau\sqrt{1-\tau^2}ik_1z}[1+\delta^2(\tfrac{1}{2}+ik_1z/\tau)].$$

As a rule the wave tilt is so nearly equal to the value τ predicted by Zenneck that present day wave tilt measurements do not distinguish between the two.

Part Three—The Wave Tilt of the Q_0-Wave

It would be but natural for a reader to ask what wave tilt would be observed at the surface of a flat earth if there were no Heaviside layer. It was shown in Part One that the asymptotic expansion of the complete wave function is $Q_0 + Q_2$, of which Q_2 is negligible. The function Q_0 there considered is the surface value of a detached wave that carries energy to infinity in all directions. One would therefore expect that at the surface of the earth the Q_0-wave would act like the detached plane wave employed in Part Two. It will now be shown that it does.

It was shown in paper 8 that in the air

$$Q_0 \sim \frac{e^{-ikr}}{r}+\frac{e^{-ikr_2}}{r_2}\left\{g_{01}(c)-\frac{g_{02}(c)}{ikr_2}+\frac{g_{03}(c)}{(ikr_2)^2}+\cdots\right\},$$

where

$$g_{01}(c) = \frac{c-\tau\sqrt{1-\tau^2+\tau^2c^2}}{c+\tau\sqrt{1-\tau^2+\tau^2c^2}},$$

$$g_{0(n+1)}(c) = \frac{n-1}{2}g_{0n}(c)-\frac{c}{n}g_{0n}'(c)+\frac{1-c^2}{2n}g_{0n}''(c),$$

$c = \cos\theta$ and r_2 and θ are shown in Fig. 2,

$$r_2 = \sqrt{\rho^2+w^2}, \qquad c = w/r_2, \qquad w = z+a.$$

We need to compute (elm. units are employed, $\mu = 1$)

$$E_\rho = \frac{-\mu i\omega}{k^2}\frac{\partial^2 Q_0}{\partial\rho\partial z},$$

$$= \frac{-i\omega}{k^2}\left(\sqrt{1-c^2}\frac{\partial}{\partial r_2}-\frac{c\sqrt{1-c^2}}{r_2}\frac{\partial}{\partial c}\right)\left(c\frac{\partial}{\partial r_2}+\frac{1-c^2}{r_2}\frac{\partial}{\partial c}\right)Q_0$$

and

$$E_z = -\mu i\omega\left[1 + \frac{1}{k^2}\frac{\partial^2}{\partial z^2}\right]Q_0,$$

$$= -\ i\omega\left[1 + \frac{1}{k^2}\left(c\frac{\partial}{\partial r_2} + \frac{1 - c^2}{r_2}\frac{\partial}{\partial c}\right)^2\right]Q_0$$

at a great distance near the interface; that is to say, retaining only the leading terms in c/r_2 and $1/r_2^2$.

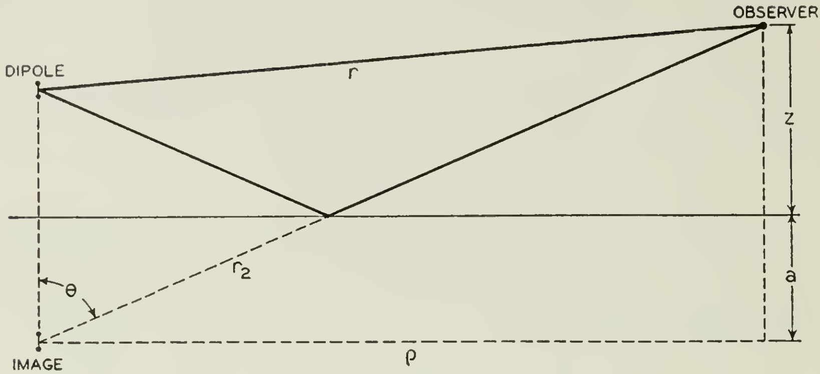


Fig. 2

The complete calculation of E_ρ and E_z is too long to be included. If $a = 0$ so that $r = r_2$

$$E_\rho = -\ i\omega\frac{e^{-ikr_2}}{r_2}\left\{-\ c\sqrt{1 - c^2}(1 + g_{01}(c)) + \frac{\sqrt{1 - c^2}}{ikr_2}[(1 - 2c^2)g_{01}'(c)\right.$$

$$- 3c(1 + g_{01}(c)) + cg_{02}(c)] - \frac{\sqrt{1 - c^2}}{(ikr_2)^2}[-\ (2 - 5c^2)g_{01}'(c)$$

$$-c(1 - c^2)g_{01}''(c) + (1 - 2c^2)g_{02}'(c) + 3c(1 + g_{01}(c))$$

$$\left. - 5cg_{02}(c) + cg_{03}(c)] + \cdots\right\},$$

$$E_z = -\ i\omega\frac{e^{-ikr_2}}{r_2}\left\{(1 - c^2)(1 + g_{01}(c)) - \frac{1}{ikr_2}[(1 - c^2)g_{02}(c)\right.$$

$$- (1 - 3c^2)(1 + g_{01}(c)) - 2c(1 - c^2)g_{01}'(c)]$$

$$+ \frac{1}{(ikr_2)^2}[-\ (1 - 5c^2)g_{02}(c) + (1 - c^2)g_{03}(c) - (1 - c^2)^2g_{01}''(c)$$

$$\left. + (1 - 3c^2)(1 + g_{01}(c)) + c(1 - c^2)(5g_{01}'(c) - 2g_{02}'(c))] + \cdots\right\}.$$

Since c is to be very small it is best to expand $g_{01}(c)$ into an ascending power series in c.

$$g_{01}(c) = -1 + \frac{2c}{\tau\sqrt{1-\tau^2}} - \frac{2c^2}{\tau^2(1-\tau^2)} + \frac{(2-\tau^4)c^3}{\tau^3(1-\tau^2)^{3/2}}$$
$$- \frac{2(1+\tau^2)c^4}{\tau^4(1-\tau^2)} + \frac{(8-12\tau^4+3\tau^8)c^5}{4\tau^5(1-\tau^2)^{5/2}} - \frac{2(1+\tau^2)^2c^6}{\tau^6(1-\tau^2)} + - \cdots.$$

The recurrence relation then gives us

$$g_{02}(c) = \frac{-2}{\tau^2(1-\tau^2)} + \frac{(6-2\tau^2-\tau^4)c}{\tau^3(1-\tau^2)^{3/2}} - \frac{(12+6\tau^2)c^2}{\tau^4(1-\tau^2)}$$
$$+ \frac{(40-24\tau^2-36\tau^4+12\tau^6+3\tau^8)c^3}{2\tau^5(1-\tau^2)^{5/2}} + \cdots,$$

$$g_{03}(c) = \frac{-6-4\tau^2}{\tau^4(1-\tau^2)} + \frac{(120-72\tau^2-108\tau^4+36\tau^6+9\tau^8)c}{4\tau^5(1-\tau^2)^{5/2}} + \cdots.$$

After dropping all but the leading terms there is left

$$E_p = -i\omega\frac{e^{-ikr_2}}{r_2}\left\{\frac{1}{ik_2r_2}\left[\frac{2}{\tau\sqrt{1-\tau^2}}\right]\right\},$$
$$E_z = -i\omega\frac{e^{-ikr_2}}{r_2}\left\{\frac{2(1+\tau\sqrt{1-\tau^2}ikw)}{ikr_2\tau^2(1-\tau^2)}\right\}.$$

The wave tilt near the surface of the ground is then

$$\frac{E_p}{E_z} = \frac{\tau\sqrt{1-\tau^2}}{1+\tau\sqrt{1-\tau^2}ikz}.$$

This is the wave tilt in the asymptotic field of a quarter wave antenna or flat top antenna.

If a is not zero but c is small the final field expressions are

$$E_p = -i\omega\frac{e^{-ikr_2}}{r_2}\left\{\frac{2+\tau\sqrt{1-\tau^2}ik[w+(2a-w)e^{ik2az/r_2}]}{ikr_2\tau\sqrt{1-\tau^2}}\right.$$
$$-\frac{3}{ikr_2^2}[(w-2a)e^{ik2az/r_2}-w]$$
$$\left.-\frac{6-6\tau^2-3\tau^4+6\tau\sqrt{1-\tau^2}ikw+2\tau^2(1-\tau^2)(ikw)^2}{(ikr_2)^2\tau^3(1-\tau^2)^{3/2}}+\cdots\right\},$$

$$E_z = -i\omega \frac{e^{-ikr_2}}{r_2}\left\{ (e^{ik2az/r_2} - 1)\left(1 + \frac{1}{ikr_2}\right) + \frac{2(1 + \tau\sqrt{1 - \tau^2}ikw)}{ikr_2\tau^2(1 - \tau^2)} \right.$$
$$+ \frac{(e^{ik2az/r_2} - 1)(1 + k^2w^2) - 6k^2a(w - a)e^{ik2az/r_2}}{(ikr_2)^2}$$
$$\left. + \frac{2 - 6/\tau^2 - (6 - 8\tau^2 + 5\tau^4)ikw/\tau\sqrt{1 - \tau^2} - 2(ikw)^2}{(ikr_2)^2\tau^2(1 - \tau^2)} \right\}.$$

If $k2az/r_2 \ll 1$ the leading terms give

$$E_\rho = -i\omega \frac{e^{-ikr_2}}{r_2} \cdot \frac{2(1 + \tau\sqrt{1 - \tau^2}ika)}{ikr_2\tau\sqrt{1 - \tau^2}},$$

$$E_z = -i\omega \frac{e^{-ikr_2}}{r_2} \cdot \frac{2(1 + \tau\sqrt{1 - \tau^2}ikz)(1 + \tau\sqrt{1 - \tau^2}ika)}{ikr_2\tau^2(1 - \tau^2)}$$

and E_ρ/E_z is the same as obtained above with $a = 0$.

Bibliography

1. J. Zenneck: *Ann. der Physik*, Bd. 23, page 846, 1907.
2. A. Sommerfeld: *Ann. der Physik*, Bd. 28, page 705, 1909. *Jahrb. d. drahtl. Tel.*, Bd. 4, page 157, 1911.
3. W. H. Wise: *Proc. I. R. E.*, Vol. 19, page 1684, 1931.
4. W. H. Wise: *Bulletin Amer. Math. Soc.*, Vol. XLI, pp. 700–706, October, 1935.
5. Bruno Rolf: *Ingeniörs Vetenskaps Akademien, Handlingar*, Nr 96, 1929.
6. Balth. van der Pol: *Zeitschrift für Hochfrequenztechnik*, Bd. 37, page 152, 1931.
7. H. Weyl: *Ann. der Physik*, Bd. 60, page 481, 1919.
8. W. H. Wise: *Physics*, Vol. 4, page 354, 1933.
9. K. A. Norton: *Proc. I. R. E.*, Vol. 24, page 1367, 1936.
10. C. R. Burrows, *Nature*, Aug. 15, 1936. A more complete description will appear in the February issue of the *Proc. I. R. E.*

Radio Propagation Over Plane Earth—Field Strength Curves

By CHARLES R. BURROWS

Curves are presented to facilitate the calculation of radio propagation over plane earth. The magnitude and phase of the reflection coefficient for all conductivities of interest and for four values of the dielectric constant are presented in the form of curves from which the significant quantities may be read with the same degree of accuracy for all conditions. Simple equations, from which the effect of raising the antennas above the earth's surface may be readily calculated, are presented.

Introduction

THIS paper is intended to facilitate the calculation of radio propagation over plane earth. In Part I curves are presented that show the decrease of field strength with distance for antennas on the surface of the earth. In these curves the results obtained by Sommerfeld [1] and Rolf [2] are corrected and certain approximations [3] introduced by Rolf to reduce the number of variables to a workable number are eliminated. For a discussion of the Sommerfeld-Rolf curves, the reader is referred to a companion paper.[4]

Part II is concerned with the more general case of antennas above the surface of the earth. The complete equation that gives the field strength for antennas at any height above the earth is reduced to a simple equation which allows the calculation of the field under conditions of practical interest.

To facilitate field strength calculations, the values of the reflection coefficient are presented in the form of curves from which the significant quantities may be read with the required degree of accuracy for all angles of incidence.

Preliminary Remarks

A rectilinear antenna in free space generates an electric field whose effective value in the equatorial plane of the antenna at a distance large compared with the wave-length and the antenna length is

$$E_0 = \frac{60\pi HI}{\lambda d}, \tag{1}$$

where HI is equal to the line integral of the current taken over the

[1] Numbers refer to bibliography at end.

antenna.* If the antenna is placed above and perpendicular to a perfectly conducting plane and the antenna current is maintained the same, the electric field will be twice as great † or

$$E = 2E_0 = \frac{120\pi HI}{\lambda d}. \tag{2}$$

To maintain the current constant, however, it is now necessary to deliver more power to the antenna.

For a short doublet antenna in free space the radiation resistance is $R_0 = 80\pi^2 H^2/\lambda^2$ and hence the effective value of the received field strength is given as a function of the radiated power by ‡

$$E = \frac{3\sqrt{5}\sqrt{P}}{d}. \tag{3}$$

If this antenna is placed perpendicular to and very near a perfectly conducting plane the field strength pattern will be unchanged in the upper hemisphere but there will be no field below the perfectly conducting plane. The power that was required to produce the field in the lower hemisphere, which because of symmetry is half the total, is no longer radiated so that the same field strength will be produced by half the power,†† or

$$E = \frac{3\sqrt{10}\sqrt{P}}{d}. \tag{4}$$

If the transmitting antenna is removed so far from the ground that the reaction of the currents in the ground on the antenna current is negligible its radiation resistance is the same as if the ground were not present. The receiving antenna, however, still "sees" the image of the transmitting antenna in the ground. At a distance large compared with the height above ground, the transmitting antenna and

* The units are volts, amperes, meters and watts. H is the effective height of the antenna as defined in the most recent "Report of the Standards Committee" of the I.R.E. (1933).

† Under the hypothetical conditions taken by Sommerfeld, namely the antenna half in the ground and half in the air, the field is the same above a perfectly conducting plane as in free space. When the antenna is entirely above a perfectly conducting plane the field is the same as it would be if the plane were replaced by the image of the antenna in it. That is, the field is the sum of two equal components, one due to the antenna itself and the other due to its image. At distances large compared with the height of the antenna above the plane these two components are in phase and their sum is equal to twice either of them.

‡ For half-wave antennas the numerical factors in equations (3), (4) and (5) are respectively 7.0, 9.9 and 14.0.

†† Let E_1 be the received field strength in free space produced by a power P_1 and let E_2 be the field strength for an antenna perpendicular to and very near a perfectly conducting plane produced by a power P_2. Then $E_2 = E_1$ when $P_2 = P_1/2$, and by equation (3), $E_2 = E_1 = 3\sqrt{5}\sqrt{P_1}/d = 3\sqrt{5}\sqrt{2P_2}/d$, which is equivalent to equation (4).

its image are substantially the same distance from the receiver so that

$$E = \frac{6\sqrt{5}\sqrt{P}}{d}. \tag{5}$$

The way in which the ground currents affect the antenna resistance is given by the following equations which follow directly from more general cases considered by Sterba.[5]

$$\frac{R_V}{R_0} = \left[1 - 3\left(\frac{\cos v}{v^2} - \frac{\sin v}{v^3}\right)\right] = 2 - \frac{v^2}{10} + \frac{v^4}{280} - \cdots, \tag{6}$$

$$\frac{R_H}{R_0} = \left[1 - \frac{3}{2}\left(\frac{\sin v}{v} + \frac{\cos v}{v^2} - \frac{\sin v}{v^3}\right)\right] = \frac{v^2}{5} - \frac{3v^4}{280} + \cdots, \tag{7}$$

where v is equal to 4π times the height of the antenna above the ground in wave-lengths, and R_V and R_H are the radiation resistances of

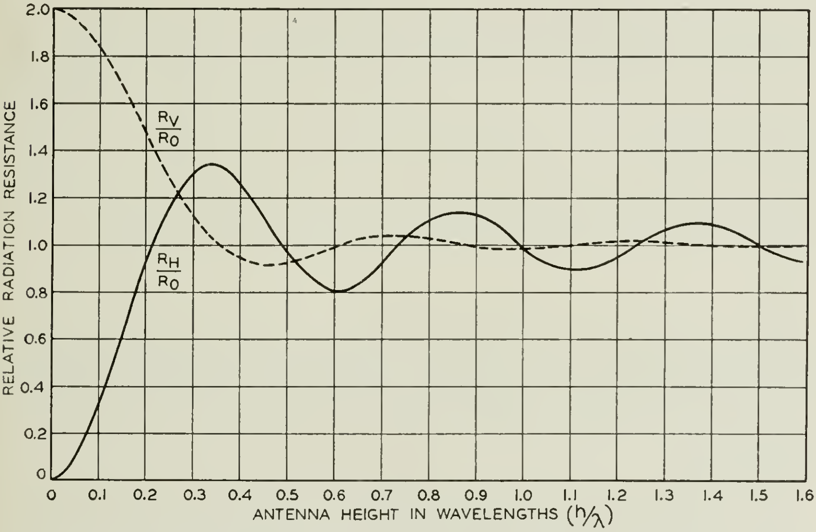


Fig. 1—Ratio of the radiation resistance of a short doublet antenna above perfectly conducting ground to that of the same antenna in free space.

short vertical and horizontal doublets above perfectly conducting earth respectively, and R_0 is the radiation resistance of the same antenna in free space. For the same input power the received field is inversely proportional to the square root of these ratios which are plotted in Fig. 1.

It is sometimes convenient to express the results in terms of the ratio of transmitted power to useful received power. The useful received power is the maximum power that can be transferred from the receiving antenna to the first circuit of the receiver. This is

$$P_r = \left(\frac{E\lambda}{8\pi\sqrt{5}}\right)^2 \tag{8}$$

for a short doublet. From equation (3) it follows that the ratio of transmitted to useful received powers for antennas in free space is *

$$\frac{P_t}{P_r} = \left(\frac{8\pi d}{3\lambda}\right)^2. \tag{9}$$

For short vertical doublets above the surface of a perfectly conducting plane this becomes,

$$\frac{P_t}{P_r} = \left(\frac{4\pi d}{3\lambda}\right)^2 \left(\frac{R_{V_1}}{R_0}\right)\left(\frac{R_{V_2}}{R_0}\right) \tag{10}$$

at distances that are large compared with the antenna heights. Here R_{V_1}/R_0 and R_{V_2}/R_0 are the ratios given by equation (6) and Fig. 1 for the transmitting and the receiving antenna respectively. When the antennas are more than a wave-length above the ground these ratios are substantially unity, and only one-fourth as much transmitted power is required as would be if the antennas were in free space. When both antennas are very near the surface of the earth, $R_V/R_0 = 2$ and the same transmitted power is required as in free space.

Part I—Vertical Antennas on the Surface of the Earth

In this section transmission between two short vertical antennas above and very near to the surface of plane earth will be considered. The attenuation factor will be taken as the ratio of the received field strength to that which would result if this plane surface had perfect conductivity.

In evaluating the electromagnetic field generated by a short vertical antenna on the surface of an imperfectly conducting plane it is convenient to first determine the auxiliary function $\mathbf{\Pi}$, called the Hertzian potential, from which the vertical component of the electric field may be obtained by means of the relationship,†

$$\mathbf{E} = -\frac{240i\pi^2}{\lambda}\left(1 + \frac{\lambda^2}{4\pi^2}\frac{\partial^2}{\partial z^2}\right)\mathbf{\Pi} \quad \text{volts per meter.} \tag{11}$$

* For half-wave antennas the right-hand side of equation (9) must be multiplied by $(73.2/80)^2 = 0.837$.

† Bold face type is used to indicate a complex quantity. The same character in light face type represents its magnitude with which the radio engineer is concerned. The imaginary unit, $\sqrt{-1}$, is represented by i.

For an antenna on the surface of a perfectly conducting plane this function may be written *

$$\boldsymbol{\Pi} = 2\boldsymbol{\Pi}_0 = 2\frac{HIe^{-2\pi iR_1/\lambda}}{4\pi R_1} \text{ amperes,} \tag{12}$$

where $R_1 = \sqrt{d^2 + z^2}$ is the distance. For an antenna on an imperfectly conducting plane

$$\boldsymbol{\Pi} = 2W\boldsymbol{\Pi}_0, \tag{13}$$

where W is the ratio of the Hertzian potential due to an antenna on an imperfectly conducting plane to that on a perfectly conducting plane. W may be expressed as the sum of two infinite convergent series, A and D, which are defined in Appendix I (page 70).

$$W = A + D. \tag{14}$$

The series D becomes unwieldy for distances greater than the order of a wave-length. In order to facilitate computation, D may be transformed into an asymptotic expansion to which it is equivalent at large distances, so that

$$W = A - B/2 + F. \tag{15}$$

At still greater distances A also becomes unwieldy and it may in turn be replaced by its asymptotic expansion, which contains the term $B/2$, so that

$$W = C + F. \tag{16}$$

When the impedance of the ground is very different from that of the air,† F is small compared with $A - B/2 \approx C$. If the conductivity of the ground is not zero, F is exponentially attenuated so that it may be neglected in comparison with C in equation (16). Even if the conductivity is zero and the relative dielectric constant is as small as 4, the only effect of F in equation (16) is to produce oscillations in W of approximately 3 per cent from the magnitude of C. Even under these extreme conditions the received field strength may be calculated from equation (15) neglecting F without introducing an error greater than 3 per cent. As the transmitting antenna is approached, the approximations involved in equation (15) become poorer and poorer but at the same time the field strength becomes independent of W so that at no distance is there an appreciable error introduced

* The factor 2 occurs in equation (12) since Π_0 and E_0 refer to transmission in free space.

† This is true when the so-called "complex dielectric constant," $\epsilon - 2i\sigma/f$, differs sufficiently from unity.

by using equation (15) to calculate the field strength. This is fortunate since series D requires laborious calculations.

The attenuation factor may be obtained from $\mathbf{W}$ by means of the relation,*

$$\frac{E}{2E_0} = \frac{W}{1+\tau^2} + \frac{1}{1-\tau^4}\left[\frac{1}{2\pi id/\lambda} + \frac{1}{(2\pi id/\lambda)^2}\right], \tag{17}$$

where

$$\frac{1}{\tau^2} = \epsilon - 2i\sigma/f = \epsilon(1 - i/Q). \tag{18}$$

In this equation $2E_0$ is the inverse distance, or radiation, component of the field that would result from transmission over a perfectly conducting plane, and Q is the ratio of the imaginary component to the real component of the admittance of the ground. In other words, Q is the ratio of the dielectric current to the conduction current.† The parameter ϵ occurring in equation (18) is the relative dielectric constant (with respect to vacuum), a pure numeric that is numerically equal to the dielectric constant measured in electrostatic units.

If the value of W from equation (16) is substituted in equation (17) and terms which involve $(1/d)$ to powers higher than the first are neglected as may be done at the greater distances, we have

$$E \to \left[\frac{1}{1-\tau^4}\frac{1+\tau^2}{2\pi\tau^2 id/\lambda}\right]\left[1 - \frac{\tau^5}{(1+\tau^2)}e^{\frac{2\pi id}{\lambda}\left(1-\frac{1}{\tau}\right)}\right]2E_0. \tag{19}$$

The magnitude of the second factor on the right differs from unity

* This expression may be obtained as follows. $\mathbf{\Pi}$ satisfies the wave equation which in cylindrical coordinates (z, θ, d) is

$$\left(\frac{1}{d}\frac{\partial}{\partial d}d\frac{\partial}{\partial d} + \frac{1}{d^2}\frac{\partial^2}{\partial\theta^2} + \frac{\partial^2}{\partial z^2} + \frac{4\pi^2}{\lambda^2}\right)\mathbf{\Pi} = 0.$$

Because of symmetry the second term is zero. Solving for the value of the last two terms and substituting it in equation (11) yields

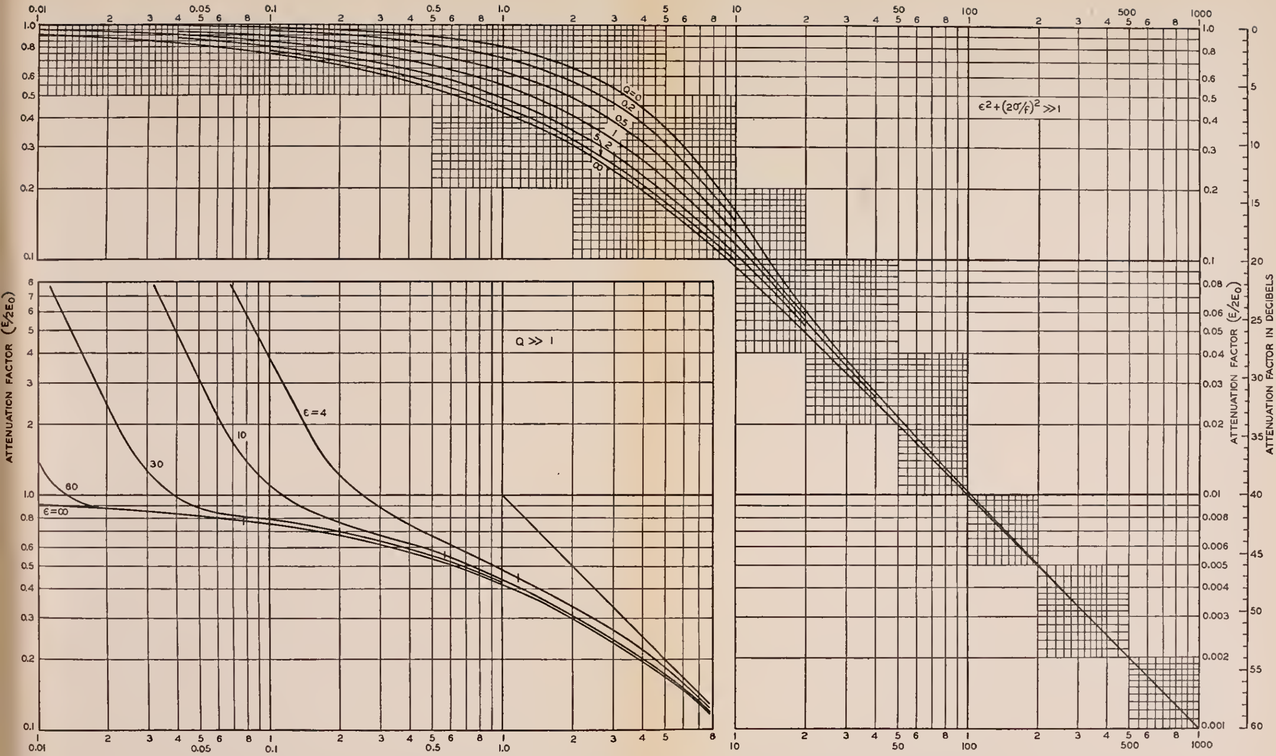
$$E = 60i\lambda\left[\frac{\partial^2}{\partial d^2} + \frac{1}{d}\frac{\partial}{\partial d}\right]\mathbf{\Pi}.$$

The differential equation given by Wise [11] for $\mathbf{\Pi}$ becomes

$$-\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2\mathbf{\Pi}}{\partial d^2} + \frac{1}{d}\frac{\partial\mathbf{\Pi}}{\partial d}\right) = \frac{\mathbf{\Pi}}{1+\tau^2} + \frac{2}{1-\tau^4}\left[\frac{1}{2\pi id/\lambda} + \frac{1}{(2\pi id/\lambda)^2}\right]\mathbf{\Pi}_0$$

when the value of $y = (1+\tau^2)\mathbf{\Pi}/2$ is substituted in his equation (7), and the result multiplied by $2/(1+\tau^2)$. Substitution of this relation in the preceding equation and division by $E_0 = -240i\pi^2\mathbf{\Pi}_0/\lambda$ gives equation (17) of the text. Since E_0 is the inverse distance component of the free space field, this relation follows from equation (11).

† In practical units $Q = 2\pi f\epsilon'/g$, where ϵ' is the dielectric constant in farads per meter and g is the conductivity in mhos per meter. On frequent occasions, the constants of the dielectric are expressed in electrostatic units; then $Q = f\epsilon/2\sigma$.



RELATIVE DISTANCE $$X = \frac{2\pi d/\lambda}{\epsilon}\,\frac{\sqrt{(1-1/\epsilon)^2+1/Q^2}}{1+1/Q^2} = \frac{2\pi d/\lambda}{\sqrt{\epsilon^2+(2\sigma/f)^2}}\sqrt{\frac{(\epsilon-1)^2+(2\sigma/f)^2}{\epsilon^2+(2\sigma/f)^2}} = \frac{2\pi d/\lambda}{\epsilon/Q}\,\frac{\sqrt{1+(1-1/\epsilon)^2 Q^2}}{1+Q^2}$$

Fig. 3. Fig. 2.

Fig. 2—Attenuation factor for radio propagation over plane earth. The number on each curve gives the value of the Q ($= \epsilon f/2\sigma$) to which it applies; σ is the conductivity and ϵ the dielectric constant in electrostatic units; f is the frequency in cycles per second; d/λ is the ratio of the distance to the wave-length.

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.

by using equation (15) to calculate the field strength. This is fortunate since series D requires laborious calculations.

The attenuation factor may be obtained from $\mathbf{W}$ by means of the relation,*

$$\frac{E}{2E_0} = \frac{W}{1+\tau^2} + \frac{1}{1-\tau^4}\left[\frac{1}{2\pi id/\lambda} + \frac{1}{(2\pi id/\lambda)^2}\right], \tag{17}$$

where

$$\frac{1}{\tau^2} = \epsilon - 2i\sigma/f = \epsilon(1 - i/Q). \tag{18}$$

In this equation $2E_0$ is the inverse distance, or radiation, component of the field that would result from transmission over a perfectly conducting plane, and Q is the ratio of the imaginary component to the real component of the admittance of the ground. In other words, Q is the ratio of the dielectric current to the conduction current.† The parameter ϵ occurring in equation (18) is the relative dielectric constant (with respect to vacuum), a pure numeric that is numerically equal to the dielectric constant measured in electrostatic units.

If the value of W from equation (16) is substituted in equation (17) and terms which involve $(1/d)$ to powers higher than the first are neglected as may be done at the greater distances, we have

$$E \rightarrow \left[\frac{1}{1-\tau^4}\frac{1+\tau^2}{2\pi\tau^2 id/\lambda}\right]\left[1 - \frac{\tau^5}{(1+\tau^2)}e^{\frac{2\pi id}{\lambda}\left(1-\frac{1}{\tau}\right)}\right]2E_0. \tag{19}$$

The magnitude of the second factor on the right differs from unity

* This expression may be obtained as follows. $\mathbf{\Pi}$ satisfies the wave equation which in cylindrical coordinates (z, θ, d) is

$$\left(\frac{1}{d}\frac{\partial}{\partial d}d\frac{\partial}{\partial d} + \frac{1}{d^2}\frac{\partial^2}{\partial\theta^2} + \frac{\partial^2}{\partial z^2} + \frac{4\pi^2}{\lambda^2}\right)\mathbf{\Pi} = 0.$$

Because of symmetry the second term is zero. Solving for the value of the last two terms and substituting it in equation (11) yields

$$E = 60i\lambda\left[\frac{\partial^2}{\partial d^2} + \frac{1}{d}\frac{\partial}{\partial d}\right]\mathbf{\Pi}.$$

The differential equation given by Wise [11] for $\mathbf{\Pi}$ becomes

$$-\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2\mathbf{\Pi}}{\partial d^2} + \frac{1}{d}\frac{\partial\mathbf{\Pi}}{\partial d}\right) = \frac{\mathbf{\Pi}}{1+\tau^2} + \frac{2}{1-\tau^4}\left[\frac{1}{2\pi id/\lambda} + \frac{1}{(2\pi id/\lambda)^2}\right]\mathbf{\Pi}_0$$

when the value of $y = (1+\tau^2)\mathbf{\Pi}/2$ is substituted in his equation (7), and the result multiplied by $2/(1+\tau^2)$. Substitution of this relation in the preceding equation and division by $E_0 = -240i\pi^2\mathbf{\Pi}_0/\lambda$ gives equation (17) of the text. Since E_0 is the inverse distance component of the free space field, this relation follows from equation (11).

† In practical units $Q = 2\pi f\epsilon'/g$, where ϵ' is the dielectric constant in farads per meter and g is the conductivity in mhos per meter. On frequent occasions, the constants of the dielectric are expressed in electrostatic units; then $Q = f\epsilon/2\sigma$.

by less than $2\frac{1}{2}$ per cent for values of ϵ greater than 4. Accordingly the magnitude of the first factor (which is equal to the first term of $\boldsymbol{C}$) gives the attenuation factor at great distances with a degree of accuracy sufficient for all practical purposes. If we choose our unit of distance such that

$$x = |2\pi\tau^2(1 - \tau^2)d/\lambda|$$

$$= \frac{2\pi d/\lambda}{\epsilon/Q} \frac{\sqrt{1 + (1 - 1/\epsilon)^2 Q^2}}{1 + Q^2} = \frac{2\pi d/\lambda}{\epsilon} \frac{\sqrt{(1 - 1/\epsilon)^2 + 1/Q^2}}{1 + 1/Q^2} \tag{20}$$

all of the attenuation curves will tend to coincide at the greater distances. This is done in Figs. 2 and 3. Figure 2 shows the variation of received field strength with distance for seven values of Q for the case where the impedance of the ground is very different from that of the air.* These curves give the correct attenuation factor for arbitrary ground constants at the greater distances. At any distance the above assumption introduces a significant error only when Q is large. Accordingly the curves of Fig. 3 have been calculated for various values of the relative dielectric constant when Q is large.† The short vertical line on each curve indicates the abscissa corresponding to a distance of one wave-length. The curves do not depart appreciably from that for an infinite dielectric constant except for distances less than this.‡ Since the error introduced in applying the curves of Fig. 2 to the general case is greatest for the conditions represented in Fig. 3, the curves of Fig. 2 may be used with confidence.

It should be emphasized that the curves of Fig. 2 give the ratio of the received field strength to that which would result from the same current in the same antenna on the surface of a plane earth of perfect conductivity. The antenna is assumed at the earth's surface so that the curves are strictly true only for short antennas. The error for half-wave doublets whose mid-points are not more than a half wave-length above the surface of the earth is negligible except in the immediate vicinity of the transmitter. The effect of height above the surface of the earth is taken up more fully in the next section.

* Since the writing of this paper, Part I of a paper by K. A. Norton on "The Propagation of Radio Waves Over the Surface of the Earth and in the Upper Atmosphere" has appeared in *Proc. I.R.E.*, **24**, 1367–1387, October, 1936. The curves of Fig. 2 in this paper are similar to those of Norton's Fig. 1, but by presenting the curves as a function of x their validity is extended to include a wider range of ground constants.

† The writer is indebted to Miss Clara L. Froelich for making these calculations.

‡ The ratio $E/2E_0$ is greater than unity at the shorter distances because E_0 is the inverse distance or radiation component of the free space field while E is the total field. At distances that are small compared with a wave-length, $E/2E_0$ is given by the second and third terms on the right of equation (17) and the effect of the ground is to increase the field by the factor $2/(1 - \tau^4)$.

The calculation of the field strength as a function of the radiated power requires a knowledge of the effect of imperfect conductivity on the resistance of the antenna. The reader is referred to papers by Barrow [6] and Niessen [7] on this subject. In the wave-length range where these curves are of greatest applicability, the practice is to minimize the ground losses by a ground system consisting of a counterpoise or a network of buried wires. When this is done the ground losses are properly part of the antenna losses and the radiated power may rightfully be taken as the rate of flow of energy past a hemisphere large enough to include the antenna and ground system. If this is done, the field strength is given by

$$E = \frac{3\sqrt{10}\sqrt{P}}{d} F(x), \tag{20a}$$

where $F(x)$ is the ratio plotted in Fig. 2.*

Part II—Antennas Above the Surface of the Earth

It is well known [8, 9] that calculations based on the physical optics of plane waves give the first approximation to the received field for radio propagation over plane earth. This approximation is accurate enough for all practical purposes if the antennas are sufficiently removed from the surface of the earth.† Under these conditions, the ratio of the received field strength to that which would be received in free space is given by ‡

$$E/E_0 = \sqrt{(1 - K)^2 + 4K \sin^2 (\gamma/2)}, \tag{21}$$

* In estimating the fraction of the total power input that is radiated the following papers may be helpful: George H. Brown, "The phase and magnitude of earth currents near radio transmitting antennas," *Proc. I.R.E.*, **23**, 168–182, February, 1935; and H. E. Gihring and G. H. Brown, "General considerations of tower antennas for broadcast use," *Proc. I.R.E.*, **23**, 311–356, April, 1935.

† This height depends upon the distance, wave-length and ground constants. The range of validity of this approximation is discussed more fully in connection with equation (27).

‡ Equation (21) gives the received field strength for either polarization for transmission along the ground. In this case the direct and reflected components are oriented in the same direction in space. It may also be used to calculate the effect of the ground for signals arriving at large angles by taking into consideration the space orientation of the components.

For horizontal antennas the orientation of the electric vector is horizontal for all angles of incidence so that equation (21) applies directly. For vertical antennas the electric vector makes the angle ξ_2 with the vertical, both in the direct and reflected wave. Hence if the ratio given in equation (21) is taken as the ratio of the vertical component of the received field to the total incident field it must be multiplied by $\cos \xi_2$. Even if the ground were not present, however, the vertical component would be reduced by this factor so that the effect of the presence of the ground on the field received by a vertical antenna is given by equation (21) as written without the $\cos \xi_2$ factor.

where K is the ratio of the amplitude of the reflected wave to that of the direct wave and $\gamma + \pi$ is their phase difference.

$$\gamma = \psi - \Delta, \tag{22}$$

where Δ is 2π times the path difference in wave-lengths and

$$\varphi = \psi \pm \pi \tag{23}$$

is the phase advance at reflection. The geometry is shown in Fig. 4. Δ may be calculated from the geometry by means of equation (47) of Appendix II (page 72).

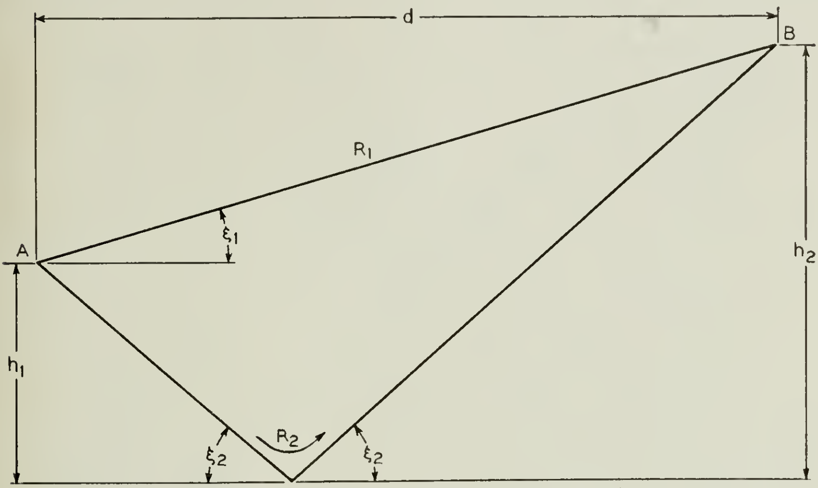


Fig. 4

The magnitude and phase of the reflection coefficient, $R = -Ke^{i\psi}$, for both polarizations are plotted in Figs. 5–12 for $\epsilon = 4$, 10, 30 and 80 and a series of values of $\epsilon/Q = 2\sigma/f$ differing by factors that are multiples of 10. The coordinate system has been chosen so that the quantities $1 - K$ and ψ that enter into the equation for the resultant field strength may be read with the same degree of accuracy for the entire range of the curves. To obtain values of $1 - K$ and ψ for smaller values of ξ_2 than shown on Figs. 5–12 use is made of the fact that both of these quantities are proportional to ξ_2 for small values of ξ_2. This linear relationship holds for the lowest cycle * of the curves so that the parts

* An exception to this occurs for large Q in the ψ-curves for vertical polarization. Under these conditions the difference between ψ and zero for values not shown on the chart is relatively unimportant. When Q is large and K is different from zero ψ is substantially 0° or 180°. For horizontal polarization ψ may be taken equal to zero for most practical purposes.

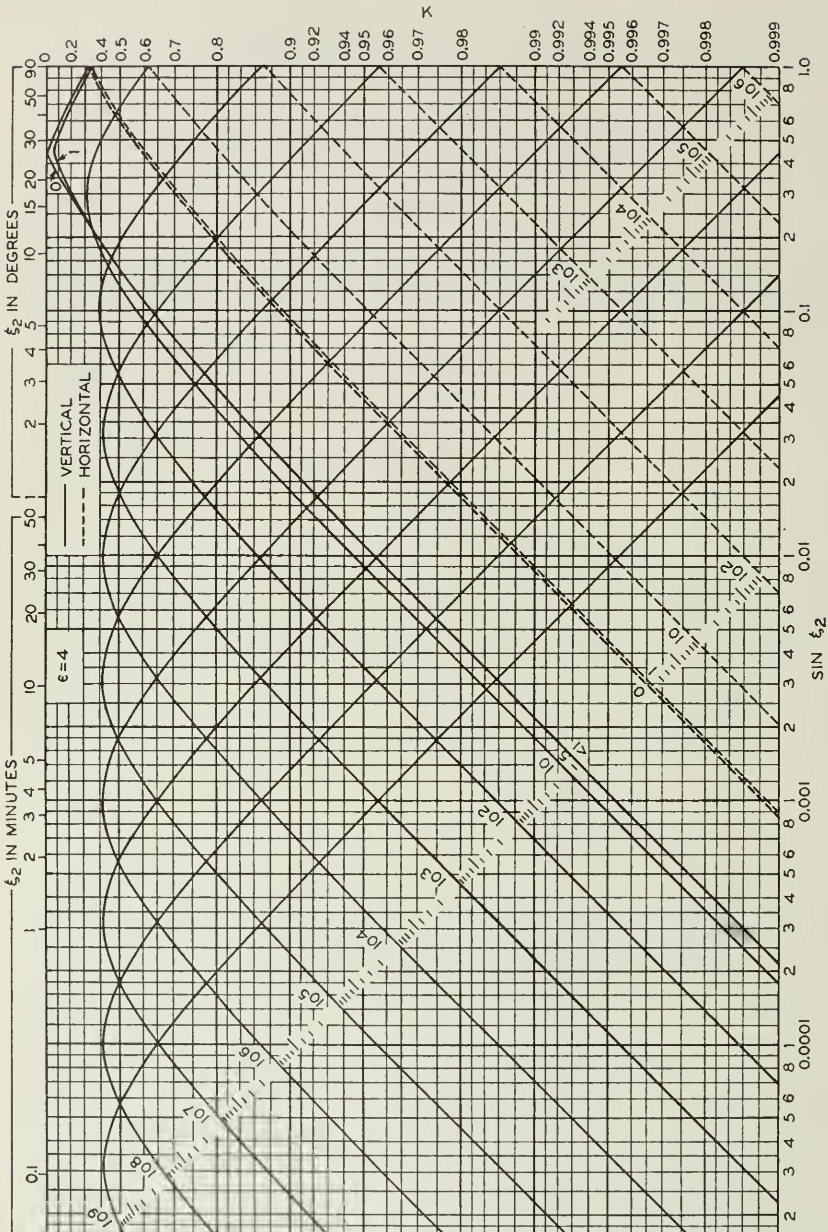


Fig. 5—Magnitude of reflection coefficient for $\epsilon = 4$. The number on each curve gives the value of q ($= 2\sigma/f$) to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

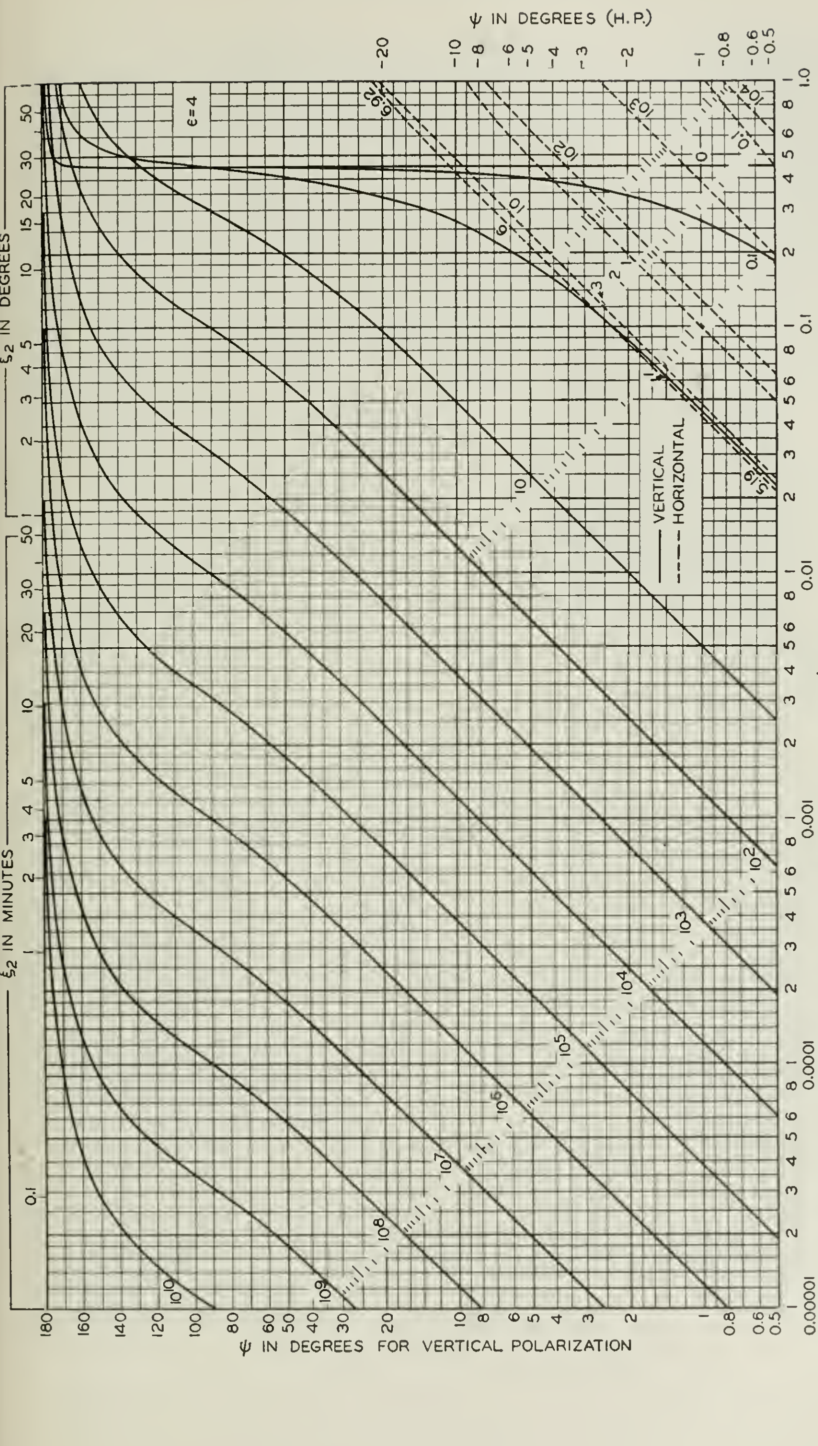


Fig. 6—Phase shift at reflection for $\epsilon = 4$. The number on each curve gives the value of q $(= 2\sigma/f)$ to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

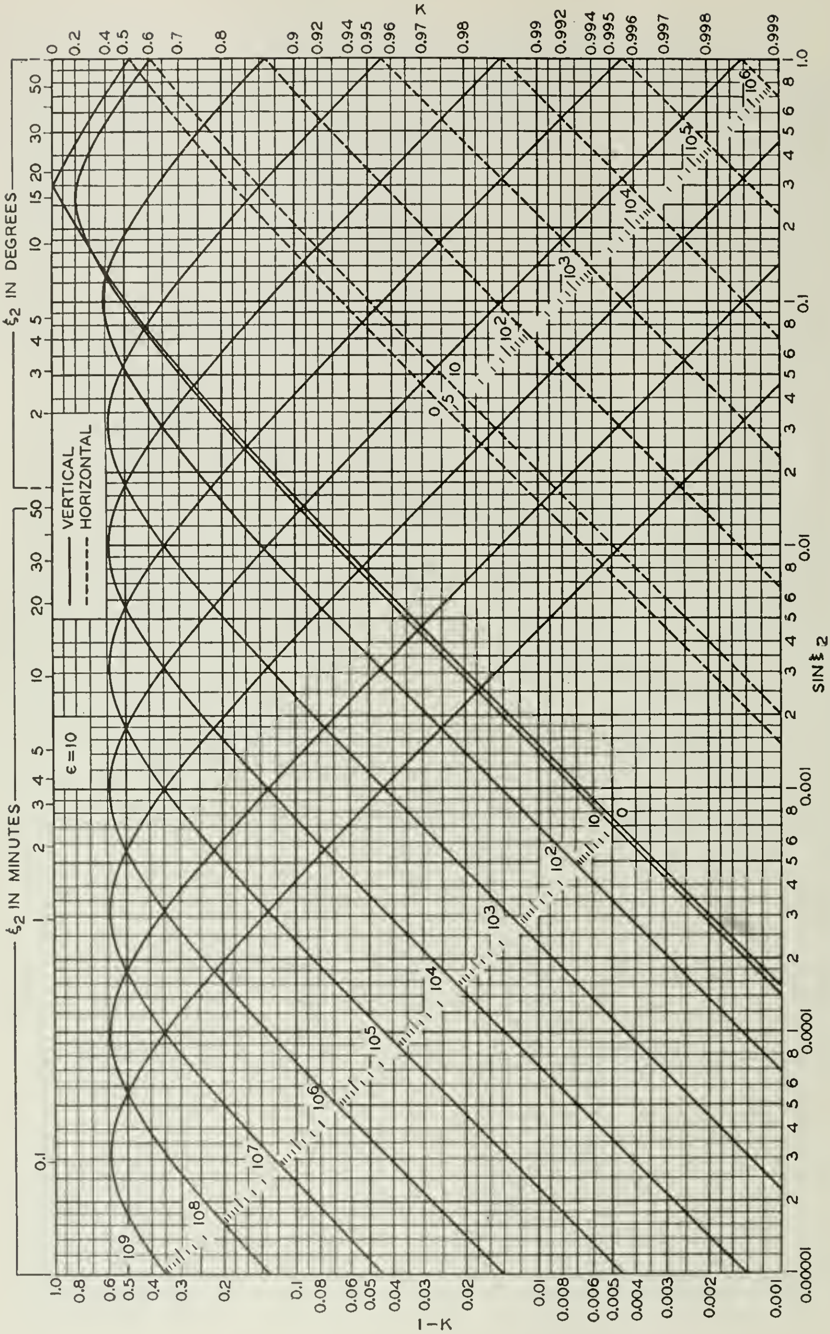


Fig. 7—Magnitude of reflection coefficient for $\epsilon = 10$. The number on each curve gives the value of q $(= 2\sigma/f)$ to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

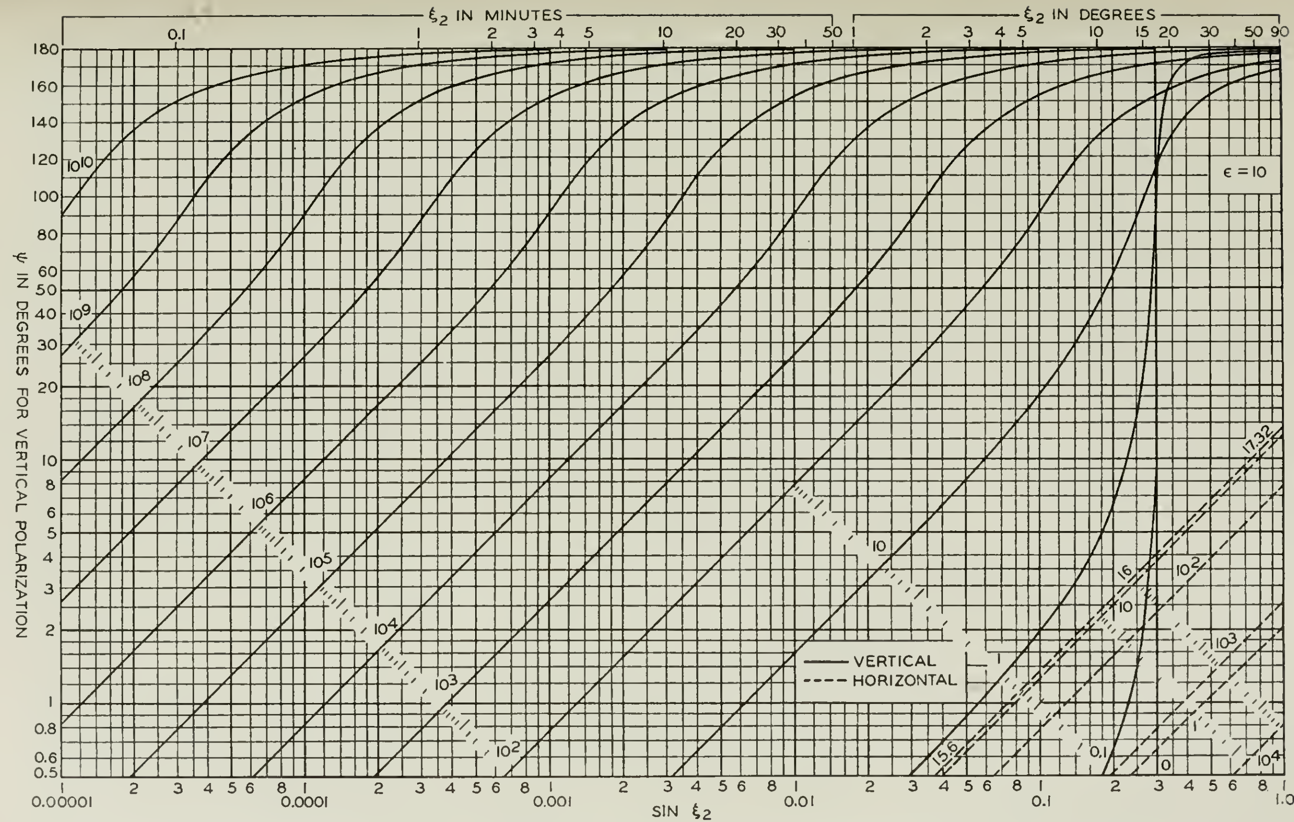


Fig. 8—Phase shift at reflection for $\epsilon = 10$. The number on each curve gives the value of q $(= 2\sigma/f)$ to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

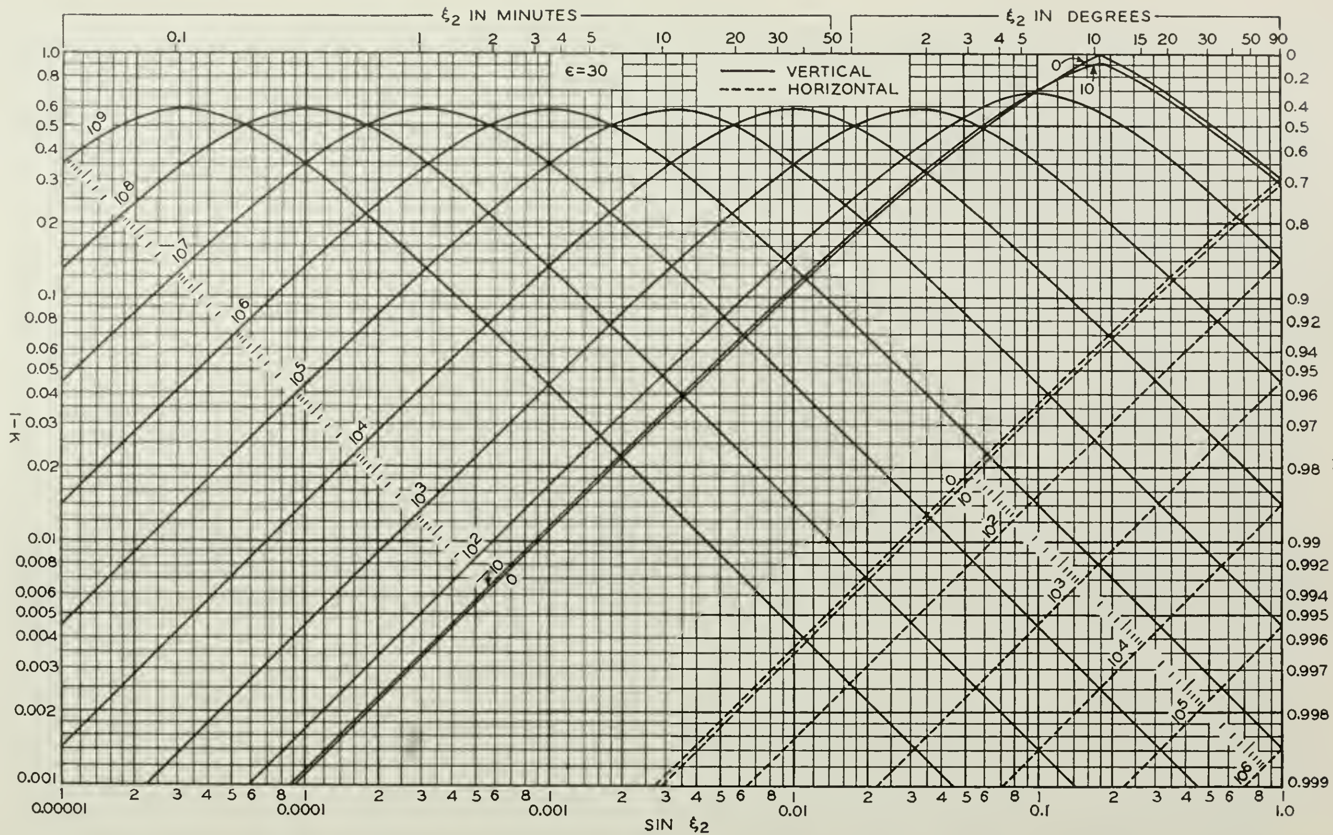


Fig. 9—Magnitude of reflection coefficient for $\epsilon = 30$. The number on each curve gives the value of q $(= 2\sigma/f)$ to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

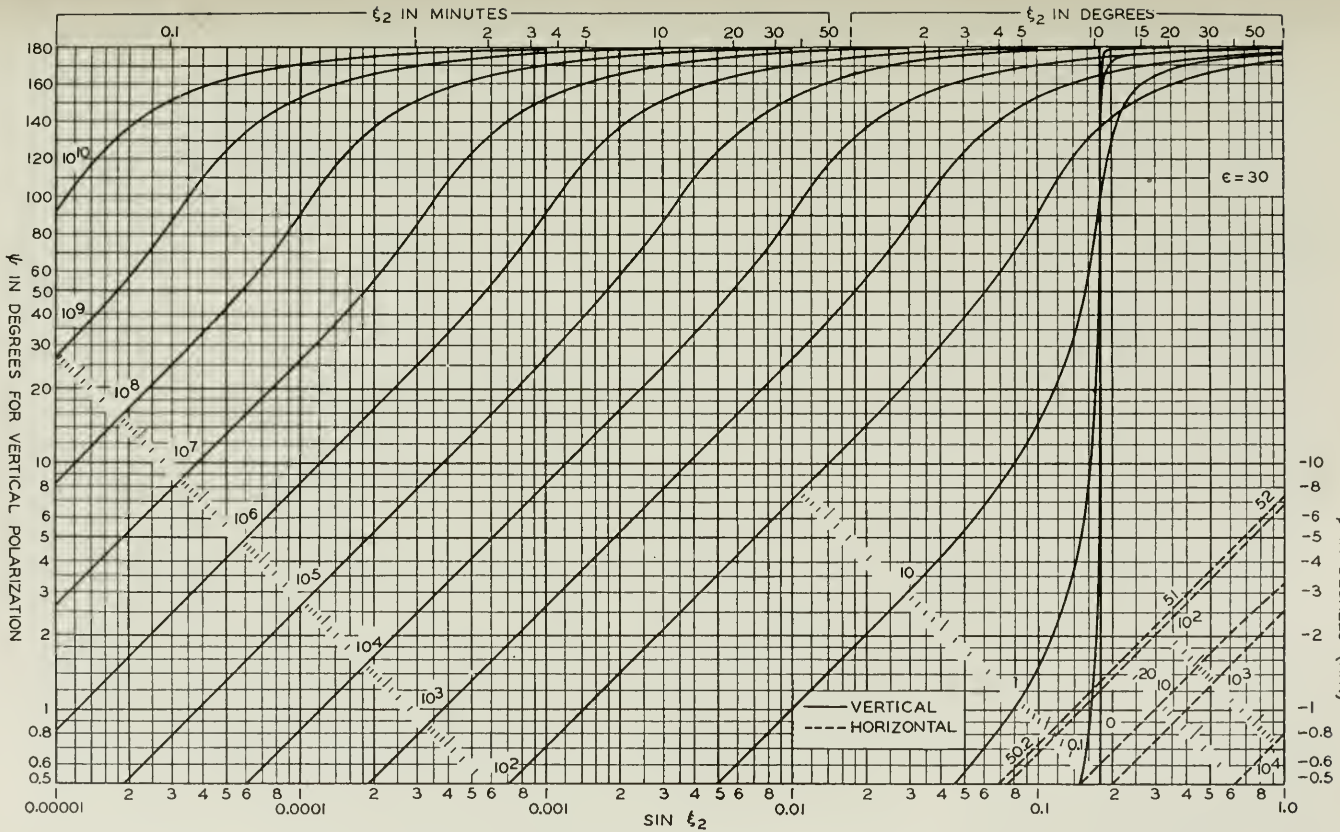


Fig. 10—Phase shift at reflection for $\epsilon = 30$. The number on each curve gives the value of q ($= 2\sigma/f$) to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

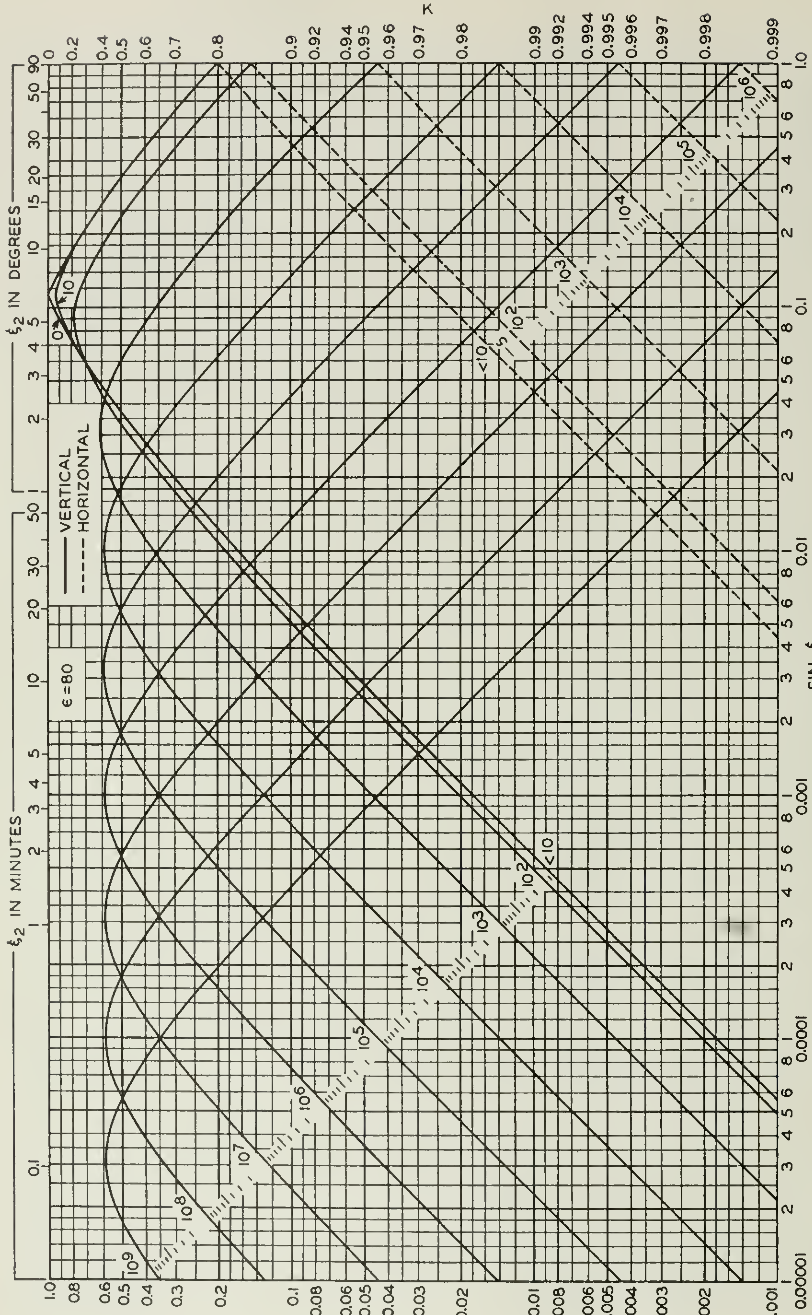


Fig. 11—Magnitude of reflection coefficient for $\epsilon = 80$. The number on each curve gives the value of q ($= 2\sigma/f$) to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

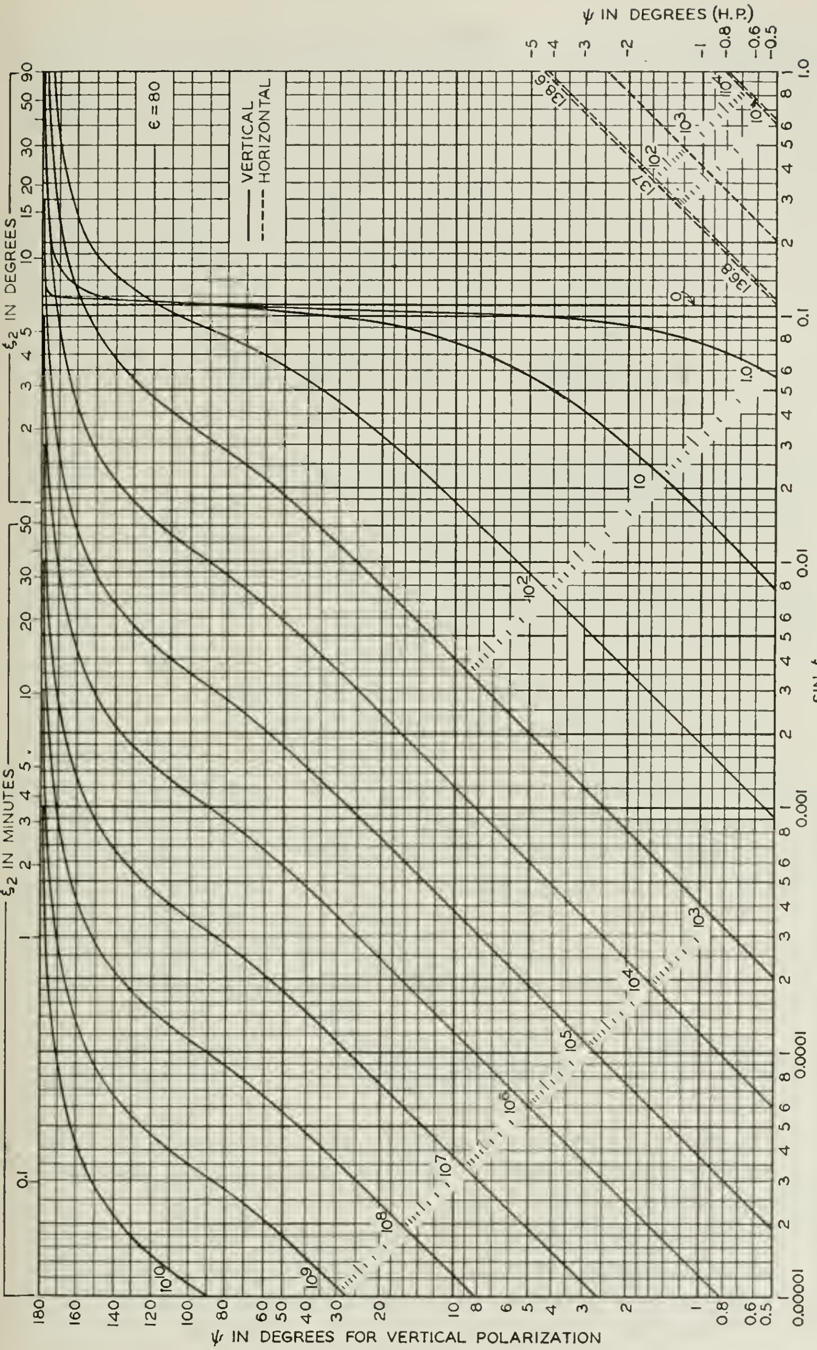


Fig. 12—Phase shift at reflection for $\epsilon = 80$. The number on each curve gives the value of q $(= 2\sigma/f)$ to which it applies. Vertical polarization is shown by solid lines, horizontal polarization by broken lines.

of the curves in this cycle may be used to obtain the values of $1 - K$ and ψ for any value of ξ_2 for which the curves go below the edge of the charts, as follows. Multiply ξ_2 by the smallest power of ten that will give a value on the chart, read the value of $1 - K$ or ψ corresponding to this new ξ_2 and divide this value of $1 - K$ or ψ by this power of ten to obtain the desired value of $1 - K$ or ψ. To obtain values of $1 - K$ and ψ for larger values of $q = 2\sigma/f$ than shown on Figs. 5–12 use is made of the fact that both of these quantities may be expressed as functions of the parameter $\sqrt{q}\,\sin\,\xi_2$ for large values of q, $(Q \ll 1)$. That is, the shape of all the curves for values of Q that are small compared with unity is the same. Hence to obtain values of $1 - K$ and ψ for values of q greater than those for which curves are shown, divide the given q by some power of one hundred that gives a value of q for which a curve is drawn, and read the desired value of $1 - K$ or ψ opposite the value of $\sin\,\xi_2$ that is the same power of ten times the given $\sin\,\xi_2$ as the power of one hundred by which q was divided.

If the following characteristics of the curves are taken into consideration, interpolation is simplified. The similarity of the K-curves suggests relabelling the abscissa so that the value of K for some intermediate value of q may be read from one of the curves that is drawn. Any curve for a large value of q, $(Q \ll 1)$, such as for $q = 10^n$ may be relabelled $q = x \times 10^n$ if the value of the abscissa is divided by $\sqrt{x}$. The same method is useful for small values of q but in this case the quantity by which the abscissa must be divided depends on the value of Q. Also in this case the shape of the curves changes with Q so that it is desirable to read the values from the curves drawn for the nearest values of q on either side of the desired value. The factor by which the abscissa must be multiplied to obtain the desired result may be inferred from the interpolation scale on the curves. When q is large the same method of interpolation may be employed for the ψ-curves as for the K-curves. When q is small, $(Q \gg 1)$, the fact that ψ is proportional to q suggests the method of interpolation. (On vertical polarization when ξ_2 is greater than the Brewster angle, $\pi - \psi$ is proportional to q.) If the value of ψ for $q = x \times 10^{-n}$ is required, read the value of ψ from the curve for $q = 10^{-n}$ and multiply by x to obtain the desired value of ψ. If greater accuracy is required the values may be calculated from the equations and Table III of Appendix II without prohibitive labor.

When the antennas approach the ground, the ratio given in equation (21) approaches zero so that more terms of the complete solution must be taken into consideration. Wise [10] has derived an expression for the effect of the ground on the Hertzian potential which when added to the primary disturbance gives the following expression. Since it is now known[4] that no exponential term must be added to this expression, it may be used to calculate the received field.

$$\Pi = \frac{HI}{4\pi}\left[\frac{e^{-2\pi iR_1/\lambda}}{R_1} + \frac{e^{-2\pi iR_2/\lambda}}{R_2}\sum_{n=1}\frac{g_n}{(-2\pi iR_2/\lambda)^{n-1}}\right], \tag{24}$$

where the geometry and nomenclature are given in Fig. 4.

$g_1 = R = -Ke^{i\psi}$ is the reflection coefficient and

$$g_{n+1} = \frac{n-1}{2}g_n - \frac{\sin\xi_2}{n}g_n' + \frac{\cos^2\xi_2}{2n}g_n'', \tag{25}$$

where $\pi/2 - \xi_2$ is the angle of incidence and the primes denote differentiation with respect to $\sin\xi_2$. Performing the operation indicated in (11) on (24) the complete expression for the received field strength on vertical polarization is found to be

$$\begin{aligned} E = -\frac{60\pi iHI}{\lambda}\Bigg\{ & \frac{e^{-2\pi iR_1/\lambda}}{R_1}\cos^2\xi_1 + \frac{e^{-2\pi iR_2/\lambda}}{R_2}g_1\cos^2\xi_2 + \frac{e^{-2\pi iR_1/\lambda}}{R_1}\,\frac{1-3\sin^2\xi_1}{2\pi iR_1/\lambda} \\ & + \frac{e^{-2\pi iR_2/\lambda}}{R_2}\,\frac{g_1(1-3\sin^2\xi_2)+2g_1'\sin\xi_2\cos^2\xi_2 - g_2\cos^2\xi_2}{2\pi iR_2/\lambda} \\ & + \frac{e^{-2\pi iR_1/\lambda}}{R_1}\,\frac{1-3\sin^2\xi_1}{(2\pi iR_1/\lambda)^2} + \frac{e^{-2\pi iR_2/\lambda}}{R_2}\Bigg[\frac{g_1(1-3\sin^2\xi_2)+5g_1'\sin\xi_2\cos^2\xi_2}{(2\pi iR_2/\lambda)^2} \\ & + \frac{-g_1''\cos^4\xi_2 - g_2(1-5\sin^2\xi_2) - 2g_2'\sin\xi_2\cos^2\xi_2 + g_3\cos^2\xi_2}{(2\pi iR_2/\lambda)^2}\Bigg] \\ & + \sum_{n=3}\frac{e^{-2\pi iR_2/\lambda}}{R_2}\Bigg[\frac{g_{n-1}(n-1)(1-[n+1]\sin^2\xi_2)}{(-2\pi iR_2/\lambda)^n} \\ & + \frac{(2n+1)g_{n-1}'\sin\xi_2\cos^2\xi_2 - g_{n-1}''\cos^4\xi_2}{(-2\pi iR_2/\lambda)^n} \\ & + \frac{-g_n(1-[2n+1]\sin\xi_2) - 2g_n'\sin\xi_2\cos^2\xi_2 + g_{n+1}\cos^2\xi_2}{(-2\pi iR_2/\lambda)^n}\Bigg]\Bigg\}. \end{aligned} \tag{26}$$

The first term on the right of equation (26) is the vertical component of the electric field radiated by a vertical electric doublet in free space. The second term is the corresponding component reflected from the earth. The third and fifth terms are sometimes referred to as the induction and electrostatic components respectively. The remaining terms complete the effect of the ground. When the antennas approach the ground $R_1 \to R_2$, $\cos\xi_1 \to \cos\xi_2 \to 1$ and $g_1 \to -1$ so that the first two terms tend to cancel. Under these conditions the sum of the first four terms of equation (26) may be written *

$$\frac{E}{E_0} = 1 + \left[R + \frac{(R+1)^2 \lambda d}{4\pi i (h_1 + h_2)^2} \right] e^{-4\pi i h_1 h_2 / \lambda d}. \tag{27}$$

At the greater antenna heights this expression also gives the correct result provided the distance, d, is large compared with the sum of the antenna heights, $h_1 + h_2$. For smaller antenna heights this expression is limited to distances for which the magnitude of the second term within the bracket is small (say less than 0.1). If this term is not small more terms of equation (26) must be taken into consideration. While equation (26) applies to vertical polarization only, equation (27) applies to both polarizations within the region for which it is valid provided the appropriate reflection coefficient is employed.

For antenna heights sufficiently small that the exponential factor of equation (27) may be replaced by the first two terms of its series expansion,

$$\frac{E}{E_0} = \frac{4\pi i h_1 h_2}{\lambda d} \left[1 + \frac{(a - ib)\lambda}{4\pi i h_1} \right] \left[1 + \frac{(a - ib)\lambda}{4\pi i h_2} \right], \tag{28}$$

where a and b are given in Table III of Appendix II (page 72). The first factor gives the well known expression for ultra-short-wave propagation over level land. The second two factors are important for antennas near the ground. When $h_1 \rightarrow h_2 \rightarrow 0$ this becomes

$$\frac{E}{E_0} = \frac{(a - ib)^2 \lambda}{4\pi i d}, \tag{29}$$

which is equivalent to the first term of the asymptotic expansion of the attenuation factor given in Part I.

A more useful form of equation (28) is

$$\frac{E}{E_0} = \frac{4\pi h_1 h_2}{\lambda d} \left[1 + a_1 \left(\frac{1}{h_1} + \frac{1}{h_2} \right) + a_2 \left(\frac{1}{h_1} + \frac{1}{h_2} \right)^2 + a_3 \frac{1}{h_1 h_2} + a_4 \frac{1}{h_1 h_2} \left(\frac{1}{h_1} + \frac{1}{h_2} \right) + a_5 \frac{1}{h_1^2 h_2^2} \right]^{1/2}, \tag{30}$$

* Equation (27) differs from equation (26) only in the dropping of terms in $1/d^3$. By leaving the exponential factor and the coefficient of reflection unexpanded the useful range of this formula is increased. The term

$$\frac{1}{2\pi i d/\lambda} \left[1 + R e^{-4\pi i h_1 h_2 / \lambda d} \right]$$

has been omitted from the right side of equation (27) since it can be shown that this term is always small compared with the remaining terms when $2\pi d/\lambda \gg 1$. In order to facilitate calculations by means of the reflection coefficient curves, $-g_2$ is replaced by $(R + 1)^2 d^2 / 2(h_1 + h_2)^2$ to which it is equal to the required order of approximation. Another form of equation (27) that may be preferred in some cases is given as equation (35) in the conclusions. This form results from substituting for $-g_2$ its value $(a - ib)^2/2$.

TABLE I

	a_1	a_2	a_3	$a_4 = a_1 a_2$	$a_5 = a_2^2$
1. Both V.P. and H.P.	$-\frac{b}{2\pi/\lambda}$	$\frac{a^2+b^2}{4(2\pi/\lambda)^2}$	$-\frac{a^2-b^2}{2(2\pi/\lambda)^2}$	$-\frac{a^2b+b^3}{4(2\pi/\lambda)^3}$	$\frac{(a^2+b^2)^2}{(4\pi/\lambda)^4}$
2. V.P. in general	$-\frac{\sqrt{2}[q\sqrt{s+r}-\epsilon\sqrt{s-r}]}{s(2\pi/\lambda)}$	$\frac{\epsilon^2+q^2}{s(2\pi/\lambda)^2}$	$-\frac{2[r\epsilon^2+(2\epsilon-r)q^2]}{s^2(2\pi/\lambda)^2}$	———	$\frac{(\epsilon^2+q^2)^2}{s^2(2\pi/\lambda)^4}$
3. V.P. $Q \ll 1$	$-\frac{\sqrt{2q}}{2\pi/\lambda}$	$\frac{q}{(2\pi/\lambda)^2}$	$-\frac{2(\epsilon+1)}{(2\pi/\lambda)^2} \to 0$	$-\frac{q\sqrt{2q}}{(2\pi/\lambda)^3}$	$\frac{q^2}{(2\pi/\lambda)^4}$
4. V.P. $Q \gg 1$	$-\frac{2q/\sqrt{\epsilon-1}}{2\pi/\lambda} \to 0$	$\frac{\epsilon^2/(\epsilon-1)}{(2\pi/\lambda)^2}$	$-\frac{2\epsilon^2/(\epsilon-1)}{(2\pi/\lambda)^2}$	$-\frac{2q\epsilon^2/(\epsilon-1)^{3/2}}{(2\pi/\lambda)^3} \to 0$	$\frac{\epsilon^4/(\epsilon-1)^2}{(2\pi/\lambda)^4}$
5. H.P. in general	$\frac{\sqrt{2}\sqrt{s-r}}{s(2\pi/\lambda)}$	$\frac{1/s}{(2\pi/\lambda)^2}$	$-\frac{2r/s^2}{(2\pi/\lambda)^2}$	$\frac{\sqrt{2}\sqrt{s-r}/s^2}{(2\pi/\lambda)^3}$	$\frac{1/s^2}{(2\pi/\lambda)^4}$
6. H.P. $Q \ll 1$	$\frac{\sqrt{2/q}}{2\pi/\lambda}$	$\frac{1/q}{(2\pi/\lambda)^2}$	$-\frac{2(\epsilon-1)/q^2}{(2\pi/\lambda)^2} \to 0$	$\frac{\sqrt{2/q^3}}{(2\pi/\lambda)^3}$	$\frac{1/q^2}{(2\pi/\lambda)^4}$
7. H.P. $Q \gg 1$	$\frac{q/(\epsilon-1)^{3/2}}{2\pi/\lambda} \to 0$	$\frac{1/(\epsilon-1)}{(2\pi/\lambda)^2}$	$-\frac{2/(\epsilon-1)}{(2\pi/\lambda)^2}$	$\frac{q/(\epsilon-1)^{5/2}}{(2\pi/\lambda)^3} \to 0$	$\frac{1/(\epsilon-1)^2}{(2\pi/\lambda)^4}$

Where $Q = \epsilon f/2\sigma$, $q = 2\sigma/f = \epsilon/Q$, $r = \epsilon - 1 + \sin^2 \xi_2$ and $s = \sqrt{r^2 + q^2}$. See Appendix II for further evaluations of a and b.

where the values of the a's are given in Table I. A similar expression results for horizontal polarization so that the a's are also evaluated for this case.

With the aid of equation (9) this may be expressed as a power ratio between short doublets:

$$\sqrt{\frac{P_r}{P_t}} = \frac{3h_1h_2}{2d^2}\left[1 + a_1\left(\frac{1}{h_1}+\frac{1}{h_2}\right) + a_2\left(\frac{1}{h_1}+\frac{1}{h_2}\right)^2 + a_3\frac{1}{h_1h_2} + a_4\frac{1}{h_1h_2}\left(\frac{1}{h_1}+\frac{1}{h_2}\right) + a_5\frac{1}{h_1^2h_2^2}\right]^{1/2}. \quad (31)$$

For antennas at the heights above the ground that are usual in the ultra-short-wave range, the bracket in equation (31) is unity so we have the useful result that within certain limitations the ratio of received to transmitted power with simple antennas is independent of the wave-length.

When the Q of the ground is large in comparison with unity, equation (30) reduces to the somewhat simpler expression

$$\frac{E}{E_0} = \frac{4\pi h_1h_2}{\lambda d}\sqrt{\left[1+\frac{a_0^2}{h_1^2}\right]\left[1+\frac{a_0^2}{h_2^2}\right]}, \quad (32)$$

where $a_0^2 = \epsilon^2\lambda^2/4\pi^2(\epsilon-1)$ for vertical polarization and $a_0^2 = \lambda^2/4\pi^2(\epsilon-1)$ for horizontal polarization. Likewise when the Q of the ground is small in comparison with unity, equation (30) reduces to

$$\frac{E}{E_0} = \frac{4\pi h_1h_2}{\lambda d}\sqrt{\left[1+\frac{b_0}{h_1}+\frac{b_0^2}{2h_1^2}\right]\left[1+\frac{b_0}{h_2}+\frac{b_0^2}{2h_2^2}\right]}, \quad (33)$$

where $b_0 = -\sqrt{2q}\,\lambda/2\pi$ for vertical polarization and $b_0 = \sqrt{2/q}\,\lambda/2\pi$ for horizontal polarization.

Equations (28), (30), (31), (32) and (33) are valid for all distances beyond those for which the received field strength begins to vary inversely with the square of the distance provided the antennas are not too high. This range of validity contains all practical distances for ultra-short-wave propagation over land and fresh water and the longer distances for ultra-short-wave propagation over sea water. For antennas at greater heights above the ground, equation (27) is required; but usually the range of antenna heights between those for which equation (21) and those for which equation (30) are valid, is small.

The applicability of the approximate equations to the problem in hand may be ascertained as follows. First calculate the parameter x of Fig. 2 to determine if the distance is sufficient for the field strength to be inversely proportional to the square of the distance. The deviation of the attenuation curve from the straight line $E/2E_0 = 1/x$ shows the degree of this approximation for antennas on the ground. If this is satisfactory then equation (27) applies.* An evaluation of the parameters in equation (27) for the greatest antenna height allows a determination of whether equations (28) and (30) apply. If $\boldsymbol{R}$ is within the range where it is a linear function of ξ_2, that is if the curves for $1 - K$ *vs* $\sin \xi_2$ (Figs. 5, 7, 9 and 11) are straight lines for this value of ξ_2, and if $\sin 4\pi h_1 h_2/\lambda d$ is approximately equal to $4\pi h_1 h_2/\lambda d$, then equations (28) and (30) apply. If also Q is very different from unity then either equation (32) or equation (33) applies.

An evaluation of the parameters in equation (27) for the lowest height will allow a determination of whether equation (21) applies. If the second term within the brackets is small compared with the first term, equation (21) applies.

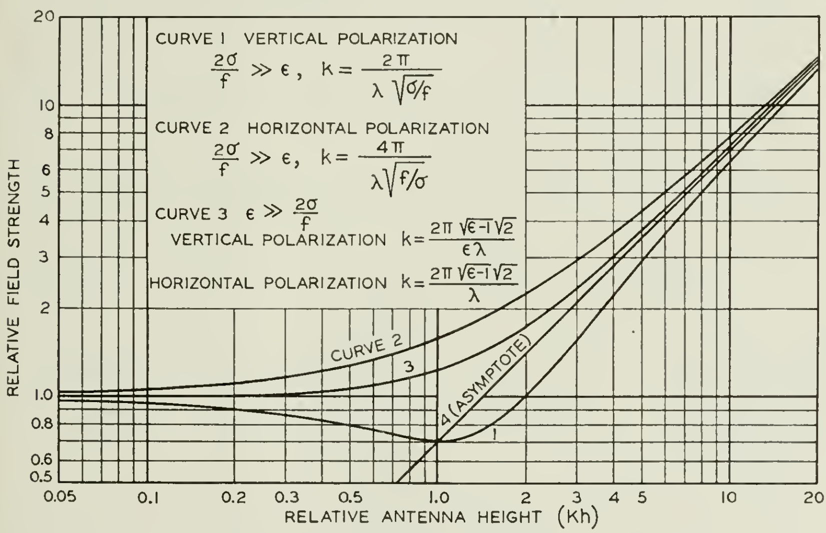


Fig. 13—Variation of received field strength with antenna height.

The variations of the received field strength with antenna height for the four cases of especial interest given by equations (32) and (33) are plotted in Fig. 13. The ordinate gives the ratio of the field strength at the height corresponding to the abscissa to that for zero height. If both antennas are off the ground the product of the ratios corresponding to the antenna heights gives the ratio of the field strength to

* As the antennas are removed from the earth's surface the error introduced because of this deviation is less.

that for both antennas on the ground. The distances between the curves and the straight line labelled " asymptote " give the magnitudes of the factors in equations (32) and (33) by which $(4\pi h_1 h_2/\lambda d)E_0$ must be multiplied to give the field strength.

For transmission over a ground of good conductivity ($Q \ll 1$) with vertical antennas there is a least favorable height for the antennas as indicated by curve 1 of Fig. 13. With both antennas at this height, which is about $1.7\lambda^{3/2}$ meters for ocean water, the received field is one-half what it would be if both antennas were on the ground.

Curve 2 for transmission on horizontal polarization over ground of good conductivity ($Q \ll 1$) shows a steady increase in the received field with increase in antenna height. If curves 1 and 2 were plotted against antenna height in meters for any given ground conditions ($Q \ll 1$), curve 2 for horizontal polarization would not depart appreciably from its asymptote until such small antenna heights were reached that curve 1 for vertical polarization would be substantially independent of antenna height. Hence curves 1 and 4 give a comparison of the received field strength at any height on the two polarizations. At the height for which the field strength is minimum on vertical polarization, the field strength is independent of polarization. For lower antennas vertical polarization gives the greater fields, while for higher antennas horizontal polarization gives the greater fields. The maximum advantage of horizontal polarization over vertical polarization occurs at twice this height and is a factor of two.

As Q increases curves 1 and 2 merge into curve 3 for transmission over a perfect dielectric. While the shape of the curves for the two polarizations is identical and the received field strength is independent of polarization at the greater antenna heights, the field strength is ϵ^2 times as great on vertical polarization as on horizontal polarization with antennas on the ground.

As an example of the use of the curves for the reflection coefficient the relative advantages of different types of ground for low-angle reception (or transmission) on vertical polarization has been calculated. With vertical antennas both the direct and the reflected components are reduced by the factor $\cos \xi_2$ so that the right-hand side of equation (21) must be multiplied by $\cos \xi_2$. The receiving antenna will be assumed to be on the ground.* Figure 14 gives the resulting curves for the indicated ground constants. For very low angles the curves are parallel, indicating that the relative advantages of different types of ground are independent of the angle at these angles. The gain in

* For higher angles of reception the relative advantages of different types of ground may be made approximately the same by properly adjusting the antenna height.

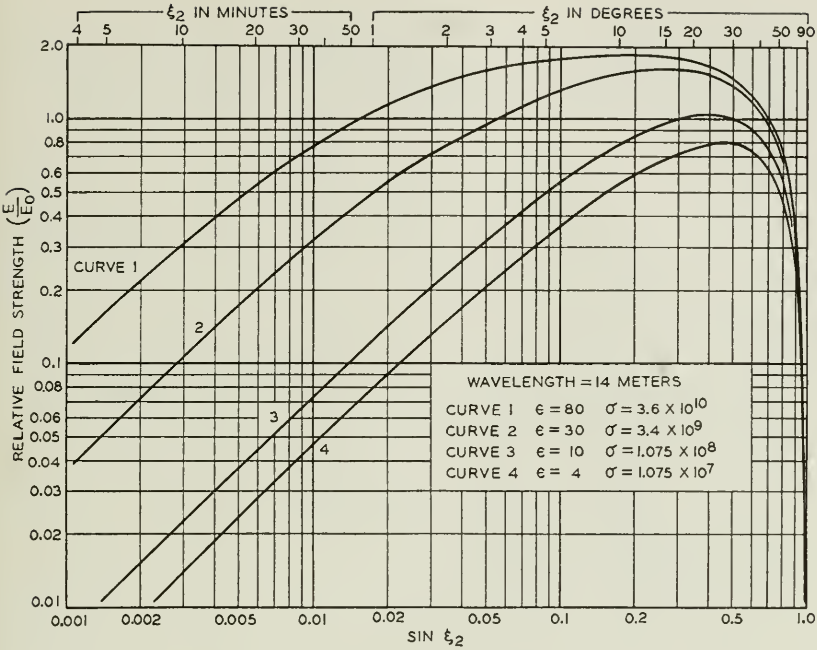


Fig. 14—Relative advantage of different types of ground for low-angle reception on 14 meters with vertical polarization.

locating an antenna above sea water instead of above the following grounds on a wave-length of 14 meters is given in Table II.

TABLE II

	Ground Constants		Gain in db for reception at			
	Dielectric Constant — ϵ	Conductivity — σ				
	Electrostatic Units	Electrostatic Units	Small Angles	1°	2°	5°
1. Sea Water . . .	80	3.6×10^{10}	0	0	0	0
2. Salt Marsh . . .	30	3.4×10^{9}	10	7	5	3
3. Dry Ground . .	10	1.075×10^{8}	24	19	15	11
4. Rocky Ground.	4	1.075×10^{7}	28	23	20	14

Acknowledgment

The value of a paper of this type depends to a large degree upon its freedom from errors. With this thought in mind the equations and tables have been checked by my associate, Mr. Loyd E. Hunt, whose cooperation is hereby acknowledged.

Conclusions

For transmission over plane earth with antennas on the ground the received field strength in volts per meter is given by the formula,

$$E = \frac{120\pi HI}{\lambda d} F(x) = \frac{3\sqrt{10}\sqrt{P}}{d} F(x), \tag{34}$$

where HI is the transmitting ampere-meters, P is the radiated power in watts exclusive of ground losses, d the distance in meters, λ the wave-length in meters and $F(x)$ is the factor plotted in Fig. 2. When the Q of the ground is large compared with unity, the factor plotted in Fig. 3 is to be preferred to that plotted in Fig. 2.

When the antennas are not on the ground the received field strength may be calculated by means of equation (27) or its equivalent,

$$\frac{E}{E_0} = 1 + \left[R + \frac{(a - ib)^2 \lambda}{4\pi i d} \right] e^{-4\pi i h_1 h_2/\lambda d}, \tag{35}$$

where the reflection coefficient,

$$R = -Ke^{i\psi} = -1 + (a - ib) \sin \xi_2 + \cdots. \tag{36}$$

The quantities K and ψ are plotted in Figs. 5–12.

When the antennas are sufficiently removed from the ground that the second term within the bracket of equation (35) may be neglected the simpler expression given in equation (21) applies.

$$E/E_0 = \sqrt{(1 - K)^2 + 4K \sin^2 (\gamma/2)}. \tag{21}$$

When the distance between antennas is sufficiently great that the exponential factor in equation (35) may be replaced by the first two terms in its series expansion, the received field strength may be calculated by equation (30) and Table I. Four special cases are given by equations (32) and (33) and Fig. 13.

Appendix I

The values of the components of

$$W = A + D \approx A - B/2 + F \approx C + F$$

are:

$$A = \frac{1}{1 - \tau^4} \left[1 + \sum_{n=1}^{\infty} \left(\frac{2\pi i \tau^2 d/\lambda}{1 + \tau^2} \right)^n \frac{a_n}{1 \cdot 3 \cdot 5 \cdots (2n - 1)} \right], \tag{37}$$

$$B = \frac{1}{1 - \tau^4} \sqrt{2\pi\sqrt{1 + \tau^2}} \sqrt{\frac{2\pi i \tau^2 d/\lambda}{1 + \tau^2}}\, e^{(2\pi i d/\lambda)(1 - 1/\sqrt{1+\tau^2})}, \tag{38}$$

$$C = -\frac{1}{1-\tau^4}\sum_{n=1}\left(\frac{1+\tau^2}{-2\pi\tau^2 id/\lambda}\right)^n[1\cdot3\cdot5\cdots(2n-1)]c_n, \tag{39}$$

$$D = -\frac{\tau^2}{1-\tau^4}e^{(2\pi id/\lambda)(1-1/\tau)}\sum_{n=0}^{\infty}D_n\left(\frac{2\pi id}{\lambda\tau}\right)^n, \tag{40}$$

$$F = \frac{\tau^2}{1-\tau^4}e^{(2\pi id/\lambda)(1-1/\tau)}\sum_{n=1}\left(\frac{1+\tau^2}{-2\pi id/\lambda\tau}\right)^n[1\cdot3\cdot5\cdots(2n-1)]f_n, \tag{41}$$

where

$$\frac{1}{\tau^2} = \epsilon - 2i\sigma/f \tag{42}$$

and τ is in the first quadrant. The positive square root of i is to be taken in equation (38).

These expressions follow from those given by Wise [11] when the sign of i is changed so that the implied time factor is $e^{i\omega t}$ in accordance with engineering practice instead of the $e^{-i\omega t}$ employed by Sommerfeld and Wise. Their expressions were derived for an antenna half in air and half in the earth. To obtain the above expressions which apply to antennas on the surface of the earth, Wise's expressions have been multiplied by $2/(1+\tau^2)$. $\boldsymbol{A}$ corresponds to his expression (5), $\boldsymbol{B}$ to his (12), $\boldsymbol{C}$ to his (8) and $\boldsymbol{D}$ to his (6). The quantities, a_n and c_n are substantially unity except when τ is not small.

$$\left.\begin{aligned} a_1 &= \frac{\tanh^{-1}\sqrt{k}}{\sqrt{k}} = \sum_{n=1}^{\infty}\frac{k^{n-1}}{(2n-1)}, \qquad a_2 = \frac{3(a_1-1)}{k}, \\ a_n &= \frac{(2n-1)(2n-3)(a_{n-1}-a_{n-2})}{(n-1)^2k}, \end{aligned}\right\} \tag{43}$$

$$\left.\begin{aligned} c_1 &= 1, \qquad c_2 = 1 - \frac{k}{3}, \\ c_n &= c_{n-1} - \frac{(n-1)^2k}{(2n-1)(2n-3)}c_{n-2}, \end{aligned}\right\} \tag{44}$$

$$\left.\begin{aligned} D_0 &= 1, \qquad D_1 = \sqrt{l}\tanh^{-1}\sqrt{l} = \sum_{n=1}^{\infty}\frac{l^n}{(2n-1)}, \\ D_2 &= D_1 - l, \quad D_n = \frac{(2n-3)D_{n-1} - lD_{n-2}}{(n-1)^2}, \end{aligned}\right\} \tag{45}$$

where

$$k = \frac{\tau^2}{1+\tau^2} = \frac{1}{\epsilon + 1 - 2i\sigma/f}, \tag{46}$$

$$l = \frac{1}{1+\tau^2} = \frac{\epsilon - 2i\sigma/f}{\epsilon + 1 - 2i\sigma/f}, \tag{47}$$

The f_n's are the same functions of l that the c_n's are of k.

Appendix II

The phase angle introduced by the path difference is:

$$\Delta = \frac{2\pi}{\lambda}\left[\sqrt{d^2 + (h_1 + h_2)^2} - \sqrt{d^2 + (h_1 - h_2)^2}\right]$$
$$= \frac{4\pi h_1 h_2}{\lambda d}\left[1 - \frac{h_1^2 + h_2^2}{2d^2} + \frac{3h_1^4 + 10h_1^2h_2^2 + 3h_2^4}{8d^4} - \cdots\right]. \quad (48)$$

The magnitude and phase of the reflection coefficient are given by the following equations.

$$R = -Ke^{i\psi}$$
$$K = \sqrt{\frac{1-\alpha}{1+\alpha}} = 1 - \alpha + \tfrac{1}{2}\alpha^2 - \tfrac{1}{2}\alpha^3 + \tfrac{3}{8}\alpha^4 + \cdots,$$
$$\alpha = \frac{mx}{1+x^2}, \quad \tan\psi = \frac{nx}{1-x^2}, \quad \sin\xi_2 = x/\sqrt{c},$$
$$m = a/\sqrt{c}, \qquad n = b/\sqrt{c}, \qquad \xi_2 = \frac{\pi}{2} - \theta,$$

where θ is the angle of incidence.

$$q = 2\sigma/f = \epsilon/Q, \quad r = \epsilon - 1 + \sin^2\xi_2,$$
$$s = \sqrt{r^2 + q^2}, \qquad \frac{r}{s} = \left[1 + \left(\frac{q}{r}\right)^2\right]^{-1/2},$$

where ϵ and σ are respectively the dielectric constant and conductivity of the ground in electrostatic units and f is the frequency in cycles per second. The values a, b and c depend on the polarization and ground constants as shown in Table III below.

For grazing incidence, $\xi_2 \to 0$, $1 - K \to a\xi_2$ and $\psi \to b\xi_2$. On vertical polarization near normal incidence for $\epsilon^2 + q^2 \gg 1$, $x \gg 1$ and the approximations $1 - K \to a/c \sin\xi_2$ and $\psi \to \pi - b/c \sin\xi_2$ are useful. These coefficients are given in Table III.

For normal incidence, $\sin\xi_2 = 1$, $\alpha = \dfrac{\sqrt{2}\sqrt{s+\epsilon}}{s+1}$ and $\tan\psi = \dfrac{\sqrt{2}\sqrt{s-\epsilon}}{1-s}$.

At Brewster's angle, $\cot\xi_2 = \sqrt{\epsilon}$ and $K = 0$ when $q = 0$ for vertical polarization. For vertical polarization the minimum value of K is $\sqrt{\dfrac{2-m}{2+m}}$ and occurs when $x = 1$ and $\sin\xi_2 = \sqrt{\dfrac{s}{\epsilon^2 + q^2}}$. For horizontal polarization the maximum value of b occurs when $q = \sqrt{3}\,r$ and is equal to $-1/\sqrt{2r}$. Under these conditions $s = 2r$ and $n = -1$.

TABLE III

	m	n	$\frac{1}{\sqrt{c}}$	$a = m\sqrt{c}$	$b = n\sqrt{c}$	$\frac{a}{c} = \frac{m}{\sqrt{c}}$	$\frac{b}{c} = \frac{n}{\sqrt{c}}$
1. Vertical Polarization in General	$\sqrt{2}\left[\sqrt{\frac{1+\frac{r}{s}}{1+\left(\frac{q}{\epsilon}\right)^2}} + \sqrt{\frac{1-\frac{r}{s}}{1+\left(\frac{\epsilon}{q}\right)^2}}\right]$	$\sqrt{2}\left[\sqrt{\frac{1+\frac{r}{s}}{1+\left(\frac{\epsilon}{q}\right)^2}} - \sqrt{\frac{1-\frac{r}{s}}{1+\left(\frac{q}{\epsilon}\right)^2}}\right]$	$\sqrt{\frac{s}{\epsilon^2+q^2}}$	$\frac{\sqrt{2}}{s}[\epsilon\sqrt{s+r} + q\sqrt{s-r}]$	$\frac{\sqrt{2}}{s}[q\sqrt{s+r} - \epsilon\sqrt{s-r}]$		
2. V. P., $Q \ll 1$	$\sqrt{2}$	$\sqrt{2}$	$\frac{1}{\sqrt{q}}$	$\sqrt{2q}$ *	$\sqrt{2q}$ *	$\sqrt{\frac{2}{q}}$	$-\sqrt{\frac{2}{q}}$
3. V. P., $Q \gg 1$	2	$\frac{2r-\epsilon}{\epsilon r}q$	$\frac{\sqrt{r}}{\epsilon} = \sqrt{\frac{\epsilon-1}{\epsilon^2-x^2}}$	$\frac{2\epsilon}{\sqrt{\epsilon-1}}$ *	$\frac{(\epsilon-2)q}{(\epsilon-1)^{3/2}}$ *	$\frac{2\sqrt{\epsilon-1+\sin^2\xi_2}}{\epsilon}$	$\frac{7\epsilon-8+8\sin^2\xi}{(\epsilon-1+\sin^2\xi)}\frac{\sqrt{\epsilon-1+\sin^2\xi_2}}{4\epsilon^2}q$
4. Horizontal Polarization in General	$\sqrt{2}\sqrt{1+\frac{r}{s}}$	$-\sqrt{2}\sqrt{1-\frac{r}{s}}$	$\sqrt{s}$	$\frac{\sqrt{2}\sqrt{s+r}}{s}$	$-\frac{\sqrt{2}\sqrt{s-r}}{s}$		
5. H. P., $Q \ll 1$	$\sqrt{2}$	$-\sqrt{2}$	$\sqrt{q}$	$\sqrt{\frac{2}{q}}$	$-\sqrt{\frac{2}{q}}$		
6. H. P., $Q \gg 1$	2	$-\frac{q}{r}$	$\sqrt{r} = \frac{\epsilon-1}{1-x^2}$	$\frac{2}{\sqrt{\epsilon-1+\sin^2\xi_2}}$	$\frac{q}{[\epsilon-1+\sin^2\xi_2]^{3/2}}$		

* For vertical polarization, $x \to 0$ requires $\xi_2 \to 0$. The tabulated values of a and b in rows 2 and 3 are true only for $\xi_2 \to 0$. For horizontal polarization x is never large.

The f_n's are the same functions of l that the c_n's are of k.

Appendix II

The phase angle introduced by the path difference is:

$$\Delta = \frac{2\pi}{\lambda}\left[\sqrt{d^2 + (h_1 + h_2)^2} - \sqrt{d^2 + (h_1 - h_2)^2}\right]$$

$$= \frac{4\pi h_1 h_2}{\lambda d}\left[1 - \frac{h_1^2 + h_2^2}{2d^2} + \frac{3h_1^4 + 10h_1^2h_2^2 + 3h_2^4}{8d^4} - \cdots\right]. \quad (48)$$

The magnitude and phase of the reflection coefficient are given by the following equations.

$$R = -Ke^{i\psi}$$

$$K = \sqrt{\frac{1-\alpha}{1+\alpha}} = 1 - \alpha + \tfrac{1}{2}\alpha^2 - \tfrac{1}{2}\alpha^3 + \tfrac{3}{8}\alpha^4 + \cdots,$$

$$\alpha = \frac{mx}{1+x^2}, \quad \tan\psi = \frac{nx}{1-x^2}, \quad \sin\xi_2 = x/\sqrt{c},$$

$$m = a/\sqrt{c}, \qquad n = b/\sqrt{c}, \qquad \xi_2 = \frac{\pi}{2} - \theta,$$

where θ is the angle of incidence.

$$q = 2\sigma/f = \epsilon/Q, \quad r = \epsilon - 1 + \sin^2\xi_2,$$

$$s = \sqrt{r^2 + q^2}, \quad \frac{r}{s} = \left[1 + \left(\frac{q}{r}\right)^2\right]^{-1/2},$$

where ϵ and σ are respectively the dielectric constant and conductivity of the ground in electrostatic units and f is the frequency in cycles per second. The values a, b and c depend on the polarization and ground constants as shown in Table III below.

For grazing incidence, $\xi_2 \to 0$, $1 - K \to a\xi_2$ and $\psi \to b\xi_2$. On vertical polarization near normal incidence for $\epsilon^2 + q^2 \gg 1$, $x \gg 1$ and the approximations $1-K \to a/c \sin\xi_2$ and $\psi \to \pi - b/c \sin\xi_2$ are useful. These coefficients are given in Table III.

For normal incidence, $\sin\xi_2 = 1$, $\alpha = \dfrac{\sqrt{2}\sqrt{s+\epsilon}}{s+1}$ and $\tan\psi = \dfrac{\sqrt{2}\sqrt{s-\epsilon}}{1-s}$.

At Brewster's angle, $\cot\xi_2 = \sqrt{\epsilon}$ and $K = 0$ when $q = 0$ for vertical polarization. For vertical polarization the minimum value of K is $\sqrt{\dfrac{2-m}{2+m}}$ and occurs when $x = 1$ and $\sin\xi_2 = \sqrt{\dfrac{s}{\epsilon^2 + q^2}}$. For horizontal polarization the maximum value of b occurs when $q = \sqrt{3}\,r$ and is equal to $-1/\sqrt{2r}$. Under these conditions $s = 2r$ and $n = -1$.

			$\frac{a}{c} = \frac{m}{\sqrt{c}}$	$\frac{b}{c} = \frac{n}{\sqrt{c}}$
Vertical Polarization in General		$r]$		
V. P., $Q \ll 1$		*	$\sqrt{\frac{2}{q}}$	$-\sqrt{\frac{2}{q}}$
V. P., $Q \gg 1$		*	$\frac{2\sqrt{\epsilon - 1 + \sin^2 \xi_2}}{\epsilon}$	$\frac{7\epsilon - 8 + 8 \sin^2 \xi}{(\epsilon - 1 + \sin^2 \xi)} \frac{\sqrt{\epsilon - 1 + \sin^2 \xi_2}}{4\epsilon^2} q$
Horizontal Polarization in General				
H. P., $Q \ll 1$				
H. P., $Q \gg 1$				

* For vertical polarization,

Appendix III

In using the equations and curves of this paper to calculate the field, the ground constants appropriate to the location of interest should be employed. The literature on ground constants is already large and is continually increasing. An exhaustive summary of this literature would be out of place here, but as an aid to those who do not have available the ground constants of the locality in which they are interested, the following table is presented.

The first four sets of values have been widely used. The conductivities of grounds 1 and 5 have been accepted by the Madrid Conference as representative of ocean water and average ground. The conductivities of grounds 5 to 8 were obtained from field strength surveys.[12] The conductivities of water 9 to 11 were obtained from sample measurements by Mr. L. A. Wooten of these laboratories at a temperature of 25° C. Both the conductivity and dielectric constant of water vary appreciably with the temperature, approximately in accordance with the relationships

$$\sigma = \sigma_{25^\circ}(1 + 0.02t),$$
$$\epsilon = 80 - 0.4(t - 20),$$

where t is the temperature in degrees centigrade.* The conductivity also varies from place to place in the ocean due to changes in its composition. The constants of grounds 12 to 15 were obtained from measurements on samples by Mr. C. B. Feldman of these laboratories. The constants of grounds 16 and 17 are typical of measurements made by Dr. R. L. Smith-Rose on English soil.[14]

In general, both the conductivity and dielectric constant of the ground vary with temperature, moisture content and frequency as well as location. For a more complete treatment and extensive bibliographies see C. B. Feldman [13] and R. L. Smith-Rose.[14]

Column 6 of Table IV gives the frequency for which $Q = 1$ for each type of ground. At higher frequencies $Q > 1$ and the ground tends to resemble a dielectric; at lower frequencies it tends to resemble a conductor.

Columns 7, 8, 9 and 10 give the values of the parameter x of Fig. 2 for a distance of 1 km. and the indicated frequency. For any other distance, x is equal to these values times the distance in kilometers. When $Q \ll 1$, x is proportional to the distance and to the square of the frequency. When the frequency is small compared with that

* The first equation was obtained from the values given for sodium chloride in the International Critical Tables. The second equation is given in the same source for pure water.

TABLE IV

Type of Ground	Relative Dielectric Constant	Conductivity in emu	Conductivity in esu	Conductivity in mhos per meter	Values of f_0 in $Q=f/f_0$ Mc.	For Fig. 2: When $Q \ll 1$, $x=c_1df^2/f_1^2$. Values of c_1 in km^{-1} for $f_1=50$ kc. (6000 m.)	$f_1=500$ kc. (600 m.)	$f_1=50$ mc. (6 m.)	For Fig. 2: When $Q \gg 1$, $x=c_2df/f_2$. Values of c_2 in km^{-1} for $f_2=50$ mc.	For Figs. 5–12: Values of $q=2\sigma/f$ for $f=50$ kc.	$f=500$ kc.	$f=50$ mc.
1	2	3	4	5	6	7	8	9	10	11	12	13
1. Sea Water	80	10^{-11}	0.9×10^{10}	1	225	2.91×10^{-6}	2.91×10^{-4}	2.91		360,000	36,000	360
2. Fresh Water	80	10^{-14}	0.9×10^{7}	10^{-3}	0.225	0.291			13	360	36	0.36
3. Wet Ground	10	5×10^{-14}	4.5×10^{7}	5×10^{-3}	9.0	0.058	5.8		94	1,800	180	1.8
4. Dry Ground	4	10^{-15}	0.9×10^{6}	10^{-4}	0.45	2.91			196	36	3.6	0.036
5. Middle West U. S.		10^{-13}	0.9×10^{8}	10^{-2}		0.029	2.9			3,600	360	3.6
6. Texas		3×10^{-13}	2.7×10^{8}	3×10^{-2}		0.01	1.0			10,800	1,080	10.8
7. Southeast U. S.		4×10^{-14}	3.6×10^{7}	4×10^{-3}		0.072	7.2			1,440	144	1.44
8. New England		2×10^{-14}	1.8×10^{7}	2×10^{-3}		0.145	14.5			720	72	0.72
9. Atlantic Ocean near N. J.		4.32×10^{-11}	3.9×10^{10}	4.32	975	0.67×10^{-6}	0.67×10^{-4}	0.67	13	1,560,000	156,000	1,560
10. Average fresh water lakes in N. J.		6.7×10^{-14}	6×10^{7}	6.7×10^{-3}	1.5	0.043	4.3		13	2,400	240	2.4
11. Lake Michigan		2.44×10^{-13}	2.2×10^{8}	2.44×10^{-2}	5.5	0.012	1.2		13	8,800	880	8.8
12. Holmdel, N. J.	15–25	$1–2\times10^{-13}$	$1–2\times10^{8}$	$1–2\times10^{-2}$	8–26				40–65			
13. Netcong, N. J.	5–10	$1–3\times10^{-14}$	$1–3\times10^{7}$	$1–3\times10^{-3}$	2–12				94–167			
14. Japan	15	2×10^{-14}	1.8×10^{7}	2×10^{-3}	2.4	0.145	14.5		65	720	72	0.72
15. Philippine Islands	12	3×10^{-14}	2.7×10^{7}	3×10^{-3}	4.5	0.1	10		80	1,080	108	1.08
16. Dry Soil	3–4	1.1×10^{-16}	10^{5}	1.1×10^{-5}	0.06					4	0.4	0.004
17. Moist Soil	30–40	$1–2\times10^{-13}$	$1–2\times10^{8}$	$1–2\times10^{-2}$	5–13							

given in Column 6 these conditions are fulfilled and the proportionality factor is given in Columns 7, 8 and 9 for three frequencies. When the frequency is large compared with that given in Column 6, $Q \gg 1$ and the parameter x of Fig. 2 is proportional to the distance and the first power of the frequency. This proportionality factor is given in Column 10 for a frequency of 50 mc.

The parameter $q = 2\sigma/f$ of Figs. 5–12 is given in Columns 10, 11 and 12 for three frequencies.

Bibliography

1. Arnold Sommerfeld, "Ausbreitung der Wellen in der drahtlosen Telegraphie. Einfluss der Bodenbeschaffenheit auf gerichtete und ungerichtete Wellenzuge," *Jahrb. d. drahtl. T. u. T.* **4**, 157–176, December, 1910.
2. Bruno Rolf, "Numerical Discussion of Prof. Sommerfeld's Attenuation Formula for Radio Waves," Ingeniörs Vetenskaps Akademien, Stockholm, 1929 and "Graphs to Prof. Sommerfeld's Attenuation Formula for Radio Waves," *Proc. I.R.E.*, **18**, 391–402, March, 1930.
3. W. Howard Wise, "Note on the Accuracy of Rolf's Graphs of Sommerfeld's Attenuation Formula," *Proc. I.R.E.*, **18**, 1971–1972, November, 1930.
4. Chas. R. Burrows, "The Surface Wave in Radio Propagation over Plane Earth," *Proc. I.R.E.*, **25**, February, 1937.
5. E. J. Sterba, "Theoretical and Practical Aspects of Directional Transmitting Systems," *Proc. I.R.E.*, **19**, 1184–1215, July, 1931.
6. W. L. Barrow, "On the Impedance of a Vertical Half-Wave Antenna above an Earth of Finite Conductivity," *Proc. I.R.E.*, **22**, 150–167, February, 1935.
7. K. F. Niessen, "Erdabsorption bei verlikalen Dipolantennen in grosser Höeh über ebener Erde," *Ann. d. Phys.* **5**, 25, 673–687, March, 1936.
8. M. J. O. Strutt, "Strahlung von Antenen unter dem Einfluss der Erdbodeneigenschaften; (*a*) elektrische Antennen, (*b*) magnetische Antennen, (*c*) Rechnung in zweiter Näherung, (*d*) Strahlungsmessungen mit Antennen," *Ann. d. Phys.*, **5**, 1, 721–772; **4**, 1–16; **9**, 67–91, 1929–1931.
9. W. Howard Wise, "Asymptotic Dipole Radiation Formulas," *Bell. Sys. Tech. Jour.*, **8**, 662–671, October, 1929.
10. W. Howard Wise, "Note on Dipole Radiation Theory," *Physics*, **4**, 354–358, October, 1933.
11. W. Howard Wise, "The Grounded Condenser Antenna Radiation Formula," *Proc. I. R. E.*, **19**, 1684–1689, September, 1931.
12. Glenn D. Gillett and Marcy Eager, "Some Engineering and Economic Aspects of Radio Broadcast Coverage," *Proc. I.R.E.*, **24**, 190–206, February, 1936.
13. C. B. Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proc. I.R.E.*, **21**, 764–801, June, 1933.
14. R. L. Smith-Rose, "Electrical Measurements on Soil with Alternating Currents," *I.E.E. Jour.* (London) **75**, 221–237, August, 1934.

The Inductive Coordination of Common-Neutral Power Distribution Systems and Telephone Circuits *

By J. O'R. COLEMAN† and R. F. DAVIS

Early installations of three-phase, four-wire power distribution systems of the multi-grounded or common-neutral type in some cases created noise problems involving neighboring telephone circuits. Operating experience, studies of specific situations and comprehensive cooperative research over a period of years have developed means of largely avoiding difficulties of this character. The relative importance of various features of the power and telephone systems which have been found to affect the noise induction problems involved is discussed and the general cooperative procedures, most helpful in conversions to or extensions of these types of power distribution systems, are outlined.

Introduction

PRIOR to about 1915, delta-connected 2300-volt, three-phase, primary circuits were used extensively for the distribution of electric current. While some distribution networks throughout the country still operate in this manner, the marked increase in load densities, starting about 1915, often made the retention of the 2300-volt delta system impracticable. In a few instances the development of the particular network was at a point where it was feasible to change from the 2300-volt delta to a 4600-volt delta arrangement but in other cases the existing equipment represented too great an investment for a complete change of this character.

From studies of various methods of caring for the augmented load densities it was found that the existing equipment could largely be saved and the capacity of the distributing networks substantially increased by converting them to a 2300/4000-volt, star-connected, four-wire primary system. By about 1925 this system had extended to most of the larger cities and most power companies had found it economical for use in at least some parts of their territories.

In using the 2300-volt equipment on the 4000-volt, four-wire system it was necessary to stabilize the neutral conductor in some way. Most of the four-wire systems had the neutral conductor grounded at the

* Published in *Electrical Engineering*, January, 1937.

† Mr. J. O'R. Coleman, joint author of this paper, is an engineer on the staff of the Edison Electric Institute, New York City.

substation only although sometimes low-voltage lightning arresters were placed on it at various points in the distribution network to aid in its stabilization in case of a break in it. In some instances, at the time of the installation of the four-wire system, the primary neutral was connected at various points in the network to driven ground rods thus resulting in a multi-grounded neutral system. In at least one instance the neutral conductor was not solidly grounded even at the substation, it being connected to ground through lightning arresters.

The experiences of the power companies with the multi-grounded neutral were generally favorable. It was found to be more reliable and to embrace some simplifications over other distribution methods. While the early multi-grounded neutral arrangements were obtained by making connections to ground along the primary neutral conductor and interconnecting it, at service transformers, to well-grounded secondary neutrals a further simplification in the arrangement was readily apparent.

It will be noted in Fig. 1 that this interconnection of the primary and secondary neutrals resulted in two grounded neutrals on the pole line in all sections where the secondary neutral existed. In extending the multi-grounded neutral arrangement or in reconstructing existing portions of the network, these two neutrals were combined into a single well-grounded conductor continuous in all portions of a feeder area and often continuous in all parts of a substation area or of several contiguous substation areas. This arrangement, called the "common-neutral," which was first extensively applied in Minneapolis * by Mr. S. B. Hood, resulted in certain savings in equipment and relief of congested pole heads and in a neutral network most effectively grounded since all secondary neutral grounds were thus made available, in addition to any driven grounds along the pole line.

The operation of this system in Minneapolis showed many advantages in the protection of secondary networks from the effects of voltage rises under abnormal conditions. In addition a paper presented in 1925 by Mr. Hood [1] pointed out that over a period of three years the rate of transformer failure was reduced to 8/10 of 1 per cent per annum. This excellent performance in transformers arose undoubtedly from the fact that with the "common-neutral" or interconnected neutral arrangements the lightning arresters are connected directly around the transformers. Later studies showed that the connection of the lightning arresters directly between the primary conductors and secondary neutral provides a degree of protection which cannot readily be obtained in any other way.[2, 3, 4, 5, 6, 7]

* Prior to applying this system in Minneapolis, Mr. Hood introduced it at Toronto, Canada.

In urban areas, the multi-grounded or common-neutral method of distribution introduced, in some instances, noise induction in nearby telephone circuits. In view of this fact an extensive cooperative investigation was undertaken by Project Committee No. 6 of the

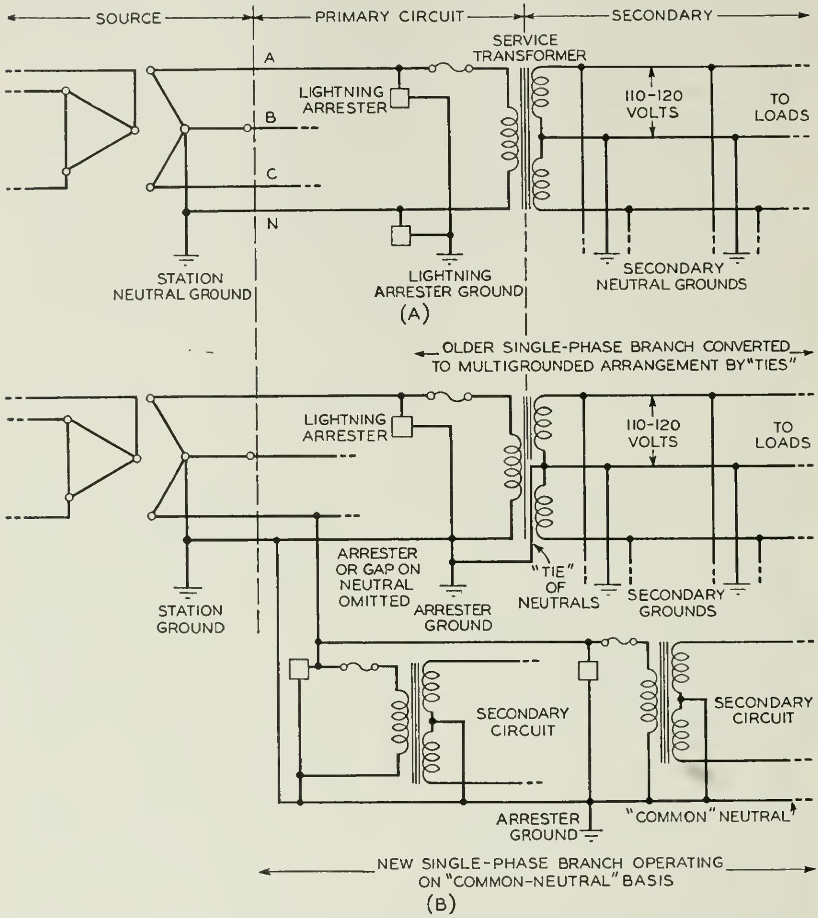


Fig. 1—*A*. Simplified feeder operating with primary neutral grounded at source only. *B*. Simplified feeder operating with multiple grounds on primary neutral-older part of feeder having ties between primary and secondary and new extensions being of common-neutral arrangement.

Joint Subcommittee on Development and Research of the National Electric Light Association and Bell Telephone System to determine the factors involved in the coordination of local power distribution systems and telephone systems. A study was carried out in Minneapolis during the years 1924 to 1926 having as its primary objective the determination of the factors involved where the telephone distribution was largely in aerial cable. The investigation was continued in Elmira, New York in 1926 to 1929 to embrace the factors introduced when the exchange telephone plant was of open-wire construction.[8] Supplementing these detailed technical studies, an investigation of certain economic features of various arrangements of power and telephone distributing methods and of their practical application under varying conditions was carried out in California in 1928 and 1929. As a result of these investigations the various factors involved in the coordination of multi-grounded or common-neutral power systems and telephone distribution systems were determined and certain practices developed for the coordination of these systems under various conditions in urban areas.[9]

The purpose of this paper is that of briefly outlining a few of the more important features of power and telephone circuits affecting noise coordination. Following such a review there is presented a list of measures which extended experience has shown will, where given proper consideration by both parties, enable multi-grounded or common-neutral power circuits and telephone circuits to live harmoniously. No attempt is made, however, to reiterate the extensive technical information obtained from the investigations outlined above as these are adequately covered in the references cited.

Recent Trends

During the past few years there has been extensive conversion from other types of urban power distribution to the multi-grounded or common-neutral system of primary distribution. Where there exists a three-phase, three-wire delta circuit the system is converted by making the secondary neutral network continuous, reinforcing it where necessary, and making the required changes in transformer connections. Where there is a three-phase, four-wire uni-grounded primary system the conversion is, as previously mentioned, made by interconnecting the primary and the secondary neutral at each load transformer generally removing the primary neutral only at the time of major rebuilding. In either case extensions are usually made using a single neutral in the secondary position.

In the urban areas most of the multi-grounded or common-neutral systems are of the 2300/4000-volt class, although there are a few instances where 4600-volt systems have been converted. At the present time there is being constructed a 6900/12,000-volt common-neutral distribution system at Wichita, Kansas.

The distinct trend in power distribution practice has been, in no small measure, influenced by the improved overall protection features readily obtained by the multi-grounded neutral arrangement as well as by certain equipment savings. The recent emphasis placed on the electrification of rural areas and the distinct need for maximum service continuity on rural power circuits has increased the interest in the use of the multi-grounded or common-neutral method of distribution in rural areas. The rural systems are generally of the 7600/13,200-volt class, although 4600/8000-volt circuits have been used to some extent.

The factors affecting inductive coordination involved in the use of the common-neutral method of distribution in rural areas, are somewhat different from those encountered in urban areas. This is largely due to the lower load densities, the greater lengths of circuits, higher operating voltages and to the somewhat different types of equipment employed in rural telephone distribution. These factors were investigated by the Joint Subcommittee on Development and Research during the summers of 1935 and 1936 and the more important considerations determined.[10]

Factors Influencing Inductive Coordination

In any problem of inductive coordination it is convenient to subdivide the factors influencing coordination into those relating to the inductive influence of the power circuit, the inductive susceptiveness of the telephone circuit and the inductive coupling between the two types of circuits. As far as urban distribution circuits are concerned the load current unbalance of the power circuit is usually the controlling influence factor. For rural distribution circuits the unbalanced charging currents are generally more important than the unbalanced load currents. Likewise, in an exchange telephone circuit the admittance and impedance unbalances of the two sides of the circuit are usually the controlling factors in its inductive susceptiveness. As far as coupling between the power and telephone circuits is concerned this is largely controlled by their relative positions and the lengths of the exposure. For urban areas their relative positions are largely fixed by the normal arrangement of conductors and cables on jointly-used poles. In rural areas power and telephone circuits are generally at roadway separation although some joint use exists. In urban areas considerable control can often be exercised over the coupling by planning the routes of the main feeds of the two services so as to avoid long sections of close exposure. In rural areas where there are no paralleling routes close together it is generally necessary for both services to use the same roads and therefore the opportunity to control the coupling by the cooperative planning of routes is much reduced.

Certain quantitative indications of the extent to which this measure of coordination is applicable in the two types of areas are shown in the illustrative examples in the Appendix.

Power Circuits

Power systems operate, for the most part, at frequencies of 60 cycles and below. Telephone circuits, on the other hand depend mainly upon frequencies above about 200 cycles for the transmission of speech. Ordinarily, therefore, the effects of induction from the fundamental frequency currents and voltages in neighboring power lines are negligible as far as telephone circuit noise is concerned. It is quite generally recognized, however, that it is impracticable commercially to build rotating machinery and transformers which are entirely free from harmonics. There are, therefore, harmonics present on all operating power systems and it is the harmonic-frequency components induced into telephone circuits from these power system harmonics that are of major importance from the noise standpoint.[11]

In any distribution circuit the harmonic currents present will fall within the following classes:—load currents, transformer-exciting currents and line-charging currents. With a uni-grounded neutral the load currents and the transformer-exciting currents are practically entirely confined to the wires of the circuit. Where the neutral is multi-grounded, the vector sum of the currents in the phase conductors (residual current) will divide between the neutral conductor and the paralleling earth path as determined by the relative impedances of these two paths. While there is some variation in the division of the return current between the neutral and ground paths, for most practical purposes this division may be assumed to be about half in each path at all the frequencies of interest.

As pointed out above, in the case of a line operating with uni-grounded neutral, the earth-return components of the load and transformer-exciting currents are ordinarily negligibly small. However, this is not true of the line-charging current which is chiefly a function of the magnitude and frequency of the impressed voltage, the circuit length, and, at non-triple harmonic frequencies, of the balance of the admittances to ground of the various phase conductors. While multi-grounding the neutral ordinarily increases the earth-return components of the load and transformer-exciting currents, it has been found, due to the parallel path provided by the neutral wire, on an average to decrease slightly the amount of charging current in the earth.

In an urban distribution system where the load density is relatively high, the load currents and transformer-exciting currents are relatively large and the line-charging currents are usually negligible. In such a system multi-grounding the neutral results in an increase in the current returning through the earth and a consequent increase in the inductive influence of the power distribution system.

In rural areas, however, where the load density is low and the load currents and transformer-exciting currents are relatively small, the line-charging currents become significant. In general, under such conditions the multi-grounding of the neutral does not increase the magnitude of the ground-return current at frequencies of interest from the noise induction viewpoint. Under certain conditions the magnitude of this ground-return current may actually be substantially decreased by the multi-grounding of the neutral. This effect is more marked for the higher voltage circuits.

The harmonics present in a distribution circuit may be divided into (1) triple harmonics, that is, the third harmonic and odd multiples of it, and (2) non-triple harmonics, that is, the odd harmonics, starting with and including the fifth, which are not multiples of three. The triple harmonics in a three-phase system are in phase in the three line conductors so that their residual value (vector sum) is the arithmetic sum of their magnitudes in the three-phase wires. The non-triple harmonics are spaced, in time phase, the same as the 60-cycle fundamental and the magnitude of the residual current (vector sum) for these harmonics is usually much less than their arithmetic sum. If these harmonics were perfectly balanced the residual current for these frequencies would be zero. In exposures involving three-phase sections of line the balance of the non-triple harmonics between phases is influenced by the degree of balance of the loads and single-phase branches and therefore has an important effect in reducing the overall influence of the power system. In exposures involving single-phase extensions, or extensions consisting only of two-phase wires and a neutral wire this advantage of the balancing of the non-triple harmonics is, of course, not obtainable.

The extent to which induction from the non-triple harmonic voltages and currents in power distribution circuits can be controlled by power circuit transpositions is ordinarily very limited. Usually, due to the large number of exposure discontinuities arising from changes in the power or telephone circuits, the power circuit transpositions are quite ineffective. This is particularly true in cases where the induction from the ground-return current is controlling. In specific cases where considerable wave shape distortion exists and the induction from the balanced voltages and currents may therefore be relatively important, transpositions in power distribution circuits may be found helpful.

Table *A* shows the average harmonics present on three-phase, four-wire industrial and residential feeders under light and heavy load conditions. The reduced magnitudes of the non-triple frequencies in the residual current (neutral and ground-return) are evident. The importance of this as regards noise induction is further indicated in the illustrative examples of the Appendix.

TABLE *A* *

Average Current and Voltage Wave Shapes of 2300/4000-Volt, 3-Phase, 4-Wire Distribution Circuits

Frequency	Order of Harmonic	Phase-to-Neutral Voltage at Bus.	Current in Industrial Feeder (In Amperes)				Current in Residential Feeder (In Amperes)			
			Light Load		Heavy Load		Light Load		Heavy Load	
			Phase	Residual	Phase	Residual	Phase	Residual	Phase	Residual
60	—	2380	65.		130		53		99	
180	*3*	*16*	*1.1*	*2.6*	*1.1*	*2.9*	*1.8*	*5.2*	*1.9*	*6.0*
300	5	21	1.0	.15	1.3	.16	.43	.21	.75	.29
420	7	6.4	.3	.03	.3	.05	.13	.06	.17	.08
540	*9*	*1.7*	*.04*	*.08*	*.04*	*.13*	*.04*	*.09*	*.04*	*.09*
660	11	1.9	.07	.01	.09	.01	.04	.01	.06	.02
780	13	1.8	.08	.01	.05	.01	.02	.01	.04	.01
900	*15*	*.42*	*.01*	*.01*	*.01*	*.03*	*.01*	*.01*	*.01*	*.01*
1020	17	.90	.03	.01	.07	.01	.02	.01	.03	.01
1140	19	.87	.02	—	.04	—	.01	.01	.02	.01
1260	*21*	*.16*	—	—	—	*.01*	—	—	*.01*	*.01*
1380	23	1.4	.05	—	.06	—	.01	—	.03	.01
1500	25	2.1	.06	.01	.09	.01	.01	.01	.03	.01
1620	*27*	*.79*	—	*.01*	*.01*	*.02*	—	—	*.01*	*.01*
1740	29	1.5	.02	—	.03	.01	.01	—	.02	.01
1860	31	1.4	.01	—	.02	—	—	—	.01	—
TIF †		9.7	11.2	—	8.9	—	6.6	—	5.8	—
Kv. T or I.T.		23.2	733	193	1160	330	346	238	571	283

Note: Triple harmonics are italicized.
* Tables 31 & 32—pp. 235 & 236 of Vol. II of Eng'g Reports of Joint Subcommittee.
† New weighting—see Engineering Report No. 33 of Joint Subcommittee.

The triple-harmonic currents present on a feeder supplied from a delta-wye substation transformer bank are generally due to the exciting currents of the single-phase load transformers. Under this condition no excessive triples are impressed on the feeder at its source as is sometimes the case where the source is a wye-connected, grounded-neutral generator directly connected to the feeder. The exciting currents flow from the individual single-phase transformers toward the delta-wye transformer in the substation. The presence on the feeder of a large three-phase wye-delta load transformer with its neutral connected to the system neutral, provides a parallel path for supplying part of these triple-harmonic exciting currents as well as part of the unbalanced non-triple and fundamental currents and under certain conditions may substantially decrease the overall inductive influence of a feeder by reducing the ground-return current flowing through an exposure. The effect of such a connection in reducing the noise is dependent upon the location of the bank with respect to the exposure and its relative impedance to the various harmonics as compared to that of the path back to the substation. From the power operating standpoint such a bank tends to supply part of the unbalanced load and also, in case of the interruption of one phase between it and the substation, tends to supply the power to that portion of the phase still connected to it. Under certain conditions, the action of such a bank may prove detrimental to the operation of the power feeder due to its action in attempting to balance the voltages at the point of its connection to the feeder. Under other conditions the neutral of an existing bank can readily be connected to the feeder neutral with distinctly beneficial effects on the inductive influence and with little or no adverse effects on the power-system operation. The tendency of such banks toward noise reduction and towards unbalanced load supply is shown in two of the illustrative examples in the Appendix.

Telephone Circuits

The voltages induced into a telephone circuit may be divided into (1) metallic-circuit induction, that is, a voltage induced between the two sides of the circuit with a resultant current flowing around the circuit, and (2) longitudinal-circuit induction, that is, a voltage induced along the conductors such that the resultant current flows in a circuit having the telephone conductors as one side and the earth as the other. This latter voltage may also result in noise, due to its action upon telephone circuit unbalances, setting up currents in the voice channel (metallic-circuit). For either type of voltage, the induction may be "electric," that is, from the voltage on the power circuit, or "magnetic" from the current in the power circuit.

The local telephone circuit may be divided into three parts: (*a*) the central office equipment, (*b*) the line conductors and (*c*) the subscriber equipment. Inter-office circuits include only the first two items.

(*a*) *Central Offices Equipment*

The central office equipment associated with a subscriber circuit consists essentially of two elements: (1) line signaling equipment connected to the circuit for indicating to the operator, or to the dial equipment, the desire of a subscriber to start a call and (2) a linking or switching circuit or circuits for interconnecting two subscriber circuits either directly or through intervening trunk circuits and providing supervision during the call.

The line signaling equipment with its associated relay is either bridged across the line or arranged so that, when two subscriber circuits are interconnected, any ground connections on the line relays are automatically opened. The line signaling equipment is not, therefore, ordinarily a factor in noise considerations. Occasionally, however, the effect on noise of the ground connection on the line signaling equipment requires specific treatment when the longitudinal-circuit induction is sufficiently high. The noise in such instances occurs either during the pre-answering period before the line relay is "cut-off" or, in certain types of switchboards, on conversations between two persons on the same line (party-line) where the use of a switching circuit in the office is unnecessary.

The linking or switching equipment in the central office may consist of a pair of wires with bridged supervisory relays as in the case of a magneto office or may be a complicated arrangement of relays, repeating coils, condensers, etc., as in the case of common-battery offices of the manual or dial type. The necessary ground connections of the latter type of apparatus introduce the possibility of the unbalances in the equipment contributing to the overall noise when the longitudinal-circuit induction on the outside conductors is impressed on the switching circuits. Ordinarily in urban areas, due either to the frequency make-up of the longitudinal-circuit induction or to the relationships of the various impedances-to-ground, the amount of noise contributed by the central office equipment is relatively low. This is readily evident from Table B which shows, at 500 and 1000 cycles, the relative proportions of overall noise due to the action of induced voltages on station, cable and central office unbalances:

TABLE *B* *

Relative Importance of Circuit Unbalances

Type of Service	500 Cycles			1000 Cycles		
	Contribution from:			Contribution from:		
	Station	Cable	Office	Station	Cable	Office
Individual (bridged ringers)	Negligible	3	3	Negligible	20	20
Party-line (grounded 8-A ringers)	100	3	3	100	20	20

* See p. 72 of Vol. I of Eng'g Reports.

Cases arise, however, quite frequently where the relative circuit impedances or the frequency make-up of the induction or both are such that the noise contribution from the unbalances in the central office equipment becomes important. Such cases usually involve long subscriber or inter-office trunk circuits and particularly where sections of open-wire construction are present. Values of the unbalances in certain types of central office equipment are given on page 91 of Volume I of the Engineering Reports of the Joint Subcommittee on Development and Research.

(*b*) *Line Conductors*

Where the telephone line conductors are in open-wire, the induced voltage between conductors (metallic-circuit) as well as along these conductors (longitudinal-circuit) must be considered. The direct metallic-circuit induction can be greatly reduced by systematic transpositions in the telephone circuit. Due to the physical limitations in a practical layout of telephone transpositions, the reduction in metallic-circuit induction is, on the average, from 60 to 80 per cent on non-pole pairs and about 90 per cent on pole-pairs. Transpositions also tend to lessen the capacitance and inductance unbalances of the two sides to ground and to other circuits, thereby reducing the effect of the longitudinal-circuit induction on such unbalances. The improved balance of the mutual impedances between the various telephone conductors is, of course, distinctly beneficial in reducing crosstalk and transpositions are generally used for limiting the crosstalk where open-wire telephone circuits extend for substantial distances.

The construction of telephone cables is such that there is inherently very close spacing between the conductors and they are frequently transposed due to the continuous twisting of the pairs in manufacture. Due to this close spacing and frequent transposing there is practically no voltage induced between the wires of a cable pair or quad (group of four conductors). The unbalance to ground of the conductors of the present type of cable is so small that it is not ordinarily a contributing factor to noise induction. It may be noted, however, from Table *B* that in cases where the central office unbalances are of importance, the effect of shunt or series unbalances in the cables also needs consideration.

The lead sheath of a telephone cable provides practically perfect shielding against induction from power system voltages when it is grounded at one or more points. The sheath likewise provides substantial magnetic shielding when it is grounded more or less continuously as in underground construction or is grounded at both ends of the aerial section or near both ends of an exposure. The degree of magnetic shielding effected varies, depending on the size of the cable

TABLE C

Magnetic Shielding for Various Frequencies
Sizes and Lengths of Telephone Cables and Various Grounding Resistances

(Calculated Values of e_r/e_i)
(e_r = Voltage remaining after shielding)
(e_i = Voltage present with sheath grounded at one point only)

Frequency	Full-Sized Cable (2.61″ Outside Dia.)						101 Pair 24 Ga. (0.85″ Outside Dia.)						51 Pair 24 Ga. (0.64″ Outside Dia.)					
	1 Mile *			3 Miles *			½ Mile *			1 Mile *			½ Mile *			1 Mile *		
	0^{ω} †	5^{ω} †	10^{ω} †	0^{ω}	5^{ω}	10^{ω}	0^{ω}	5^{ω}	10^{ω}	0^{ω}	5^{ω}	10^{ω}	0^{ω}	5^{ω}	10^{ω}	0^{ω}	5^{ω}	10^{ω}
180 Cycles	14%	82%	94%	14%	50%	71%	60%	95%	99%	60%	89%	95%	71%	96%	99%	71%	92%	96%
300 "	8.5	65	85	8.5	33	52	42	89	96	42	77	89	57	91	96.5	57	82	91
420 "	6	52	76	6	24	40	31.5	81	93	31.5	65	81	44.5	84	93.5	44.5	71	84
540 "	5	43.5	71	5	19	32	25	73	88	25	55.5	73	37	77	90	37	63	77
660 "	4	37	64	4	16	27	21	66	84	21	47.5	66	31	71	86	31	55	71
1000 "	3	25.5	44.5	3	11	18	15	51	72	15	35	51	22	56	74.5	22	41	56

* Refers to distances along cable sheath between the two grounding points.
† Refers to total grounding impedance (approx. d.c. resistance) of the two ground connections—expressed in ohms.

and the resistance of the ground connections, reaching optimum values of over 90 oer cent. Table *C* gives the magnitude of this shielding for various selected sizes and lengths of cable. Table *C* brings out distinctly the variation in the magnetic shielding due to the factors mentioned above. The effect of cable sheath shielding in several typical cases is further indicated in the Appendix.

TABLE *D*

Relative Susceptiveness of Several Types of Station Sets

General Description of Station Set	Noise in Receiver Branch for 100 Noise Units to Ground—Average Power Wave Shape (*One Station on Line-Effect of Set Only*)
Class 1—Types of sets in most common use today	
a. Sidetone type of party-line set using 8A ringers or equivalents (one end of ringer grounded). (Common-battery talking and signaling)	350 Noise Units Approximately
b. Same type—local-battery talking	120 Noise Units Approximately
c. Magneto party-line set (52A Ringer or equivalent)	120 Noise Units Approximately
d. Individual-line set—any type	Negligible
Class 2—Types of sets frequently encountered	
a. Sidetone type of 4-party full-selective or 8-party semi-selective set (using relay or cathode tube to connect ringer to circuit during ringing period)	Negligible
b. Four-party selective or 8-party semi-selective sets employing high impedance ringers or relays connected to ground	About 30
c. Eight-party selective (harmonic ringing) sets employing ringers connected to ground and tuned to 4 different ringing frequencies	Limited data indicate that, depending on frequency for which ringer is tuned, noise will range from about 100 to about 400 units
d. Ground-return rural circuits (usually of magneto type and having code ringing)	3500 or more noise units
Class 3—Special types of sets	
a. Sidetone type of party-line set using split-condenser and higher impedance ringer (one end of ringer grounded)	About 20 noise units
b. Type of party-line set using split condenser arrangement with 8A ringer or equivalents (one end of ringer grounded).	About 90 noise units

(*c*) *Station Apparatus*

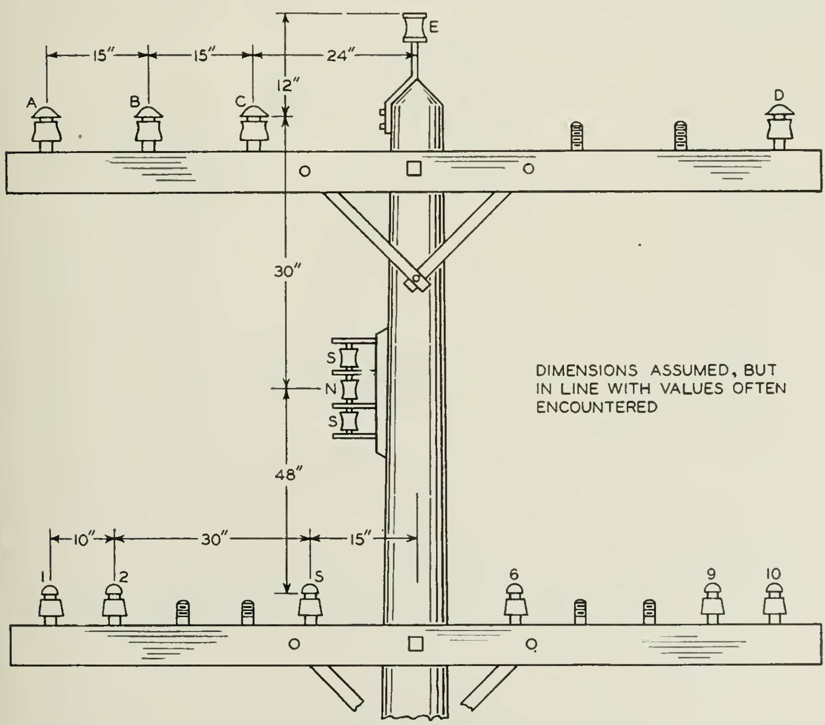
Individual-line stations employ a "bridged" ringer connection, i.e., the ringer is, in effect, connected between the two line conductors.

However, for selective signaling purposes party-line stations frequently have the ringer connected, in effect, between one of the line conductors and ground. The unbalance of the party-line station equipment is therefore affected by the impedance of the ringer and its point of connection in the station equipment. The relative susceptiveness of several types of station sets to noise-frequency induction is shown in Table *D*.

Table *D* shows that, with advance planning in areas where noise induction is or may likely become a matter of importance, much can be accomplished by the use of station sets of decreased inductive susceptiveness. Where such types of apparatus are substituted in existing plant, except in gradual replacements or in connection with general rearrangement programs, the expense is, naturally, increased.

Inductive Coupling

As stated above, the inductive coupling between exchange telephone plant and power distribution circuits in urban areas is largely controlled by such factors as the street layouts and the joint use of poles.



(See following page)

Direct Metallic Circuit Induction in Untransposed Telephone Circuits

Magnetic Induction, 1000 Cycles

Volts Metallic per Ampere of Power Circuit Current per 1000 Feet of Exposure

Type Power Circuit	Induction Component	Power Conductors*	Pair 1–2	Pair 5–6	Pair 9–10	Avg.
Single Phase	Residual	A & N	.021	.043	.0044	.022
" "	"	C & N	.0044	.018	.0053	.0092
" "	"	E & N	.0002	.023	.0013	.0082
Two Phase Wires & Neut.	Residual	A, C, & N	.013	.031	.005	.016
	"	A, D, & N	.008	.018	.009	.012
Three Phase	Residual	A, B, C, & N	.012	.033	.005	.017
	"	A, C, D, & N	.007	.01	.004	.007
Two Phase Wires & Neut.	Balanced	A, C, & N	.017	.026	.0009	.015
	"	A, D, & N	.026	.13	.026	.061
Three Phase	Balanced	A, B, C, & N	.015	.023	.012	.017
	"	A, C, D, & N	.023	.123	.026	.057

* In each case, conductor "N" is a multi-grounded neutral, the other wires being phase conductors. Assumed 50% of the residual current in the neutral and 50% in the ground.

Electric Induction

Volts per Kilovolt *Vm/Kv*

Type Power Circuit	Induction Component	Power Conductors†	Pair 1–2	Pair 5–6	Pair 9–10	Avg.
Single Phase	Residual	A & N	7	19	3	9.7
" "	"	C & N	2	7.3	2.3	3.9
" "	"	E & N	.4	9	.2	3.2
Two Phase Wires & Neut.	Residual	A, C, & N	5	13	2.6	6.9
	"	A, D, & N	2	7	2	3.7
Three Phase	Residual	A, B, C, & N	2.8	10	2	5.0
	"	A, C, D, & N	1.4	2.2	.7	1.4
Two Phase Wires & Neut.	Balanced	A, C, & N	3	7	.4	3.5
	"	A, D, & N	6	30	5	13.7
Three Phase	Balanced	A, B, C, & N	2.7	5.8	.4	3.0
	"	A, C, D, & N	5.6	27	5.4	12.7

† Wires S & S assumed continuous through exposure and grounded.

Fig. 2—Effect of relative positions on joint-use pole of power and telephone conductors on coefficients of induction for voltages and currents.

However, by cooperative planning of routes it is frequently practicable to secure lower coupling by avoiding long exposures between the main feeds of the two plants. As shown by the illustrative examples this procedure is, where applicable, very beneficial.

In rural areas where both distribution services must ordinarily be carried along the highways the opportunity for controlling the coupling between the two classes of circuits by cooperative planning of routes is much reduced.

Some benefit may be gained, however, in the case of open-wire construction particularly at joint-use separations, by arrangements of the conductors on the pole so as to avoid excessive spacings. As shown on Fig. 2 certain arrangements tend to minimize the amount of noise induction arising from the power circuit voltages and currents. This beneficial effect is, however, much less noticeable at roadway separations.

Summary and Conclusions

Since about 1915 there has been a continued increase in the use of the multi-grounded or common-neutral arrangement of power distribution in this country. At the present time, approximately half of the distribution is by 4000-volt multi-grounded or common-neutral circuits. A large part of the higher-voltage rural distribution is also operating with this arrangement.

In general it may be said that for the lower-voltage 2300/4000-volt distribution circuits, the use of the multi-grounded or common-neutral arrangement may be expected to increase the inductive influence of the power circuits. Unless attention is given to cooperative planning to secure features beneficial from the inductive coordination standpoint, noise problems may result either in restricted or extensive areas. With proper attention to the coordination features [9] such noise situations as develop are largely in the nature of isolated cases and can usually be cared for by relatively minor changes or adjustments in either or both plants.

For the higher voltage (11–13 kv.) rural distribution circuits, there seems to be little difference, from the noise induction standpoint, between the uni-grounded four-wire system and the multi-grounded or common-neutral arrangement.[10] Under many conditions the placing of multiple grounds on the neutral will result in noise reductions due to the effect, previously mentioned, of the multi-grounded neutral on the line charging currents. It is interesting to note that experience to date with the multi-grounded or common-neutral in rural areas has shown that many of the measures of coordination applicable in urban areas will prove similarly helpful in rural communities.

The measures of coordination which investigations and operating experience have shown to be practicable and effective include:

1. Cooperative planning by both parties to avoid not only severe exposure conditions but also types of equipment likely to aggravate the possible noise induction situation.

2. A reasonable degree of balance of the loads between the three phases of the power circuit. In the higher-voltage rural circuits this also includes the lengths of branches consisting of one or two phase wires and neutral.
3. The avoidance of unnecessarily heavily loaded branches consisting of one or two phase wires and neutral.
4. The prevention of excessive over-excitation of transformers.
5. The grounding, where necessary, of aerial telephone cables at or near both ends of an exposure to obtain the benefits of magnetic shielding.
6. The use of adequately coordinated telephone transpositions on open-wire extensions and the avoidance of severe unbalances in the open-wire conductors.
7. The correction of badly distorted voltage or current wave shape on the power system.
8. The connection of the neutral point of three-phase wye-delta load banks to the system neutral conductor.
9. The use of telephone station apparatus, on party-line service, of lower susceptiveness.
10. Occasionally the use of arrangements or apparatus to minimize the effects from unbalances in central office equipment.

It is, of course, essential in successfully coordinating the power distribution and telephone circuits that, as in other coordination situations, the power and telephone people view the matter as a mutual responsibility and fully cooperate in the application of the tools available. Experience over a period of years has now shown that where this is done adequate overall coordination can be readily secured.[12]

The authors wish to acknowledge their indebtedness to their many coworkers who aided in carrying on the various investigations on which this paper is based.

Bibliography

1. "Improvement in Distribution Methods," S. B. Hood—p. 1038 of 1925 *Trans. A.I.E.E.*
2. "Interconnection of Primary Lightning Arrester Ground and the Grounded Neutral of the Secondary Main," C. F. Harding and C. S. Sprague—p. 234 of 1932 *Trans. A.I.E.E.*
3. "Lightning Protection for Distribution Transformers," K. B. McEachron and L. Saxon—p. 239 of 1932 *Trans. A.I.E.E.*
4. "Lightning Protection for Distribution Transformers," A. M. Opsahl, A. S. Brooks and R. N. Southgate—p. 245 of 1932 *Trans. A.I.E.E.*
5. "Studies of Lightning Protection on 4000-volt Circuits—III," D. W. Roper—p. 252 of 1932 *Trans. A.I.E.E.*
6. "Lightning Protection for Distribution Transformers," T. H. Haines & C. A. Corney—p. 259 of 1932 *Trans. A.I.E.E.*

7. "Distribution System Lightning Studies by Philadelphia Electric Company," H. A. Dambly, H. N. Ekvall, and H. S. Phelps—p. 265 of 1932 *Trans. A.I.E.E.*
8. Engineering Reports 6, 9, 13 and 15—Vols. I and II of Engineering Reports of Joint Subcommittee on Development and Research of National Electric Light Association and Bell Telephone System.
9. Common-Neutral Practice Confirmed in California—p. 980—June 4, 1932 issue of *Electrical World*—(See also May 15, 1932 issue of *Electrical West.*)
10. Provisional Report No. 18 of Joint Subcommittee on Development and Research of Edison Electrical Institute and Bell Telephone System.
11. "Measurement of Telephone Noise and Power Wave Shape," J. M. Barstow, P. W. Blye and H. E. Kent—p. 1307 of 1935 *Trans. A.I.E.E.*
12. Reports of Joint General Committee of National Electric Light Association and Bell Telephone Systems on Physical Relations between Electrical Supply and Signal Systems—Edition of Dec. 9, 1922.

Appendix

For the sake of brevity, the detailed calculations and some of the minor assumptions for the following examples have been omitted.

Illustrative Example 1

The purpose of this example is to show, for average power system wave shapes:

1. The noise induction problem that might be created by the exposure of a reasonably long aerial telephone cable in an urban area with a heavily loaded single-phase feeder. It features:
   a. The relative importance of triple and non-triple harmonic induction, and
   b. The extent to which planning of routes, grounding of cable sheaths, etc. might improve the situation.
2. The changes in the noise magnitudes for the same situation with the various single-phase loads well distributed among all three phases. Under this condition, attention is directed to:
   a. The change in the relative importance of the triple and non-triple harmonic induction.
   b. The amount of reduction obtained by the same remedial measures tried in 1-*b* above.

Figure 3 shows a possible method of supplying the single-phase loads in a rather extensive part of an urban area. The general layout shown on Fig. 3 is such that all of the current for the feeder area traverses a considerable part of the exposure. Under this quite extreme condition —essentially single-phase supply for a relatively large area—the noise at location C under heavy load conditions would be about as shown on Table I.

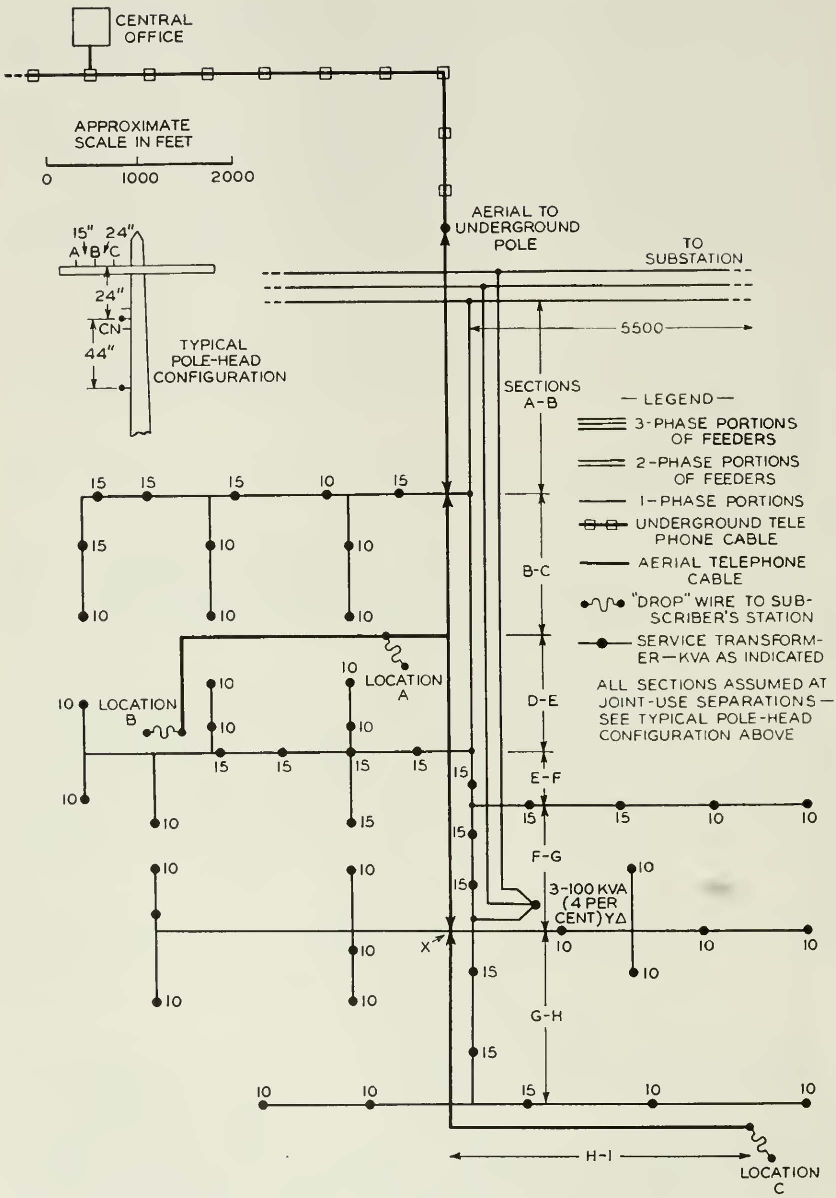


Fig. 3—Example of possible arrangement of feeder layout in an urban area where long aerial telephone cable is exposed to a heavily loaded one-phase feeder.

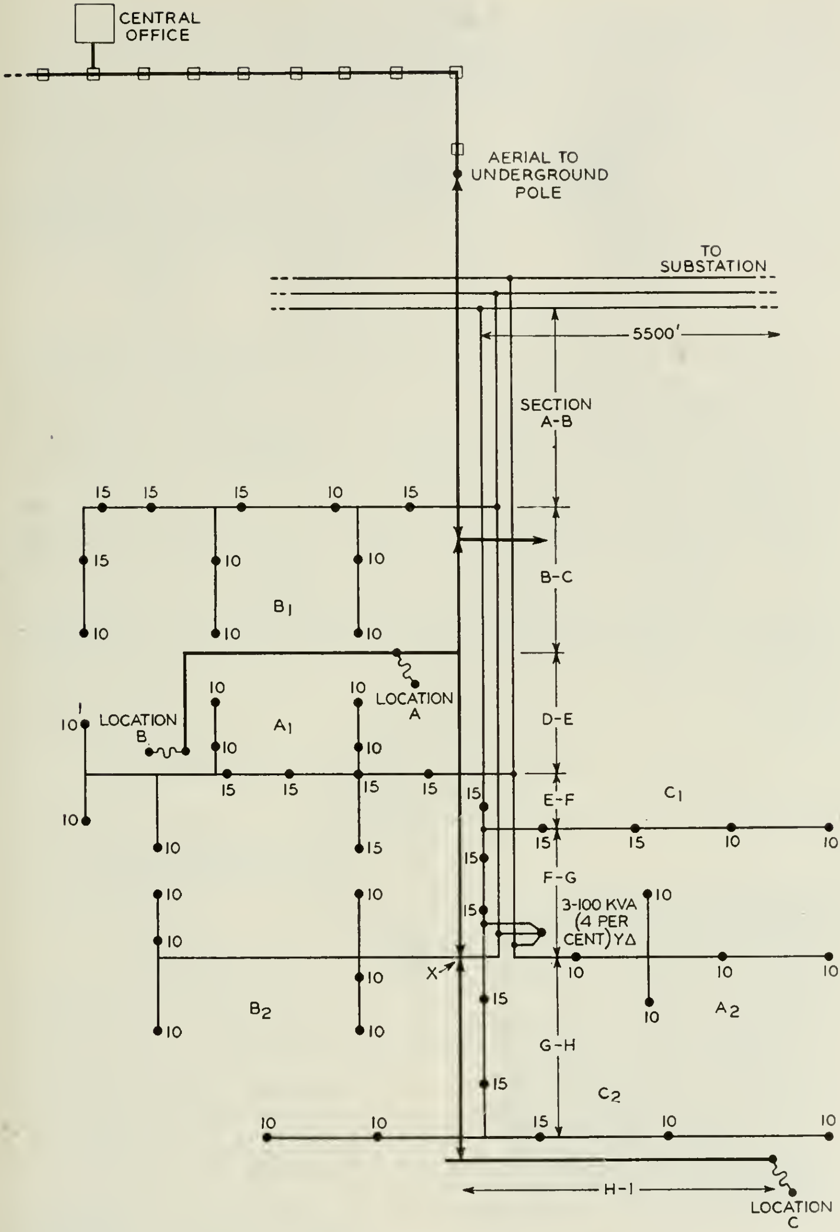


Fig. 4—Example of possible rearrangement of feeder layout shown in Fig. 3 to reduce magnitudes of "ground-return" currents.

TABLE I

NOISE CONTRIBUTIONS (FIG. 3) OF VARIOUS HARMONICS AND SECTIONS OF EXPOSURE

Section of Exposure	RMS Magnitude of Residual Current	Approximate Total Noise Contribution *	Approximate Noise Contribution From:	
			180 Cycles	300 Cycles and Higher Frequencies
A–B..........	210 amperes	1240 noise units	270 noise units	1215 noise units
B–E..........	157 "	1130 " "	250 " "	1100 " "
E–I...........	30–107 "	500 " "	100 " "	490 " "
	Total	2870 " "	620 " "	2800 " "

* Location *C*—For party-line service using 8A ringers, during heavy power loads.

It is evident from Table I that most of the party-line stations fed by the aerial telephone cable would need treatment. It will be noted from Table II that completely replacing the existing party-line stations with special station apparatus will, to a large extent, care for the situation since, in this case, the amount of noise contributed by the cable and central office unbalances would aggregate less than 150 units. Other measures either singly or in combination, probably more economical in their application, would provide substantial reductions in noise but would not be adequate for the more severely exposed stations.

TABLE II

COMPARISON OF EFFECTIVENESS OF VARIOUS REMEDIAL MEASURES (Fig. 3 Conditions)

Type of Remedial Measure	Approximate Noise on Party-Line Stations (Heavy Power Loads)	
	Location *A* or *B*	Location *C*
1. Before applying remedial measures........................	1800 noise units	2870 noise units
2. Using special telephone station sets........................	75 " "	125 " "
3. Avoiding exposure in section *A–B* by cooperative planning of routes...................	560 " "	1630 " "
4. Cable sheath shielding by tying aerial and underground telephone cables at junction pole and connecting aerial sheath to 1 ohm ground at point *X*.	1160 " "	1700 " "
5. Interconnecting 300 kva wye-delta bank with system neutral	1430 " "	1600 " " (170 kva of unbalanced load)
6. Combination of measures 3, 4, and 5.....................	365 " "	420 " "

Assume, however, that instead of supplying the single-phase loads in the area shown on Fig. 3 from one phase only, the single-phase loads were distributed reasonably uniformly among the phases. This would be advantageous not only by the noise reduction possibilities, which will be more fully discussed but also by the improved regulation attainable on the feeder. Figure 4 shows a possible rearrangement of Fig. 3 along these lines and Table III shows the noise conditions with the feeder arrangements of Fig. 4.

TABLE III

Noise Contributions (Fig. 4) of Various Harmonics and Sections of Exposure

Section of Exposure	RMS Magnitude of Residual Current	Total Noise Contribution *	Noise Contribution From:	
			180 Cycles	300 Cycles and Higher
A–B	6 amperes	280 noise units	270 noise units	75 noise units
B–E	45 "	370 " "	250 " "	275 " "
E–I	5–45 "	150 " "	100 " "	115 " "
	Total	735 " "	620 " "	375 " "

* Location *C*—For party-line service using 8A ringers, during heavy power loads.

A comparison of Tables I and III shows that the noise from the non-triple harmonics has been very materially reduced by the balancing of loads made possible by the more favorable feeder arrangement of Fig.

TABLE IV

Comparison of Effectiveness of Various Remedial Measures (Fig. 4)

Type of Remedial Measure	Approximate Noise on Party-Line Stations (Heavy Power Loads) Location *A* or *B*	Location *C*	
1. Before applying remedial measures	480 noise units	735 noise units	
2. Using special telephone station sets	25–30 " "	40–50 " "	
3. Avoiding exposure in section *a–b* by cooperative planning of routes	190 " "	460 " "	
4. Interconnecting 300 kva wye-delta bank with system neutral	330 " "	535 " "	(26 kv. of unbalanced load)
5. Cable sheath shielding—grounding at jct. pole and to 1 ohm ground at *X*	325 " "	420 " "	
6. Combinations of measure 3, 4, and 5	85 " "	245 " "	

4, although that from the triple harmonics has been inappreciably changed. The net effect has been a reduction of nearly 75 per cent in the noise on the party line stations served by the telephone cable. The reductions afforded by various remedial measures are shown in Table IV.

It is evident from Table IV that, by the application of various of the measures of coordination, the need for an extensive rearrangement of either plant is avoided.

Illustrative Example 2

The purpose of this example is to show the extent to which remedial measures of the type generally applicable in urban areas (see Example 1), may be applied in a less thickly settled area where exposures to 2.3/4 kv. multi-grounded neutral arrangements are encountered under average conditions of power system wave shape. In detail the example covers:

a. The extent to which such measures as cooperative planning of routes and use of wye-delta load banks may be ineffective.

TABLE V

Noise at Various Locations (Fig. 5) and for Various Types of Telephone Service

Location	Type of Telephone Service	Total Noise	Contribution From:		Remarks
			Cable Exposures	Open Wire Exposures	
A	1. Common-Battery Party-line stations (Class 1-*a* Table *D*)	225–345	220–270	45–215	Open-wire noise dependent on effectiveness of telephone transpositions.
	2. Magneto party-lines (Class 1-*c* Table *D*)	85–225	75 app	40–215	
	3. Individual Line (Class 1-*d* Table *D*)	40–210	25–35*	About 10–210	Lower values of noise on circuits controlled by effects of station sets—higher values by effectiveness of telephone transpositions.
B	1.	170–1200	10–100	170–1190	
	2.	75–1175	20–35	70–1175	
	3.	60–1175	About 20*	55–1175	
C	1.	400–975	285–325	275–925	
	2.	150–870	110–125	110–860	
	3.	100–860	60*	75–860	*Noise in cable section due to office and cable unbalances.

b. The importance, particularly under joint-use conditions, of noise directly induced into the metallic circuit of open-wire telephone pairs and the importance, therefore, of suitable telephone circuit transpositions.

Figure 5 shows exposure conditions such as may be encountered in a small community serving a nearby rural area and Table V shows the noise conditions at several locations during heavy power loads.

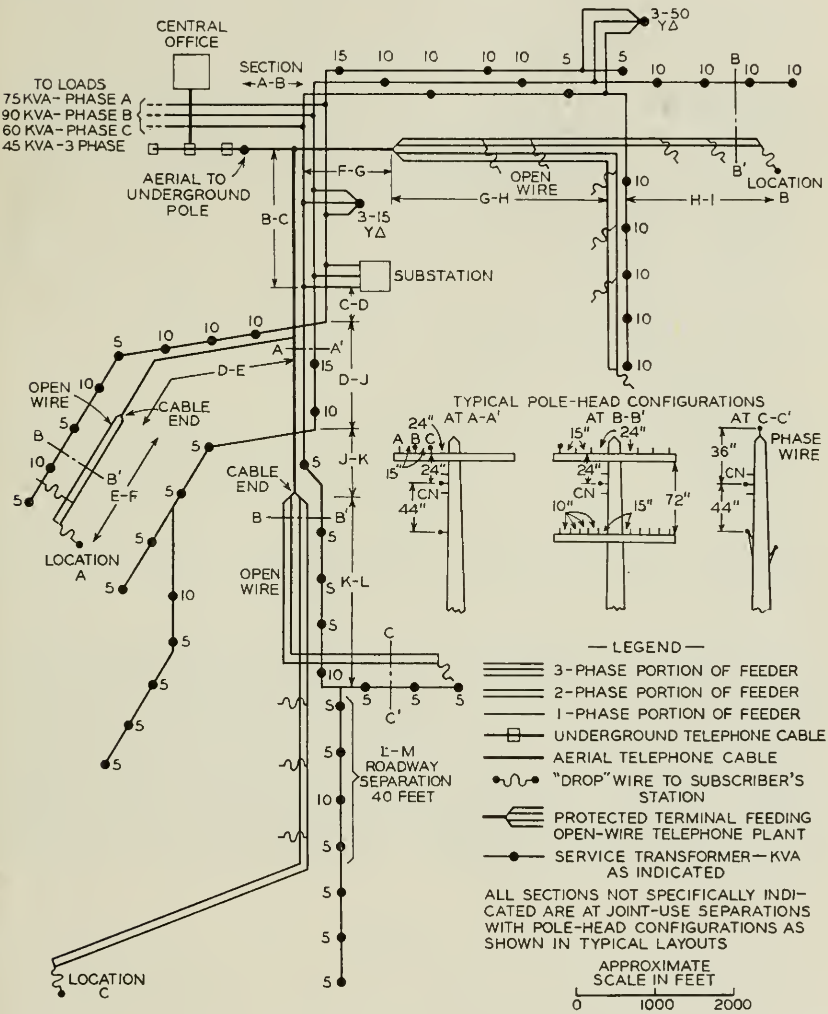


Fig. 5—Example of exposure conditions between 2300/4000-volt distribution feeder and exchange telephone plant in suburban and rural area.

The extent to which various measures of coordination could be applied to reduce the noise induction or to restrict the extent of special arrangements is shown in Table VI.

TABLE VI

COMPARISON OF EFFECTIVENESS OF VARIOUS REMEDIAL MEASURES

Type of Remedial Measure	Approximate Noise at:								
	Location *A*			Location *B*			Location *C*		
	1*	2	3	1	2	3	1	2	3
1. Before applying measures	225–345	85–225	40–210	170–1200	75–1175	60–1175	400–975	150–870	100–860
2. Using special tel. sets	40–210	40–210	Neg. Change	60–1175	60–1175	Neg. Change	100–860	100–860	Neg. Change
3. Avoiding exposure section *B–C* by cooperating planning	130–250	65–220	Neg. Change	Neg. Change			350–950	115–860	Neg. Change
4. Interconnecting neutral of 150 kva wye-delta bank	180–310	Neg. Change		150–1175	Neg. Change		370–950	Neg. Change	
5. Average degree of coordinated tel. transpositions	250	80	50	200	175	175	340	175	130
6. Tel. transpositions + cable sheath shielding †	210	70	50	180	175	175	320	165	130
7. Combination of 4, 5 and 6	185	65	50	160	155	155	310	160	125

* Type of station apparatus shown on Table V.

† Cable was assumed to be grounded at junction pole at end of Section F–G to 2.5 ohm ground; at other junction poles to grounds exceeding 10 ohms.

It is evident from this table that, for the conditions assumed, the use of reasonably coordinated telephone circuit transpositions will be necessary to care for the stations served by open-wire. Ordinarily the use of such transpositions in combination with such other measures as are reasonably effective would serve to take care of the stations served by telephone cable and would limit the extent to which special telephone station apparatus might be needed for the stations served by the longer open-wire extensions.

Series for the Wave Function of a Radiating Dipole at the Earth's Surface

By S. O. RICE

In this paper three series expansions are derived for the wave function of a vertical dipole placed at the surface of a plane earth. Two convergent series and one asymptotic series are obtained. A remainder term for the latter series is given which enables one to set an upper limit to the amount of error obtained by stopping at any particular stage in the series.

Introduction

THE wave function above the earth of a vertical dipole placed at the surface of a plane earth is [1]

$$\Pi_1(r, z) = (k_1^2 + k_2^2) \int_0^\infty \frac{J_0(\xi r) e^{-z\sqrt{\xi^2 - k_1^2}} \xi d\xi}{k_2^2 \sqrt{\xi^2 - k_1^2} + k_1^2 \sqrt{\xi^2 - k_2^2}}, \tag{1}$$

where r and z are the horizontal and vertical distances from the dipole. k_1 and k_2 are constants depending upon the electrical properties of the air and ground, respectively.[2] We shall be concerned with the value of this function at the surface of the earth. Setting $z = 0$ gives us an integral for $\Pi_1(r, 0)$ which is the function of r to be investigated here.

Although the electric and magnetic intensities are the properties of an electromagnetic field which have the greatest physical significance, writers on this subject often deal with the wave function because of its simpler form and because in many cases of practical interest it is nearly proportional to the electric intensity. However, the electromagnetic field may be obtained from the wave function by differentiation. If the real parts of $He^{-i\omega t}$ and $Ee^{-i\omega t}$ represent the electric and magnetic intensities the field above the earth produced by the dipole is

$$H_r = H_z = 0, \qquad H_\varphi = -\frac{\partial \Pi_1(r, z)}{\partial r},$$

$$E_r = \frac{ic^2}{\omega} \frac{\partial^2 \Pi_1(r, z)}{\partial r \partial z}, \qquad E_\varphi = 0, \qquad E_z = -\frac{ic^2}{\omega} \frac{1}{r} \frac{\partial}{\partial r}\left(r \frac{\partial \Pi_1(r, z)}{\partial r} \right).$$

[1] A. Sommerfeld, *Ann. der Physik*, vol. 28, pp. 665–736, No. 4 (1909).
[2] The symbols used here are defined in a list at the end of the paper.

From these expressions it will be observed that if we obtain expressions for $\Pi_1(r, 0)$ we shall be able to compute the field at the earth's surface except for the radial component E_r, which is small compared to E_z.

Statement of Results

The asymptotic expression for $\Pi_1(r, 0)$ is

$$\Pi_1(r, 0) = -\frac{1}{(1-\tau^2)r}\left[e^{ik_1 r}\sum_{n=1}^{N}\frac{n!\,P_n(k_2/s)}{(i\tau s r)^n} + R_{1N} - \tau^2 R_{20}\right], \quad (9)$$

where R_{1N} and R_{20} satisfy the inequalities

$$|R_{1N}| < \left|\frac{(N+1)!\,e^{ik_1 r}\sqrt{\csc\theta}}{[r(k_1 - s)\sin\theta]^{N+1}}\right|, \qquad |R_{20}| < \left|\frac{e^{ik_2 r}}{rk_2 - rs}\right|,$$

$\theta = \pi/2 - \arg(k_1 - s)$ being an angle slightly greater than $\pi/2$.

The convergent series for $\Pi_1(r, 0)$ are [3]

$$\Pi_1(r, 0) = \frac{1}{(1-\tau^2)r}\sqrt{\frac{\pi\tau}{2}}\left[e^{ik_1 r}\sum_{n=0}^{\infty}(-is\tau r)^n P_{-1/2}^{1/2-n}(k_1/s) \right.$$
$$\left. - \tau e^{ik_2 r}\sum_{n=0}^{\infty}\left(\frac{sr}{i\tau}\right)^n P_{-1/2}^{1/2-n}(k_2/s)\right] \quad (14)$$

and

$$\Pi_1(r, 0) = \frac{1}{(1-\tau^2)r}\left[\sum_{n=0}^{\infty}\frac{(ik_1 r)^n}{n!}F(1, -n/2; 1/2; s^2/k_2^2)\right.$$
$$\left. - \tau^2\sum_{n=0}^{\infty}\frac{(ik_2 r)^n}{n!}F(1, -n/2; 1/2; s^2/k_1^2)\right]. \quad (19)$$

The quantities τ and s are defined by $\tau = k_1/k_2$ and $1/s^2 = 1/k_1^2 + 1/k_2^2$, and the numbers on the right are the equation numbers in the text. W. H. Wise [4] has obtained series which are equivalent to those appearing in (9) and (14).

Procedure

The results given here depend upon a transformation of the integral obtained by setting $z = 0$ in equation (1). This integral can be expressed in the following way as has been shown by B. van der Pol: [5]

$$\Pi_1(r, 0) = -\frac{\tau}{1-\tau^2}\int_{k_1/s}^{k_2/s}\frac{e^{isrw}}{r}\,d(w^2-1)^{-1/2}, \quad (2)$$

[3] The Legendre functions are discussed by E. W. Hobson, "Th. of Spherical and Ellipsoidal Harmonics." Hypergeometric functions are discussed in Chap. XIV, "Modern Analysis," by Whittaker and Watson.

[4] W. H. Wise, *Proc. I.R.E.*, vol. 19, pp. 1684–1689, September 1931.

[5] *Jahrbuch der drahtlosen Telegraphie Zeitschr. f. Hochfrequenz Techn.*, *37* (1931), p. 152.

which becomes, after integration by parts,

$$= + \frac{\tau}{1 - \tau^2} \left\{ \left[\frac{- e^{isrw}}{r\sqrt{w^2 - 1}} \right]_{k_1/s}^{k_2/s} + is \int_{k_1/s}^{k_2/s} \frac{e^{isrw} dw}{\sqrt{w^2 - 1}} \right\}. \tag{3}$$

The path of integration is the straight line in the complex w plane joining the points k_1/s and k_2/s. Arg $(w - 1)$ and arg $(w + 1)$ are taken to be zero at the point this contour crosses the real axis. The Argand diagram for a typical case is shown in Fig. 1. From the definitions of k_1, k_2, and s it follows that $|s| < |k_1| < |k_2|$, and $0 = \arg k_1 < \arg s < \arg k_2 < \pi/4$.

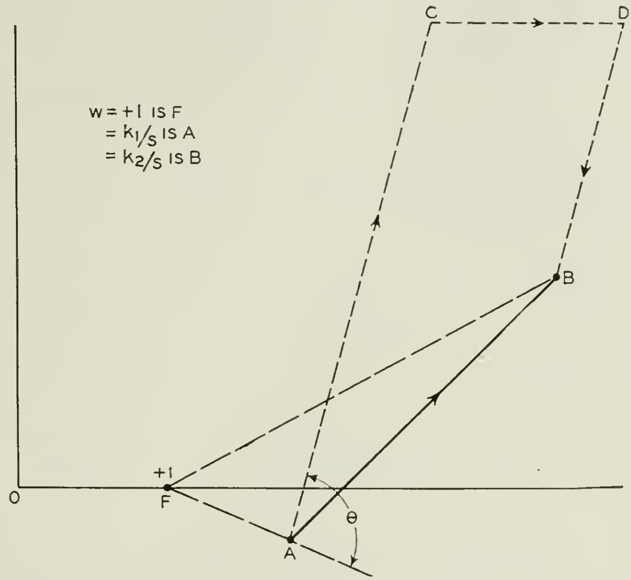


Fig. 1—Paths of integration in the w plane.

Asymptotic Expansion

To obtain an asymptotic expansion for $\Pi_1(r, 0)$ we deform the linear path joining A and B into the path $ACDB$ as is shown in Fig. 1. The lines AC and BD are both inclined to the real axis at the angle arg (is^*) where s^* is the conjugate of s. This is the direction in which the exponential term e^{isrw} decreases most rapidly since along it the variable part of the exponent is real and negative.[6] The section CD may be displaced to infinity where its contribution to the value of the integral becomes zero because of this exponential decrease.

[6] To show this for the line AC we set $w = k_1/s + is^*u$. As w goes from A to C u is real and increases from zero. The exponent then becomes $isrw = ik_1r - |s|^2ru$ since $ss^* = |s|^2$.

The integral $\Pi_1(r, 0)$ is then composed of two components consisting of the integrals along AC and DB, respectively, and we may write

$$\Pi_1(r, 0) = -\frac{\tau}{(1-\tau^2)r}[I(k_1) - I(k_2)], \tag{4}$$

where

$$I(k) = \int_{k/s}^{\infty is*} e^{isrw} d(w^2-1)^{-1/2}. \tag{5}$$

We integrate (5) by parts N times and find

$$I(k) = \left[-e^{isrw} \sum_{n=1}^{N} \frac{(-)^n}{(isr)^n} \frac{d^n}{dw^n}(w^2-1)^{-1/2} \right]_{k/s}^{\infty is*} + (-)^N \int_{k/s}^{\infty is*} \frac{e^{isrw}}{(isr)^N} \frac{d^{N+1}}{dw^{N+1}} (w^2-1)^{-1/2} dw.$$

The derivatives may be expressed in terms of Legendre polynomials by means of the relation

$$(-)^n \frac{d^n}{dw^n}(w^2-1)^{-1/2} = n!(w^2-1)^{-n/2-1/2} P_n\left(\frac{w}{\sqrt{w^2-1}}\right).$$

When the limits in the integrated portion are inserted and the definition of s used we see that

$$I(k_1) = \frac{k_2 e^{ik_1 r}}{k_1} \sum_{n=1}^{N} \left(\frac{k_2}{ik_1 sr}\right)^n n! P_n(k_2/s) + R_{1N}\frac{k_2}{k_1}, \tag{6}$$

where

$$R_{1N} = -\frac{k_1(N+1)!}{k_2(isr)^N} \int_{k_1/s}^{\infty is*} P_{N+1}\left[\frac{w}{\sqrt{w^2-1}}\right] \frac{e^{isrw}}{(w^2-1)^{N/2+1}} dw. \tag{7}$$

An inequality for R_{1N} may be obtained by using [3]

$$|P_{N+1}(t)| \leqslant |t + \sqrt{t^2-1}|^{N+1}$$

which holds for all values of t in the t plane cut from -1 to $+1$, if $\arg \sqrt{t^2-1} = 0$ when t is real and greater than $+1$. For then the absolute value of the Legendre polynomial in the integrand is seen to be less than $|(w+1)/(w-1)|^{(N+1)/2}$ when $R(w) > 0$, and R_{1N} may be compared with an integral having $|e^{isrw}|$ and powers of the factors $|w+1|$ and $|w-1|$ in the integrand. On the path AC we

[3] E. W. Hobson, loc. cit., p. 60.

have $|w - 1| \geqq |(k_1/s - 1)\sin\theta|$ where

$$\theta = \arg is^* - \arg\left(\frac{k_1}{s} - 1\right) = \frac{\pi}{2} - \arg(k_1 - s) > \frac{\pi}{2}\cdot$$

Similarly we have $|w + 1| \geqq |k_1/s + 1|$. These inequalities enable us to deal with the integral of $|e^{isrw}|$ which may be integrated to show that

$$|R_{1N}| < \left|\frac{(N+1)!e^{ik_1r}\sqrt{\csc\theta}}{[r(k_1 - s)\sin\theta]^{N+1}}\right|. \tag{8}$$

By interchanging k_1 and k_2 in (6) and (7) we obtain expressions for $I(k_2)$ and R_{2N}. An inequality for $|R_{2N}|$ is obtained from (8) by setting $\theta = \pi/2$ and interchanging k_1 and k_2. By combining these expressions in accordance with equation (4) we obtain an asymptotic expansion for $\Pi_1(r, 0)$.

In general, $I(k_2)$ is negligible in comparison with $I(k_1)$ because k_2 has a positive imaginary part which causes e^{ik_2r} to decrease rapidly. Since $I(k_2) = R_{20}$, R_{20} being the remainder after zero terms, we may obtain an inequality for $I(k_2)$ by setting $N = 0$, $\theta = \pi/2$, and interchanging k_1 and k_2 in (8). Then from (4) we have the result

$$\Pi_1(r, 0) = -\frac{1}{(1 - \tau^2)r}\left[e^{ik_1r}\sum_{n=1}^{N}\frac{n!P_n(k_2/s)}{(i\tau sr)^n} + R_{1N} - \tau^2 R_{20}\right], \tag{9}$$

where R_{1N} satisfies the inequality (8) and $|R_{20}| < |e^{ik_2r}/(rk_2 - rs)|$.

Series for $\Pi_1(r, 0)$ in Ascending Powers of r

Put

$$K(k_1) = e^{ik_1r} - i\frac{k_1sr}{k_2}\int_1^{k_1/s}\frac{e^{isrw}dw}{\sqrt{w^2 - 1}} \tag{10}$$

and define $K(k_2)$ as being obtained from (10) by interchanging k_1 and k_2. By referring to equation (3) we see that I may be written in the form

$$\Pi_1(r, 0) = \frac{1}{(1 - \tau^2)r}[K(k_1) - \tau^2 K(k_2)]. \tag{11}$$

We write

$$\begin{aligned} K(k_1) &= e^{ik_1r}\left[1 - \frac{ik_1sr}{k_2}\int_1^{k_1/s}\frac{e^{irs(w - k_1/s)}dw}{\sqrt{w^2 - 1}}\right] \\ &= e^{ik_1r}\left[1 - \frac{ik_1sr}{k_2}\sum_{n=1}^{\infty}\frac{(irs)^{n-1}}{(n-1)!}\int_1^{k_1/s}\frac{(w - k_1/s)^{n-1}dw}{\sqrt{w^2 - 1}}\right], \end{aligned} \tag{12}$$

the infinite series being uniformly convergent.

From Hobson's contour integral definition [3] of $P_n^m(t)$ it can be shown that if $R(m) < 1/2$

$$P^m_{-1/2}(t) = \sqrt{\frac{2}{\pi}} \frac{e^{-\pi i(m+1/2)}(t^2 - 1)^{m/2}}{\Gamma(1/2 - m)} \int_1^t \frac{(w - t)^{-m-1/2}}{\sqrt{w^2 - 1}} dw,$$

where $\arg(w - 1) = \varphi$, $\arg(w - t) = -\pi + \varphi$, where φ is the angle measured counter-clockwise from the positive direction of the real axis to the line directed from $w = 1$ to $w = t$. Setting $m = +1/2 - n$ where n is a positive integer we obtain

$$\int_1^t \frac{(w - t)^{n-1}}{\sqrt{w^2 - 1}} dw = (-)^{n-1}(n - 1)! \sqrt{\frac{\pi}{2}} (t^2 - 1)^{n/2-1/4} P^{1/2-n}_{-1/2}(t).$$

Thus Equation (12) becomes

$$\begin{aligned} K(k_1) &= e^{ik_1r}\left[1 + \sum_{n=1}^{\infty} \left(\frac{srk_1}{ik_2}\right)^n \cdot \sqrt{\frac{\pi k_1}{2k_2}} P^{1/2-n}_{-1/2}(k_1/s)\right] \\ &= e^{ik_1r} \sqrt{\frac{\pi k_1}{2k_2}} \sum_{n=0}^{\infty} \left(\frac{srk_1}{ik_2}\right)^n P^{1/2-n}_{-1/2}(k_1/s), \end{aligned} \tag{13}$$

where in passing from the first to the second line we have set $n = 0$ in [3]

$$P^{1/2-n}_{-1/2}(t) = \frac{1}{\Gamma(1/2 + n)} \left(\frac{t - 1}{t + 1}\right)^{n/2-1/4} F\left(1/2, 1/2; n + 1/2; \frac{1 - t}{2}\right),$$

and have summed the resulting series to show that $P^{1/2}_{-1/2}(k_1/s) = \sqrt{2/\pi\tau}$. The function $K(k_2)$ may be obtained from (13) by interchanging k_1 and k_2.

Combining (13) and (11), and using $\tau = k_1/k_2$ gives the convergent series for $\Pi_1(r, 0)$ given in the statement of results as equation (14).

Another Power Series for I

Here we obtain an expression for I somewhat similar to the one obtained in the previous section. The first step is to deform the contour joining the points A and B ($w = k_1/s$ and $w = k_2/s$). The deformation is carried out in two steps shown in Figs. 2*a* and 2*b*, respectively.

In Fig. 2(*a*) the contour joining A to B has been pulled around the point $+1$ and looped over itself. The point H is destined to move

[3] E. W. Hobson, loc. cit., p. 188.

over to B and G is to move over to A. This deformation of the contour does not alter the value of the integral as long as we pay attention to the arguments of $w - 1$ and $w + 1$. In Fig. 2(b) the deformation is almost completed; all that remains is for G to coincide with A and H to coincide with B.

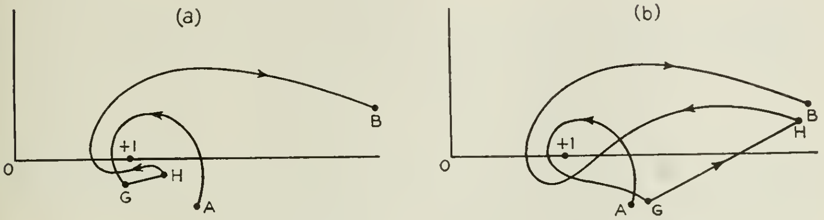


Fig. 2—Deformation of contour in w plane.

Using this deformation of the contour we may write equation (2) as follows:

$$\Pi_1(r, 0) = + \frac{\tau}{1 - \tau^2}\left[\int_A^G + \int_G^H + \int_H^B \frac{e^{isrw} w\,dw}{r(w^2 - 1)^{3/2}}\right]$$
$$= \frac{\tau}{(1 - \tau^2)r}\left[\int_{k_1/s}^{(1+)} - \int_{k_1/s}^{k_2/s} - \int_{k_2/s}^{(1+)} \frac{e^{isrw} w\,dw}{(w^2 - 1)^{3/2}}\right]$$

with the understanding that $\arg w - 1$ and $\arg w + 1$ have their principal values at the beginning of each integration. Upon referring to (2) we see that the middle integral is $-\Pi_1(r, 0)$ and hence

$$\Pi_1(r, 0) = \frac{\tau}{2r(1 - \tau^2)}\left[L(k_1) - L(k_2)\right], \tag{15}$$

where

$$L(k_1) = \int_{k_1/s}^{(1+)} \frac{e^{isrw} w\,dw}{(w^2 - 1)^{3/2}} = \sum_{n=0}^{\infty} \frac{(isr)^n}{n!} \int_{k_1/s}^{(1+)} \frac{w^{n+1}\,dw}{(w^2 - 1)^{3/2}} \tag{16}$$

and $L(k_2)$ is obtained from $L(k_1)$ by interchanging k_1 and k_2.

Let $w^2 - 1 = \tau^2(1 - t)$, or $sw = k_1\sqrt{1 - ts^2/k_2^2}$, then

$$\begin{aligned}\int_{k_1/s}^{(1+)} \frac{w^{n+1}\,dw}{(w^2 - 1)^{3/2}} &= -\frac{k_2}{2k_1}\left(\frac{k_1}{s}\right)^n \int_0^{(1+)} \frac{\left[(1 - (s^2t/k_2^2)\right]^{n/2}}{(1 - t)^{3/2}}\,dt \\ &= 2\frac{k_2}{k_1}\left(\frac{k_1}{s}\right)^n F(1, -n/2; 1/2; s^2/k_2^2),\end{aligned} \tag{17}$$

where it is understood that at the initial point of the contour

$\arg(1 - t) = 0$, $\arg(1 - (s^2/k_2^2)t) = 0$. This may be verified by expanding the numerator of the integrand and using

$$\int_0^{(1+)} t^m(1 - t)^\nu dt = (1 - e^{2\pi i\nu}) \frac{\Gamma(m + 1)\Gamma(\nu + 1)}{\Gamma(m + \nu + 2)},$$

where m is a positive integer or zero, with $\nu = -3/2$.

Expression (16) now becomes

$$L(k_1) = 2\frac{k_2}{k_1}\sum_{n=0}^{\infty} \frac{(ik_1r)^n}{n!} F(1, -n/2; 1/2; s^2/k_2^2) \tag{18}$$

and the series converges for all finite values of r since the series integrated termwise in Equation (16) is uniformly convergent.

We obtain the series for $\Pi_1(r, 0)$ given in statement of results as equation (19) by putting (18) and the corresponding expression for $L(k_2)$ in equation (15).

Notation

The following symbols are used. C.G.S. electromagnetic units are used throughout the paper.

c = velocity of light, 3×10^{10} cm./sec.

$F(a, b; c; x)$ = The hypergeometric function

$$1 + \frac{ab}{1!c}x + \frac{a(a+1)b(b+1)}{2!c(c+1)}x^2 + \cdots.$$

$J_0(\xi r)$ = Bessel function of the first kind, zero order.

$k_1 = \omega/c$.

$k_2 = \sqrt{\epsilon\omega^2 + i4\pi\sigma\omega}$. The real and imaginary parts are positive.

$P_n(t)$, $P_{-1/2}^{1/2-n}(t)$ = Legendre's polynomial, and associated Legendre's function of the first kind.

R_{1N}, R_{20} = Remainder terms in asymptotic series.

r = horizontal distance of representative point from dipole.

$s = k_1k_2/\sqrt{k_1^2 + k_2^2}$ or $1/s^2 = 1/k_1^2 + 1/k_2^2$. The real and imaginary parts of s are positive.

s^* = The complex conjugate of s.

t = time in the introduction, otherwise a complex variable.

w = complex variable.

z = height of representative point above ground.

ϵ = dielectric constant of the ground in e.m.u. The dielectric constant of air in e.m.u. is $1/c^2$. If the dielectric constant in e.s.u. is ϵ', then $\epsilon = \epsilon'/c^2$. The dielectric constant of air in e.s.u. is 1.

$\Pi_1(r, z)$ = Wave function for $z \geqq 0$ for a vertical unit dipole centered at the interface between air and ground. The wave function for a unit dipole wholly in air is obtained by multiplying the wave function given here by $2/(1 + \tau^2)$. By a unit dipole is meant the system obtained by letting the length l of a conductor approach zero while the current in the conductor approaches infinity in such a way that Il = unity, where the current equals the real part of $Ie^{-i\omega t}$ and does not vary with position along the conductor.

$\Pi_1(r, 0)$ = Value of wave function at earth's surface.

σ = conductivity of the ground in e.m.u. If the conductivity is σ' mhos per meter cube then $\sigma = 10^{-11}\sigma'$.

$\tau = k_1/k_2$.

ω = angular velocity, radians/sec.

Currents and Potentials along Leaky Ground-Return Conductors *

By E. D. SUNDE

THE problem of current and potential propagation for long conductors having large leakance to ground arises in connection with certain railway electrification problems, such as inductive effects in exposed communication lines, and voltages to ground of the tracks and of nearby underground cables. Impressed voltages in exposed lines are due partly to induction and partly to earth potential differences, and the latter in particular depend to a considerable extent on the mode of propagation of the track current. Voltages to ground of underground telephone cables depend on the mode of propagation along the cables of current produced in these by nearby railway electrification. These voltages may under certain conditions raise questions as to the possibility of hazard or, in case of direct current, give rise to electrolytic effects.

In considering propagation along earth-return conductors of the above kind, it is necessary to include certain effects which can be neglected in circuits having small leakance. One of the quantities involved in the differential equation for the current at a point of a ground-return conductor is the axial electric force produced in the ground adjacent to the conductor. For a given frequency and earth resistivity, this electric force at the point under consideration depends partly on the axial current distribution and partly on the leakage current distribution along the entire length of the conductor. The component depending on the axial current distribution is the vector potential multiplied by $-i\omega$, ω being the radian frequency, while the other component is the negative gradient of the earth potential, which depends on the leakage current distribution along the conductor. In the customary treatment, applying to conductors of small leakance, the axial current is assumed practically constant for great distances along the conductor, so that the first component becomes the negative product of axial conductor current at the point under consideration and the external earth-return impedance of the conductor. Furthermore, the earth potential is neglected or assumed constant along the conductor, so that the second component vanishes.

* Digest of a paper to be presented at the Winter Convention of the American Institute of Electrical Engineers, New York, January 25–29, 1937, and published in full in *Electrical Engineering*, Vol. 55, No. 12, pp. 1338–1346, December, 1936.

For conductors with large leakance, these simplifying assumptions are not justified. When the earth resistivity is uniform or varies with depth only, the electric force may be formulated in terms of the current along the entire length of the conductor, in which case the usual differential equation for the current is replaced by an integro-differential equation. A general solution of this equation and for conductor and earth potentials has been obtained. For homogeneous earth, rigorous as well as approximate solutions of special cases of interest in connection with the railway electrification problems mentioned above have also been derived.

One of these cases is of general interest since it may be regarded as fundamental to the solution of the general case of an arbitrary impressed electric force along the conductor. In this case a voltage $2V(0)$ is impressed across a break in the conductor at a certain point, which may be taken as the origin. The conductor current and potential are given by rather complicated integrals, which, in order to obtain practical formulas, may be expanded in series as:

$$I(x) = I_1(x) + I_2(x); \qquad I_2(x) = I_{21}(x) + I_{22}(x), \tag{1}$$

$$V(x) = V_1(x) + V_2(x); \qquad V_2(x) = V_{21}(x) + V_{22}(x), \tag{2}$$

where $I_{22}(x)$ and $V_{22}(x)$ may again be written as the sum of two terms, etc. The first terms in these expansions are:

$$I_1(x) = I_1(-x) = V(0)\frac{G(\Gamma)}{\Gamma}e^{-\Gamma x} = I(0)e^{-\Gamma x}, \tag{3}$$

$$V_1(x) = -V_1(-x) = V(0)e^{-\Gamma x} = I(0)\frac{\Gamma}{G(\Gamma)}e^{-\Gamma x}, \tag{4}$$

where $x \geq 0$ and:

$\Gamma = \sqrt{Z(\Gamma)G(\Gamma)}$.

$Z(\Gamma) = z + \frac{i\omega\nu}{2\pi}\log_\epsilon\frac{1.85\cdots}{a\alpha}\cdot$

$G(\Gamma) = \left[g^{-1} + \frac{\rho}{\pi}\log_\epsilon\frac{1.12\cdots}{a\Gamma}\right]^{-1}\cdot$

$\alpha = (i\omega\nu\rho^{-1} + \Gamma^2)^{1/2}$.

$\omega = 2\pi f$.

f = frequency in cycles per second.

x = distance from origin, in meters.

a = conductor radius, in meters.

z = internal impedance in ohms per meter.

g = leakage conductance in ohms per meter.

ρ = earth resistivity in meter-ohms.

$\nu = 4\pi\cdot 10^{-7}$ henries per meter.

The transcendental equation defining the propagation constant Γ may be solved by successive approximations; a convenient first approximation is $\Gamma = \Gamma_1 = \sqrt{gZ(0)}$, $Z(0)$ being the earth-return self-impedance of the conductor. Equations (3) and (4) are of the same form as the solution for conductors of small leakance, except that the propagation constant for the latter is taken equal to Γ_1 above. The effect of the earth potential appears in a first approximation as the second term in the expression for $G(\Gamma)$.

For earth resistivities within the usual range and for electric railway tracks or underground cables the two terms in the expression for $G(\Gamma)$ are frequently of the same order of magnitude. Appreciable errors may therefore be obtained by neglecting the second term, and in correlating the results of measurements this must be kept in mind.

The second terms in the expansions are given below, but may be neglected in the range of most practical applications.

$$I_{21}(x) = I_{21}(-x) = -I(0)\frac{G(\Gamma)\rho}{4\pi}\{(1+\Gamma x)e^{\Gamma x}Ei(\Gamma x) - (1-\Gamma x)e^{-\Gamma x}[Ei(-\Gamma x) + i\pi]\}, \quad (5)$$

$$V_{21}(x) = -V_{21}(-x) = I(0)\frac{\Gamma\rho}{4\pi}\{\Gamma x e^{\Gamma x}Ei(\Gamma x) - \Gamma x e^{-\Gamma x}[Ei(-\Gamma x) + i\pi] - 2\}, \quad (6)$$

where $Ei(u) = \int_u^\infty \frac{e^{-u}}{u}du$ is the exponential integral.

For sufficiently large values of Γx the bracket terms of expressions (5) and (6) vanish as $-8/(\Gamma x)^3$ and $4/(\Gamma x)^2$, respectively, so that in this case the second terms in the expansions predominate.

Abstracts of Technical Articles from Bell System Sources

Electricity in Gases.[1] KARL K. DARROW. The material in this paper was presented as the Joseph W. Richards Memorial Lecture delivered before the Electrochemical Society at Cincinnati, April 23, 1936. This lecture presents in a vividly descriptive manner some of the material published in past issues of the *Bell System Technical Journal.*

Electron Diffraction Experiments Upon Crystals of Galena.[2] L. H. GERMER. Cleaved surfaces of galena crystals yield electron diffraction patterns made up of Kikuchi lines, and spots which are drawn out into streaks by refraction. After etching, the spot pattern predominates and the individual spots are sharp. The lines are then rather diffuse and ill-defined. Rocking curves upon various Bragg reflections from the surface plane prove that the imperfection of a certain crystal does not exceed about 15 minutes, and that the projections through which the electrons pass are relatively thick. Estimates of imperfection and thickness made from rocking curves are in approximate agreement with those obtained from widths of Kikuchi lines.

A galena crystal which has been filed or ground parallel to a cube face exhibits two different sorts of surfaces. There are smooth "mirror" surfaces from which large blocks of the crystal have been mechanically torn, and there are very deeply scratched portions of the surface. The "mirror" surfaces give diffraction patterns which are qualitatively similar to patterns from cleaved surfaces, although there are notable differences. From mirror surfaces produced by filing, Kikuchi lines are very diffuse or are entirely missing, and diffraction spots form an extended array. The diffuseness of the lines and the extent of the array of spots correspond to great crystal imperfection, or to exceedingly thin projections. Reasons are advanced for believing in imperfection rather than extreme thinness.

The deeply scratched portions of the surface of a galena crystal give diffraction patterns which are entirely unlike patterns from cleaved surfaces. Before etching, Debye-Scherrer rings are produced. After a light or moderate etch a complex pattern appears, the nature of which is related to the angle between primary beam and direction of filing.

[1] *Transactions Electrochemical Society*, Vol. LX1X, 1936.
[2] *Phys. Rev.*, October 1, 1936.

The pattern is that of a mass of minute crystallites which have been rotated about an axis in the surface normal to the direction of filing, and in the sense determined by imaginary rollers which would be turned by slipping on the (0 1 0) plane. The magnitude of the rotation varies for different crystallites over a range from 5 to about 35 degrees. By alternate etching and examination by electron diffraction it is found that this layer of rotated crystallites extends beneath the surface to a depth of 0.003 mm.

Rotation of crystallites accompanying slip along slip planes is the mechanism reported to account for strain hardening in metals. This same rotation is observed in the present experiments on galena. It seems altogether possible that the simple technique of these experiments can be applied directly to study the disturbance in surface layers of metal crystals produced by abrasion. It may thus be a useful way of studying strain hardening in metals.

The Photoelectric Cell and Its Method of Operation.[3] M. F. Jamieson, T. E. Shea, and P. H. Pierce. This paper gives a simple description of the laws governing the release of electrons from photoelectric surfaces, their collection at anodes, and the creation of ions in photoelectric cell gases by the "ionization" process, and discusses questions of spectral selectivity of various photoelectric surfaces, the influence of spectral characteristics of illumination, and the dynamic characteristics of vacuum and gas-filled cells.

Modified Sommerfeld's Integral and Its Applications.[4] S. A. Schelkunoff. The purpose of this paper is to obtain a certain integral expressing the fundamental wave function and with the aid of this integral to calculate the radiation resistances of small doublets and small loops placed inside an infinite hollow cylinder. Some applications of this integral to calculation of radiation from parallel wires in free space are also discussed.

Diffusion of Water Through Insulating Materials.[5] R. L. Taylor, D. B. Herrmann, and A. R. Kemp. Data are presented on the rate of water diffusion through various organic materials. A diffusion constant based on Fick's linear diffusion law is calculated for each material. Several equations are derived from Fick's law to show how valuable information can be obtained in connection with practical problems.

[3] *Jour. S. M. P. E.*, October, 1936.
[4] *Proc. I. R. E.*, October, 1936.
[5] *Indus. and Engg. Chem.*, November, 1936.

The effect of variations in methods and conditions of test is studied. The rate of diffusion through a water-sorbing material such as rubber does not obey Fick's law when under diffusion conditions favoring high water sorption.

Various concepts involving sorption and diffusion processes are discussed as bearing upon the mechanism of the diffusion of water through organic substances.

Contributors to this Issue

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.

R. F. Davis, B.E.E., Cornell University, 1921. American Telephone and Telegraph Company, Department of Operation and Engineering, 1921–. Mr. Davis' work has been largely concerned with the electrical protection of communications circuits and with the electrical coordination of such circuits with power transmission and distribution circuits.

S. O. Rice, B.S. in Electrical Engineering, Oregon State College, 1929; California Institute of Technology, 1929–30, 1934–35; Bell Telephone Laboratories, 1930–. Mr. Rice has been concerned with various theoretical investigations relating to telephone transmission 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.

Nelson E. Sowers, B.S. in Engineering Physics, 1924, University of Illinois; M.A., Columbia University, 1927; Engineer-Physicist (Professional), University of Illinois, 1936. Engineering Department, Western Electric Company, 1924–25. Bell Telephone Laboratories, Inc., 1925–. Since 1931, Mr. Sowers has been engaged in studies pertaining to amplifiers for ultra-high radio frequencies.

M. E. Strieby, A.B., Colorado College, 1914; B.S., Harvard, 1916; B.S. in E.E., Massachusetts Institute of Technology, 1916; New York

Telephone Company, Engineering Department, 1916–17; Captain, Signal Corps, U. S. Army, A. E. F., 1917–19. American Telephone and Telegraph Company, Department of Development and Research, 1919–29; Bell Telephone Laboratories, 1929–. Mr. Strieby has been associated with various phases of transmission work, more particularly with the development of long toll circuits. At the present time, in his capacity as Carrier Transmission Research Engineer, he directs studies of new and improved methods of carrier frequency transmission over existing or new facilities.

Erling D. Sunde, E. E., Technische Hochschule Darmstadt, 1926. American Telephone and Telegraph Company, Department of Development and Research, 1927–34; Bell Telephone Laboratories, 1934–. Mr. Sunde's work has been mainly concerned with inductive effects of electric railways.

W. Howard Wise, B.S., Montana State College, 1921; M.A., University of Oregon, 1923; Ph.D., California Institute of Technology, 1926. American Telephone and Telegraph Company, Department of Development and Research, 1926–34; Bell Telephone Laboratories, 1934–. Dr. Wise has been engaged in various theoretical investigations relating to transmission theory and telegraphy.

VOLUME XVI APRIL, 1937 NUMBER 2

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

The Bell System Technical Journal

Vol. XVI *April, 1937* *No. 2*

Recent Trends in Toll Transmission in the United States *

By EDWIN H. COLPITTS

YOUR country is advancing industrially and commercially with tremendous strides. Adequate telephone communication is of such great importance under these conditions that I felt a general statement as to methods in process of being applied to the plant of the Bell System would be interesting and possibly helpful. I am fully aware that, in some or even many respects, your problems will differ from ours. Much of what I have presented may, therefore, serve only to suggest research and development to meet your own requirements. Perhaps also, in some small way, this statement of progress in communication may stimulate research and development in other industries and services.

In the year 1885, only nine years after the telephone was invented, a Telephone Company was chartered in the United States for the following purpose: "to connect one or more points in each and every city, town, or place in the State of New York with one or more points in each and every other city, town or place in said State and in each and every other of the United States and in Canada and Mexico; and each and every of said cities, towns, and places is to be connected with each and every other city, town, and place in said states and countries, and also by cable and other appropriate means with the rest of the known world."

This was an ambitious program, and thirty years passed before the results of research and development could be embodied in apparatus and equipment to make it possible to talk between cities on the Atlantic and Pacific coasts of the United States, and about forty years passed before the establishment of telephone service between America and Europe. Telephone service was later extended to other parts of the world including your country, and it is now possible for a telephone subscriber in the United States to converse with a person at any one of thirty-four odd million subscriber stations in a large number of countries of the world. Further, it is possible to talk with persons on suitably equipped ships at sea. Expanding still further beyond the goal set in 1885, and departing from the idea of two-way conversation,

* One of three Iwadare Foundation lectures delivered during the past month in Japan by Dr. Colpitts.

this same company provides a portion of the facilities which make it possible for a person speaking at one place almost anywhere in the civilized world to have his voice heard at almost any other. I refer to broadcasting.

Figure 1 shows a wire map of a few of the principal toll lines in the United States. This toll plant affords facilities that, in connection with the local plant, enable any telephone subscriber at any point to communicate promptly with a subscriber at any other point in the United States, Canada, or Mexico.

With the growth of radio broadcasting, a service with which you are all familiar, it became necessary to provide circuits to interconnect radio broadcasting stations. Figure 2 shows a wire map of such interconnecting circuits commonly spoken of as "program circuits." Figure 3 shows schematically the radio-telephone circuits that connect the United States to foreign countries. Another extension of the service rendered by the Bell System is indicated by Fig. 4, which shows a wire map of circuits devoted to the telephotograph service.

It is not my purpose to take your time to discuss these past developments, since they have already been described quite fully in technical literature, which I know is available to you. I propose rather to discuss some of the more recent trends in toll circuit development in the United States, but the subject is so large that I can touch only upon the more salient factors, indicating to you the direction in which the art is moving.

This new art, or perhaps more accurately this extension of an older art, utilizes the results of continuing researches on vacuum tubes and their uses as amplifiers, modulators, and oscillators, on filters as a means of splitting up broad bands of frequencies into the relatively narrow bands required for telephony or the still narrower bands required for telegraphy, and on methods of electrically isolating a particular circuit so as to avoid crosstalk and noise. These are not all the factors requiring research, but are merely some of the more important ones.

In this connection, it should be emphasized that these new systems or methods are still under development, and that their development for commercial application will require continued effort over a considerable period. We have come to group these new systems or methods under the term "high-frequency broad-band wire transmission." Instead of confining ourselves to a frequency range extending to about 30,000 cycles, as used for our present carrier systems, we are setting for our objective the transmission and utilization of bands of frequencies a million or more cycles wide in the case of

Fig. 1—Principal toll lines of the United States.

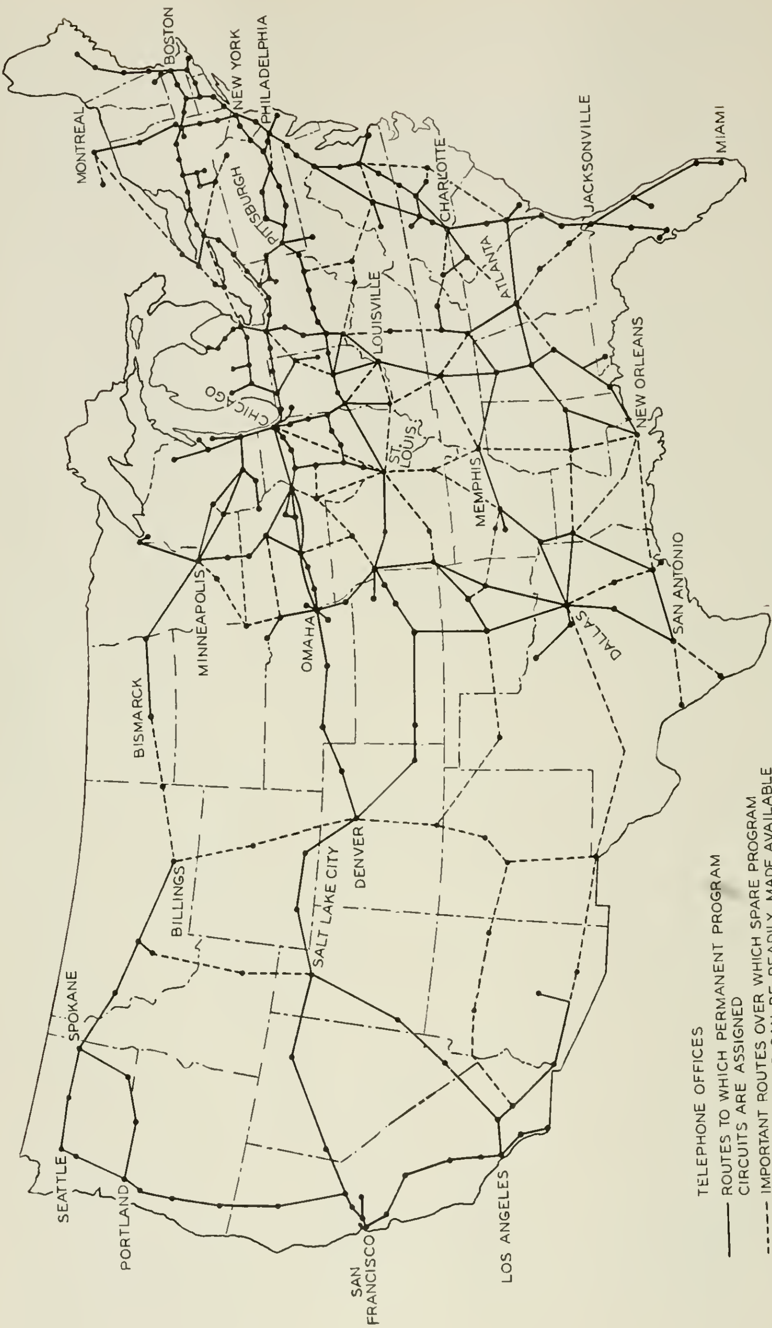


Fig. 2—Permanent and spare program circuits of the United States.

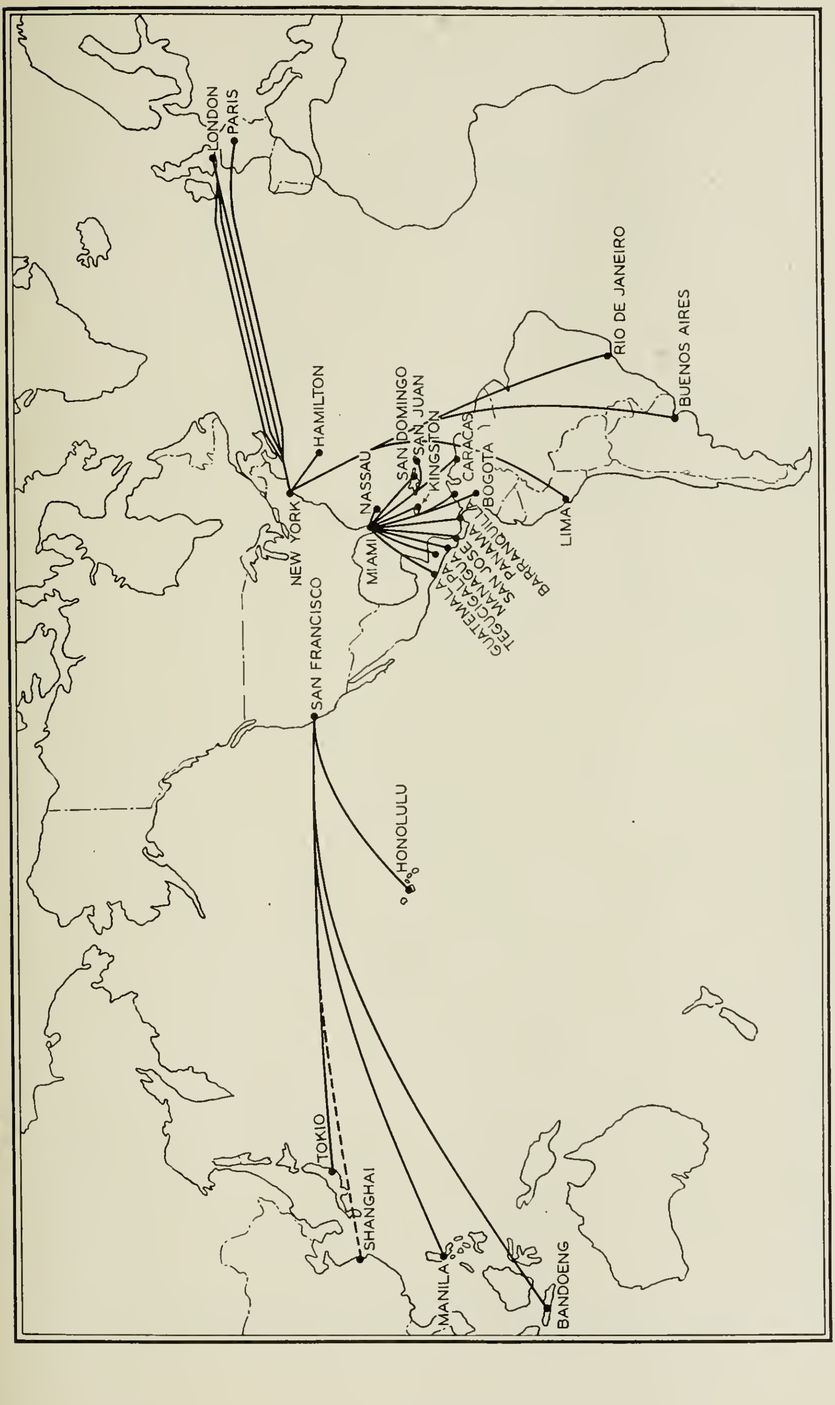


Fig. 3—Transoceanic radio telephone circuits operated by the Bell System.

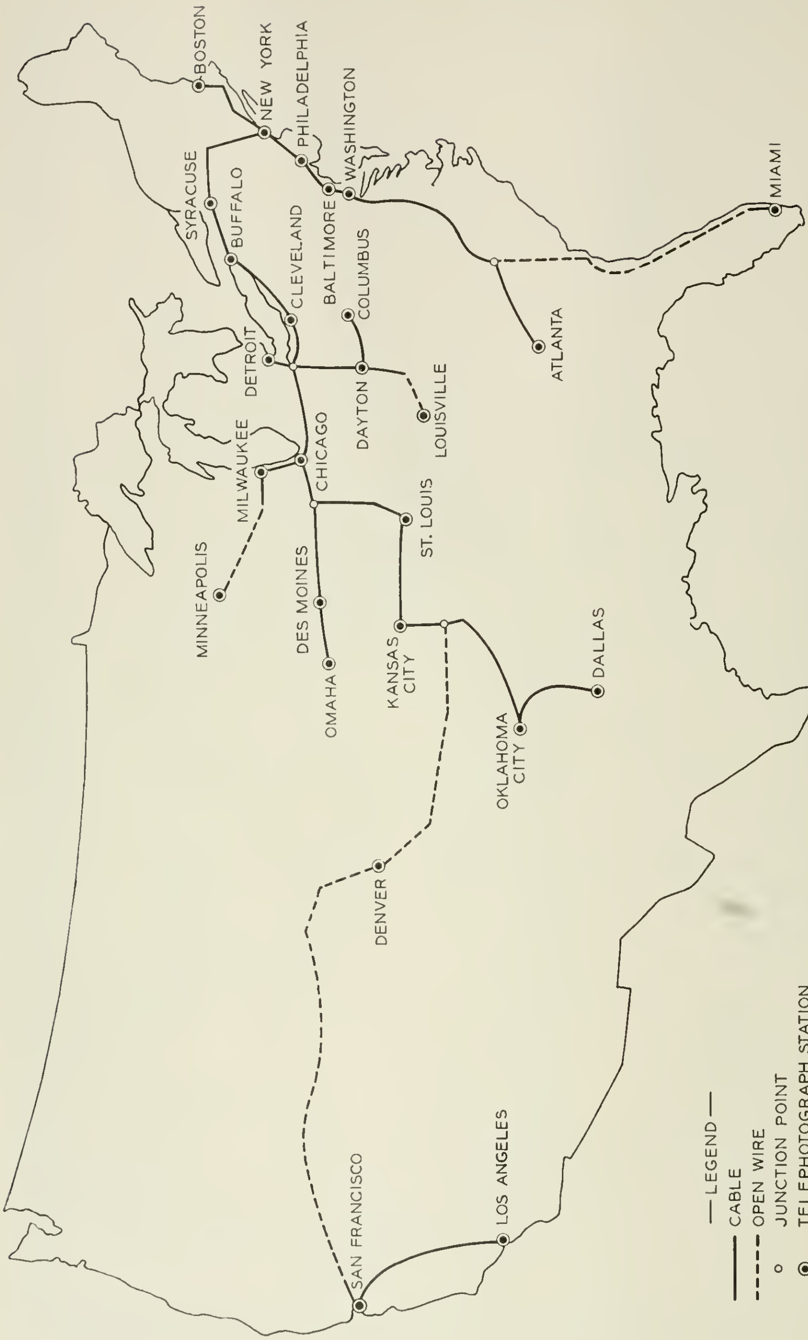


Fig. 4—Telephotograph circuits of the United States.

certain line structures. For other structures, the frequency range is narrower, but for all these systems the frequency range is transmitted as a single band, and split into communication channels for telephone or telegraph only at the terminals. If the transmission of television signals should become necessary, a very broad band—one or more million cycles—would, of course, be required. Although I have referred to television, our primary interest in broad-band wire transmission is for telephone transmission purposes, where the wide transmission band can be used to give a large number of talking channels.

You will recall that the idea of deriving more than one communication channel from a single pair of conductors, by what we now call carrier methods, is old—in fact, as old as the telephone itself. Until quite recently, however, physical devices and methods have not been available to make the carrier method utilizable in practice. Beginning about fifteen years ago, the Bell System began to install carrier systems, and since that time this method has had continued growth on open-wire lines, with the result that a substantial amount of toll traffic is now carried over carrier systems on open-wire lines. A relatively simple form of carrier equipment provides one two-way telephone circuit in addition to the usual voice frequency circuit, while more elaborate and refined equipment adds three two-way telephone circuits.

In addition to the economic urge to obtain the largest possible number of telephone channels over a given pair of wires, there is an additional factor that has influenced the development of broad-band systems, and that is, the speed of transmission. Even in the lowest-speed telephone circuits, the speed of transmission of voice waves, as judged by ordinary standards, is high, being from ten to twenty thousand miles per second (32,000 km. per sec.) in the loaded cable circuits that are now used for many of the long toll lines. For ordinary distances, moreover, the speed of transmission is not of any particular moment, but when the voice must be transmitted over distances of thousands of miles, it becomes important. Echo effects become exaggerated, and in a long connection, the actual time for speech to reach from the first subscriber to the second subscriber, added to the time required for the second subscriber to answer the first subscriber, may become an annoying factor. The broad-band transmission method furnishes circuits, however, in which the speed of transmission is raised from about 20,000 miles per second, as on loaded cable circuits, to a speed approaching that of light.

In developing a new toll system, there are many other factors, of course, that must be considered. In our case, just as in yours, there is first, an existing toll telephone plant, which must be utilized to the maximum advantage. Also, distances between toll offices or toll centers vary, and particularly the number of circuits required between given toll centers varies over a wide range. It follows, therefore, that there is no one type of construction or method which can be economically utilized in all situations. Figure 5, for example, showing a pole line carrying open-wire circuits and circuits in cable, illustrates some of our present methods.

The high-frequency broad-band transmission development is being proposed for three uses: (1) for application on telephone toll cables already in existence, or on future toll cables of very similar type of

Fig. 5—A typical pole line carrying both open wire and cable.

construction; (2) for an extension to higher frequencies on open-wire telephone circuits, so as to secure more telephone channels on a given pair; (3) for application to new types of conductors capable of transmitting a very wide frequency band, such as the "coaxial" conductor, now being tried experimentally.

I need hardly point out to you that as the frequency of transmission is raised, the attenuation or line loss is greatly increased. This is due more particularly to two factors: an increase in series resistance due to skin effect, and an increase in shunt conductance due to increased dielectric losses. As the frequency increases, the currents transmitted tend more and more to avoid the inner parts of the conductor and to concentrate on the surface, so increasing the effective series resistance. Dielectric losses in the insulation between the conductors also increase with the frequency, and so increase the effective shunt conductance. Figure 6 shows graphically the increase in attenuation with frequency,

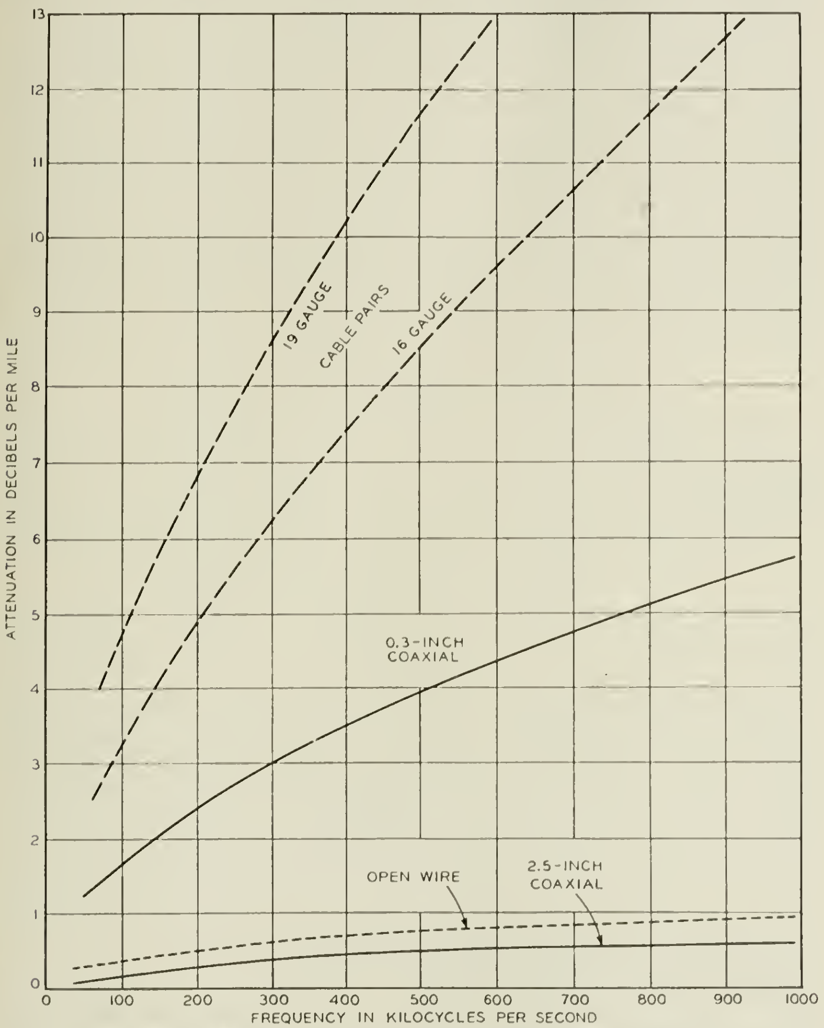


Fig. 6—Attenuation-frequency characteristics of various types of telephone circuits.

and also shows the relative attenuations of various types of construction.

Until the vacuum tube amplifier became available, the only practical method of overcoming high attenuation in a given type of line construction was to provide larger conductors. With the development of the vacuum tube, high amplification became available as an alternative method. Since increasing the size of the conductor in order to decrease attentuation involves large expense, we are naturally led to consider the use of as much amplification as possible.

Before referring further to the utilization of high amplification, I wish to point out that at the present time for distances greater than about 150 miles (240 km.) in cable, we utilize the so-called 4-wire method to obtain two-way telephone transmission; that is, transmission in one direction is carried by one pair of wires, and transmission in the other direction is carried by a second pair of wires. What is in effect the same method is employed in our present carrier systems, transmission in one direction being superposed on one frequency, and transmission in the other direction being superposed on a different frequency.

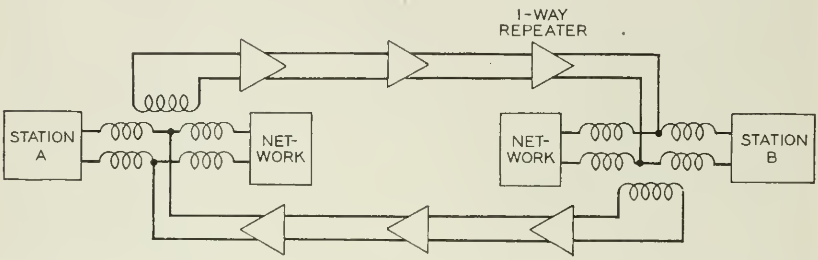


Fig. 7—Block schematic of a four-wire circuit showing two two-wire circuits with one-way repeaters.

Figure 7 shows diagrammatically a 4-wire telephone circuit in cable. You will note that one-way amplifiers are introduced in each pair of wires at points which in present practice are about fifty miles (80 km.) apart. The question naturally arises: Why not increase the distance between amplifiers and at the same time increase their amplification, and so reduce the cost? The answer is that two sources of noise disturbance have to be considered: first, induction from neighboring circuits; second, the noise of thermal agitation.

The line circuit, depending in degree upon the type of construction, receives unwanted interference from the outside, such as induced currents from power lines, lightning, and crosstalk from adjacent circuits, and it is not possible, as a result, to allow the transmitted speech signals to be attenuated below a certain level with respect to such noise interference. As a consequence, the amount which we can allow a speech signal to be attenuated before it reaches an amplifier, depends on how completely the transmission system is free from external interference.

Two methods are commonly employed to minimize external interference: shielding, and a geometrical arrangement of the conductors of the circuit so as to balance out certain forms of interference. The open-wire line, which has no shielding, depends wholly on the symmetry of its conductors and the transpositions to balance out interference. Conductors surrounded by metal sheath, such as cable pairs, are less subject to interference than open-wire lines. The conductors of a pair are close together, are well transposed by twisting, and are shielded to a certain extent by the outside lead covering. As a result of this, the noise due to outside sources has a low level, and the telephone speech currents can be permitted to become attenuated to a relatively low value before reaching an amplifier where they are stepped up to their original value. It is evident, of course, that such cable pairs, being made up of small conductors close together and having paper dielectric, have correspondingly high attenuations.

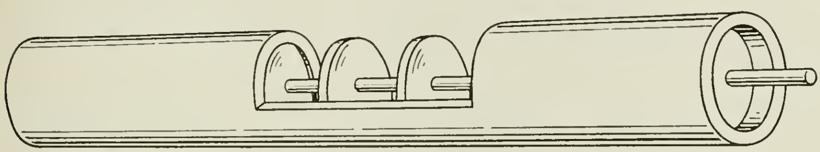


Fig. 8—Diagrammatic representation of the coaxial structure.

The ideal conductor would be one for which the attenuation over the whole operative frequency range was not too great, and at the same time one completely shielded from the influence of external electric or magnetic fields. The so-called coaxial conductor approximates to these requirements. This conductor transmits efficiently over a wide frequency band, and at the same time is well shielded from external influences, the degree of shielding being higher at the higher frequencies where greater amplification is needed to overcome the greater attenuation.

The coaxial conductor upon which we are experimenting consists of an inner wire and outer tube separated by spacing insulators. It is desirable to separate the two conductors by a minimum amount of solid insulation to the end that the dielectric will be largely air, and the losses at high frequencies be at a minimum. Figure 8 shows diagrammatically a coaxial structure. At high frequencies the current travels chiefly on the outside of the inner conductor, and on the inside of the outer conductor. It will be obvious to you that there is a wide latitude of choice in the dimensions of coaxial structures.

Although we are experimenting with a structure capable of transmitting a band of one or two million cycles in width, a coaxial structure capable of transmitting a band twenty million cycles in width is apparently not wholly unfeasible.

The question naturally arises whether, with interference from outside sources almost wholly or entirely eliminated, it is possible to allow speech currents to be attenuated to an unlimited degree before the introduction of amplification to bring them back to their original value. With all outside interference eliminated, however, noise arising within the conductor itself sets the limit. This interference is termed resistance noise or sometimes thermal-agitation noise because it is a function of the temperature of the conductor. It is apparently due to the continual moving around of the free electrons which exist in all conductors. Our Laboratories have investigated this phenomenon and determined its characteristics. This resistance noise varies in amount with the resistance of the conductor and with the temperature. It is uniformly distributed over the whole frequency range from lower voice frequencies up to the highest frequency which we have considered using. One ready means of observing this phenomenon is to provide an amplifier covering the voice range, with its input connected across the resistance, and to listen on the output of the amplifier with a telephone receiver. If the amplifier has an amplification of about 140 db, the noise heard in the telephone receiver is about as loud as would be heard in the receiver were it connected directly across the output from a telephone substation.

To prevent this thermal or resistance noise from being noticeable in a telephone conversation, we must limit the amount of amplification used at any one point in a long system, even though it were perfectly shielded, to an amount considerably less than 140 db. These considerations have led us to conclude that for a long circuit with many amplifiers distributed along the route, the amount of amplification at any one point should not exceed about 70 db.

The amount of amplification involved in present-day telephone circuits is illustrated by the 4-wire cable circuit, in which amplifiers are located in each pair at intervals of about 50 miles (80 kilometers) and each amplifier is set to give an amplification of about 25 db, or a power amplification of 300 times. For such a circuit between, say, New York and Chicago, a distance of about 900 miles (1450 km.), the total amplification is about 500 db, or a power amplification of 10^{50}. It is obviously necessary that these amplifiers must be made very *stable* so that the cumulative variations in the many amplifiers may not make it impossible to obtain the required degree of overall stability of transmission. The total amplification is affected by the requirement that the New York-Chicago circuit is expected to have a net attenuation of not over 9 db, and to be stable within about ± 2 db.

These figures may seem large and the requirements difficult to meet, but with the systems under development, the magnitude of the high-amplification problem is even greater. In the carrier on cable development, the circuits will consist of non-loaded pairs, and it will be necessary to so space the amplifiers and adjust their amplification that the total amplification on, for example, a New York-Chicago circuit will be about 3000 db at the center of the frequency band, or a power ratio of 10^{300}. Obviously, the stability requirement has been made much more rigid. With a typical coaxial circuit, the overall amplification at a million cycles for a thousand-mile circuit (1600 km.) may well be 6500 db or a power ratio of 10^{650}.

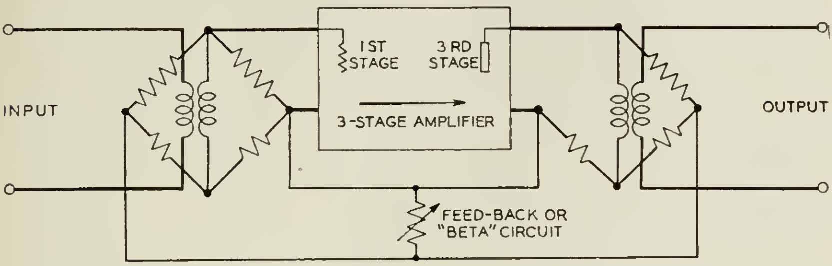


Fig. 9—Simplified schematic diagram of a feedback amplifier.

Furthermore, with the relatively simple circuit shown in Fig. 7, the amplifiers are called upon to handle merely the currents corresponding to one telephone conversation, while in the broad-band system an amplifier is required to handle simultaneously a large number of carrier telephone channels. To avoid the generation of extraneous frequencies or intermodulation products which would cause interference between the channels, an amplifier must be adopted which is more nearly perfect in this respect than any heretofore standard.

This problem of amplifier stability and perfection was solved some little time ago by an invention of one of our engineers. This engineer devised a new amplifier circuit which has been termed the "stabilized feedback amplifier." Some older types of amplifiers took some of their output and fed it back to the input for the purpose of *increasing* the amplification. This new feedback circuit controls the phases of the currents in the amplifier and feedback circuits so that the amplification is *decreased*. As a result, we have available an amplifier which is remarkably stable and closely *linear* in its performance. Figure 9 shows schematically the circuit of such an amplifier, but the actual detailed design is not simple and involves great technical skill.

Since the high amplifications just discussed are employed to offset the equally high attenuation of the line structures, careful consideration must be given to the stability or constancy of the attenuation caused by the line structure. The fact is that as the temperature changes, the attenuation of any line structure varies correspondingly. If the line structure is underground in cable, the temperature changes are slow in rate and the variations in line attenuation correspondingly slow, but if the structure is in aerial cable or consists of open wire, not only do we have variations in temperature with the season of the year, but large daily variations as well. With an aerial cable, for example, the change in loss of 19-gauge B & S non-loaded pairs throughout the year for a circuit 1,000 miles (1600 km.) in length amounts to approximately 500 db in the frequency band we propose to use. For our long cable circuits operated on voice frequencies, we have developed automatic regulating means, so that amplification is varied to compensate for changes in attenuation. With the much higher attenuations and equally higher amplifications involved in broad-band systems, more refined methods of compensating for temperature changes are under development. This is a very serious problem and sets one limit to the use of such systems.

I spoke of the new type of amplifier, employing negative feedback, which became available at a most fortunate time. It is almost equally fortunate that, due to continued research and development, new and simpler forms of electric wave filters became available. Time does not permit me to go into details, but in these newer types of electric filters suitably cut quartz crystals are utilized. Developments have also made it possible to use inductance coils with iron cores. As a result of these two changes the filter structure is simplified and its size reduced.

Fundamental to the whole broad-band transmission development, there are many other elements which have required much research and development effort such, for example, as modulators and demodulators, but I shall be able to discuss only the more striking problems underlying all broad-band systems as they pertain to certain specific applications.

Cable Carrier Systems

We plan as a first step to apply the cable carrier system to pairs in existing cables. These cables were designed and manufactured with the expectation that they would be required to transmit frequencies only up to a few thousand cycles, that is, these cables were not designed to meet crosstalk requirements at high frequencies. Crosstalk between pairs in a cable arises from a lack of geometrical symmetry between each pair and every other. As a result of very careful research, we have developed a method of connecting small mutual inductance coils, or condensers, or both, between all the pairs concerned, and in this way reducing crosstalk to a satisfactory degree. At present coils alone are being used. Figure 10 shows schematically the balancing method.

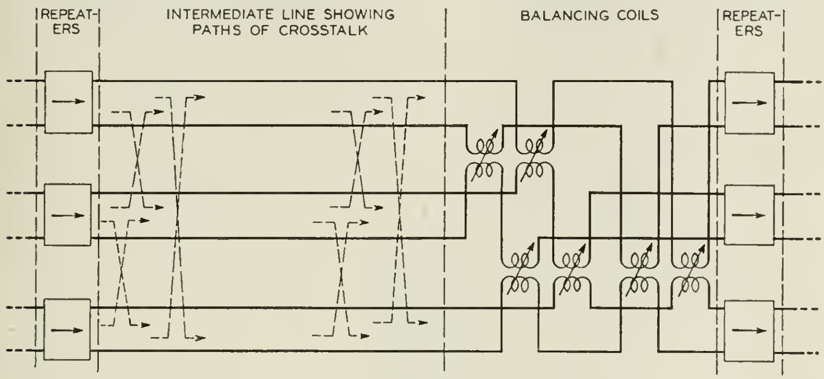


Fig. 10—Diagrammatic representation of method of balancing out cross-talk.

It is obvious that in this process of balancing out the crosstalk, the value given to the adjusting coil or condenser must be determined for each particular unit involved. When the balancing units have been installed and once adjusted, however, we feel that they will remain permanent. Present indications are that by adopting this procedure, we can employ frequency ranges up to 60,000 cycles upon our toll cables, and this will permit us to secure 12 one-way channels on each pair of conductors. With at least our present type of cables, we anticipate that it will be necessary to use separate cables in opposite directions to avoid the additional crosstalk that arises when adjacent pairs are used to transmit in opposite directions. Referring to our present cables, if the pairs which it is desired to utilize are loaded, it is necessary first to remove the loading coils. Repeaters will be spaced about 17 miles (27 km.) apart. Since the present repeater points on cables are about 50 miles (80 km.) apart, it will be necessary to add on the average two new repeater points between existing repeaters. The total amount of apparatus at these new repeater points, however, does not bulk very large, because a repeater handling 12 channels will be no larger than the older type voice repeaters

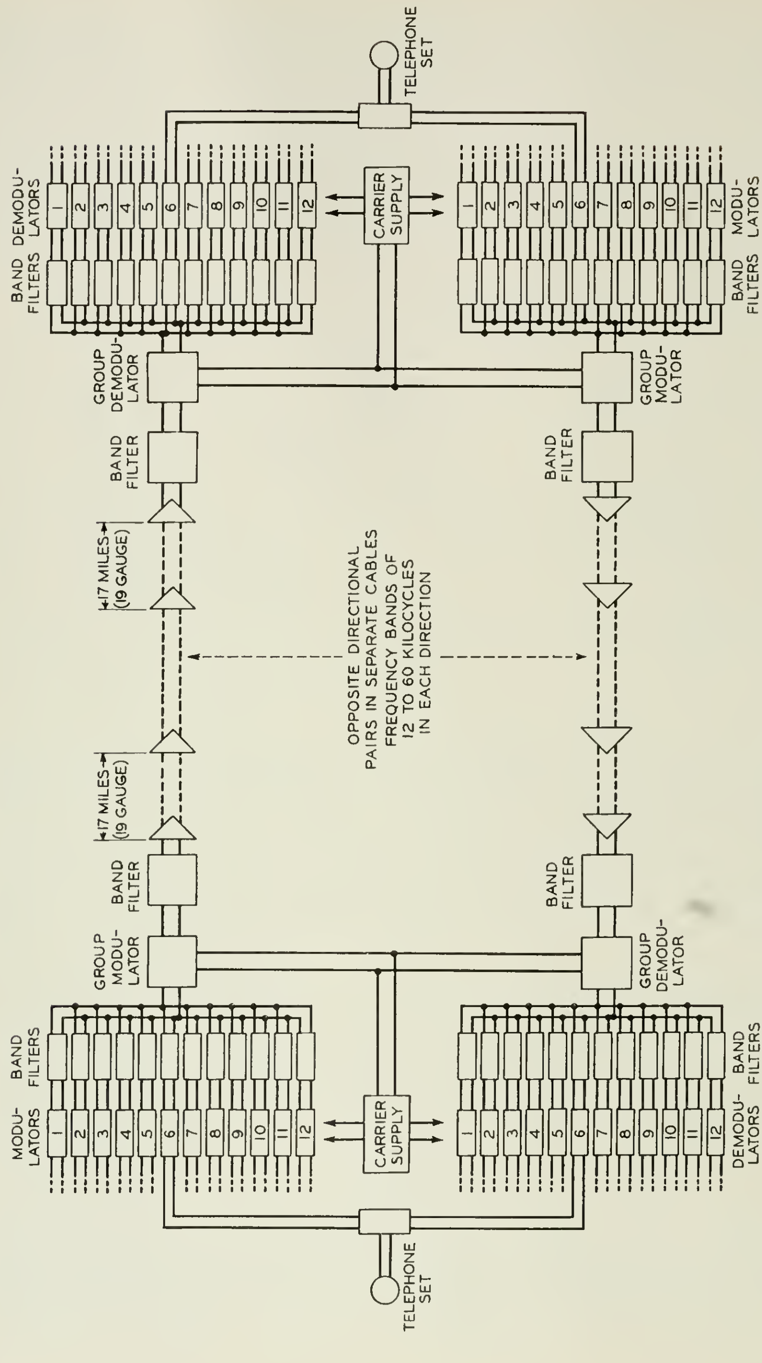


Fig. 11—Schematic representation of the broad-band cable carrier system, which supplies 12 channels in each direction.

handling only one channel. The present plans call for these intermediate repeater stations to be substantially non-attended. The electrical units employed in a cable carrier system are shown schematically in Fig. 11.

Some idea of the possibilities of this carrier cable system may be formed from the results of a trial installation made on a laboratory-scale. In one case, we had circuits as long as 7,500 miles (12,000 km.) set up, over which we carried on satisfactory conversations. The total attenuation over some of these circuits was such as to require power amplifications of 12,000 db, which corresponds to a power ratio of 10^{1200} to 1. This amplification was applied at nearly 400 points.

Broad-Band System for Open-Wire Lines

In the Bell System, as you know, along many toll routes, there is still much open-wire construction, aggregating tens of thousands of miles. At the present time, many thousands of miles of open-wire lines are equipped with 3-channel two-way carrier systems. These systems employ frequencies up to about 30,000 cycles, and with the regular voice frequency circuit provide facilities for four simultaneous conversations over one pair of wires. This might appear to be an efficient use of the wire plant, but the proposed system employs an additional frequency range from about 30,000 cycles to perhaps 150,000 cycles, adding 12 channels in each direction to a pair of wires. This will furnish a total of sixteen simultaneous conversations over a pair of wires.

Extending the frequency range accentuates the problem of crosstalk and some of the other problems of interference, but it is our present feeling that a substantial number of the pairs on a suitably constructed pole line can be rearranged for operation by this broad-band method. Figure 12 shows schematically the arrangement of apparatus at a terminal to provide these sixteen talking circuits, and Fig. 13 shows diagrammatically the arrangement of apparatus at a repeater station.

Since increased frequency means higher line attenuations with corresponding higher amplification, the use of higher frequencies means additional repeaters on the line, so that the line currents will not at any point be attenuated below a certain level. The present proposal is to provide repeater spacing of approximately 75 miles, instead of the 150 miles used on the present open-wire carrier systems.

Coaxial Cable System

This is the most radical of the broad-band developments that we have attempted to develop practically. Instead of several pairs

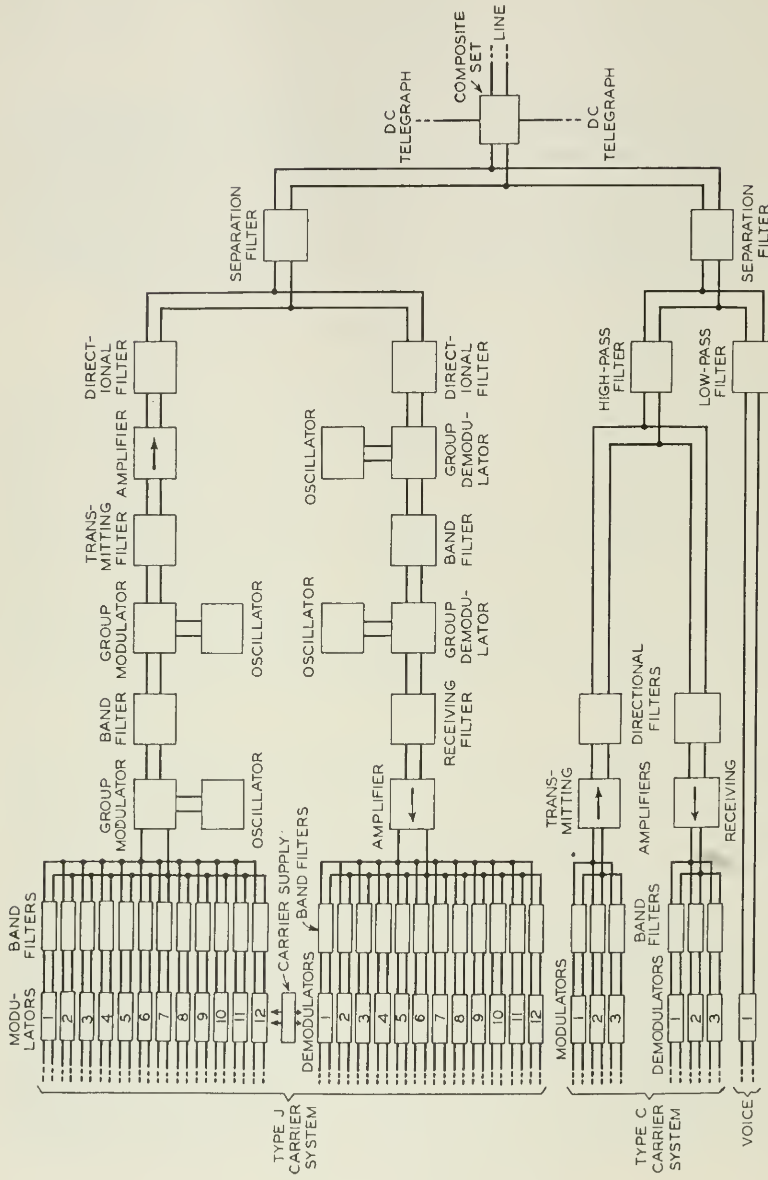


Fig. 12—The broad-band open-wire circuit superimposes twelve two-way circuits on a single pair of wires that also carries three two-way carrier channels at lower frequencies and a two-way voice circuit.

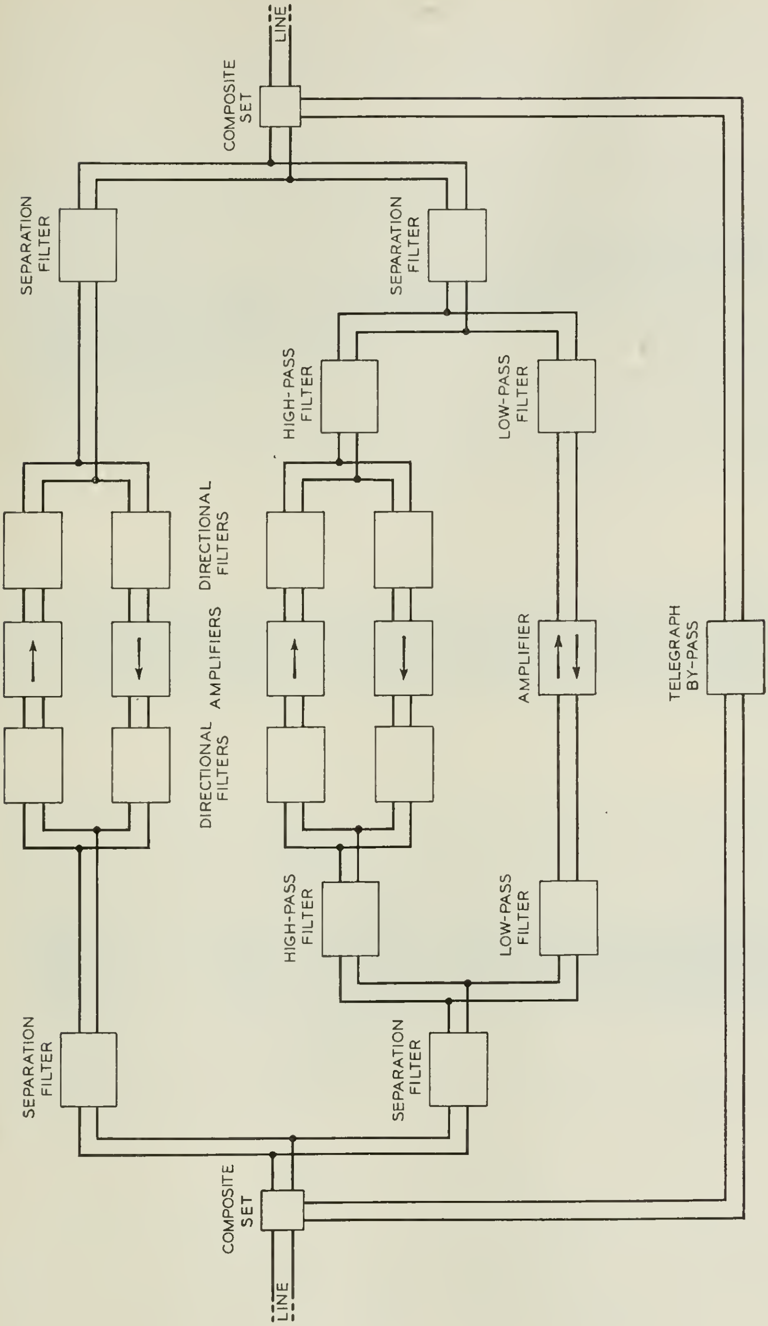


Fig. 13—Block schematic of a repeater station for a broad-band open-wire circuit.

carrying what are now considered moderately high frequencies, this new system employs a single circuit carrying a very broad frequency band. There are many ways in which conducting systems of this type may be constructed. In this connection, of course, facility of

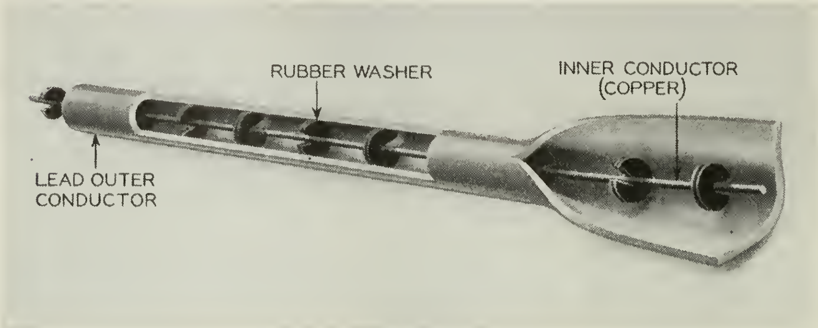


Fig. 14—One of the experimental coaxial structures.

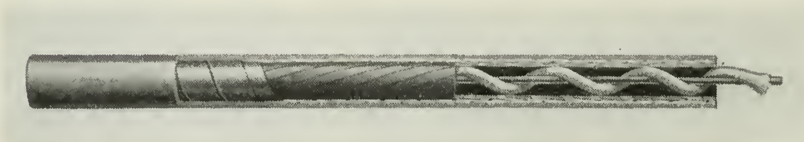


Fig. 15—Another experimental coaxial structure.

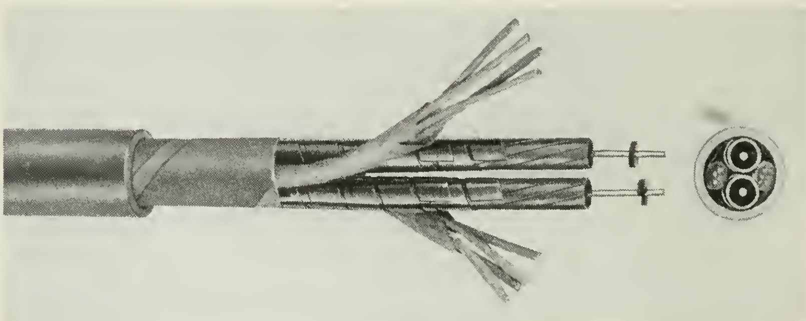


Fig. 16—The coaxial cable employed on the experimental installation between New York and Philadelphia.

manufacture and ability to withstand the handling incidental to installation, must be considered as well as electrical performance. Some of the experimental types of structure are shown in Figs. 14 and 15. In a field trial between New York and Philadelphia, over a

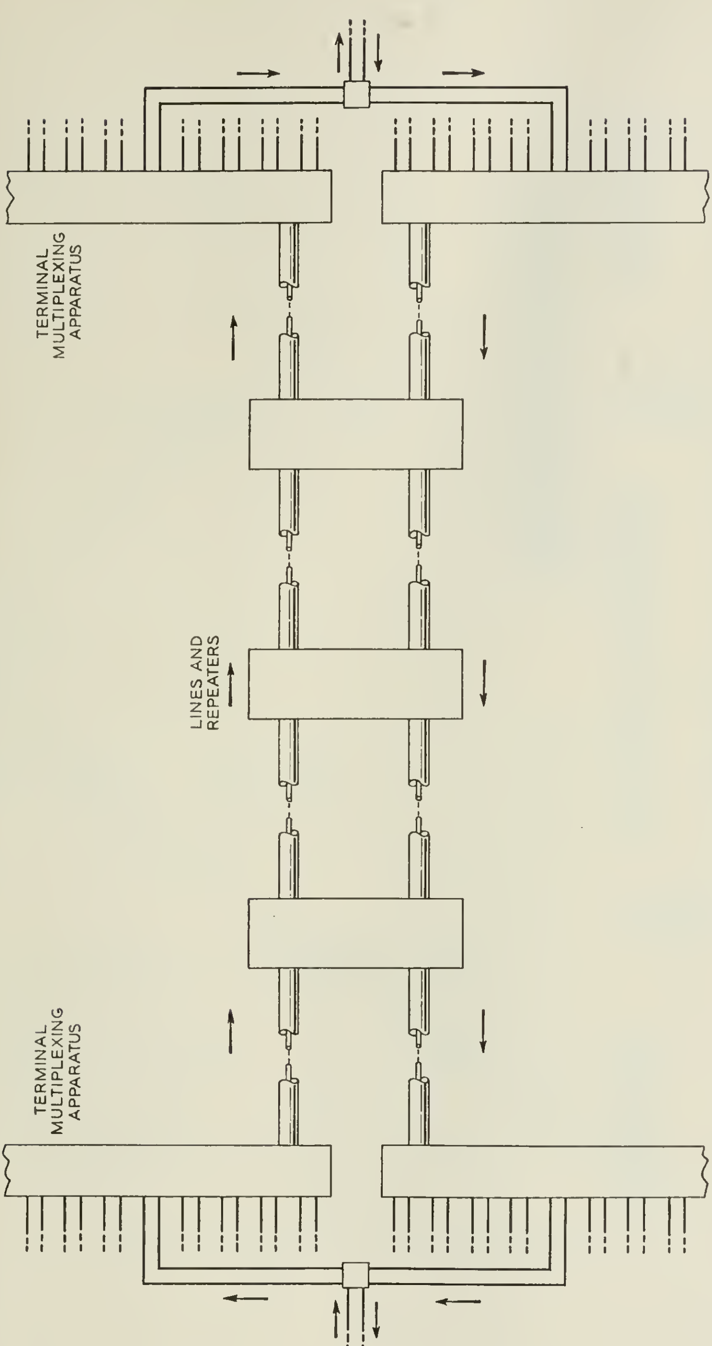


Fig. 17—Block schematic of the coaxial system.

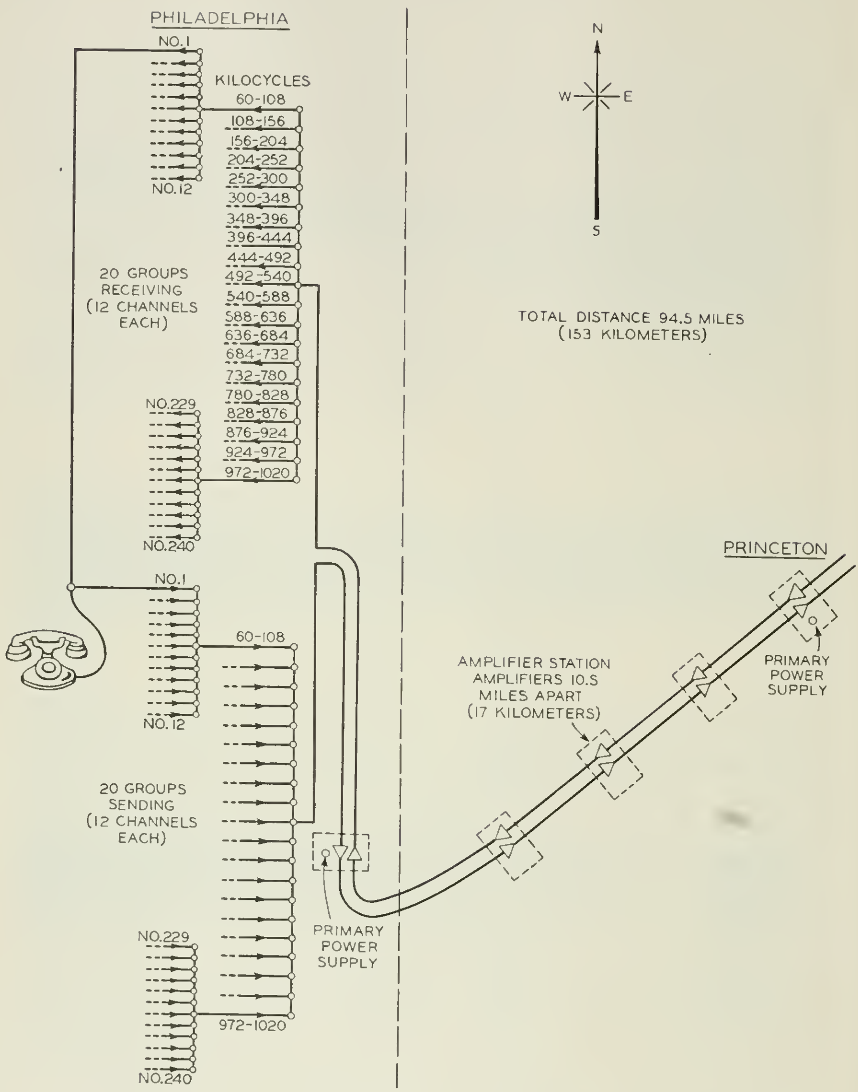


Fig. 18—Schematic representation of the coaxial system.

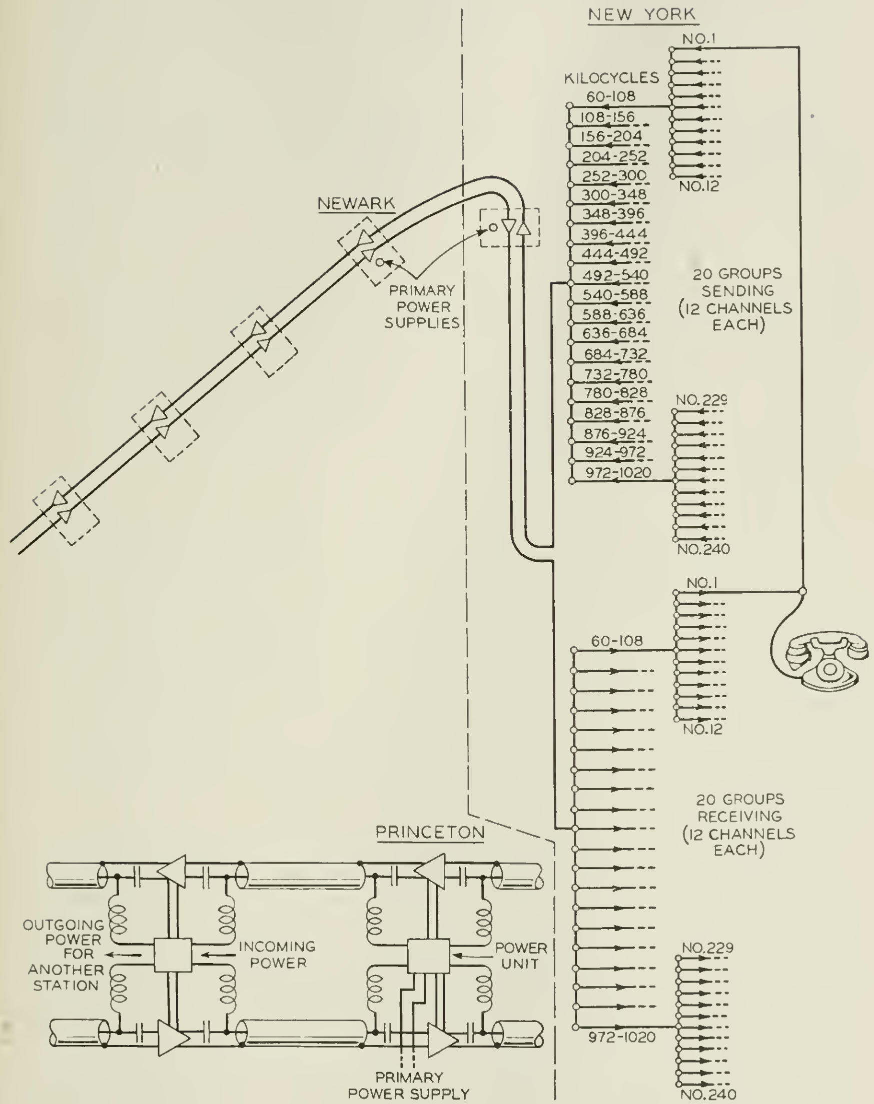


Fig. 18—*Continued from page 140.*

distance of about 95 miles (153 km.), we are employing two coaxial structures within a single lead sheath so as to provide transmission in each direction. The complete structure is shown in Fig. 16. Some of the space within the circular lead sheath is filled with ordinary cable pairs which are tapped out at repeater points for testing and trial purposes. Repeaters are being provided along this coaxial circuit at intervals of about 10 miles (16 km.), which allows frequencies up to about 1,000,000 cycles to be transmitted. Figures 17 and 18 show schematically this coaxial broad-band system.

At each repeater point there is a single amplifier for each coaxial unit, that is, one for each direction of transmission. This amplifier handles the entire number of simultaneous conversations obtainable, which is 240 for the million-cycle band. The amplifiers are equipped with automatic regulating arrangements to adjust the amplification to correspond to the attenuation of the cable as the temperature changes. As you would expect, the attenuation variations with temperature are not the same at different frequencies, but the regulating system meets this condition.

Development of means for combining and separating the channels at the terminals is an interesting feature. Some of our engineers have termed the means employed "unit group." The twelve carrier channels employed in both the cable and open-wire broad-band systems are provided in a unit called a "12-channel terminal." The coaxial system employs essentially the same 12-channel terminal, but it employs twenty of them to provide the total of 240 channels. The output of one 12-channel terminal is put directly on the line, but the outputs of the other nineteen are modulated a second time, and raised to successive positions in the frequency spectrum.

The use of double modulation has two principal advantages. In the first place it simplifies the apparatus by requiring fewer different carrier frequencies, and since it employs a channel terminal that is used by all broad-band systems, considerable economies in production are secured. The chief advantage of double modulation in the coaxial system, however, is that it simplifies the separation of the side-bands resulting from modulation, which is necessary because only one of the side-bands is transmitted. If only a single modulation were employed, the two sidebands of the upper carrier frequency would be separated by only about .05% of the carrier frequency, while with double modulation the narrowest separation is about ten times this amount.

With such a coordinated program of broad-band telephony, toll transmission takes on a new appearance. Not only will the provision of additional channels be simplified by the availability of these new systems, but the cost per channel should be somewhat decreased. At the same time the quality of the circuits, from the standpoint of voice transmission, has been improved, so that a multiple gain is obtained.

Lecturer's Note: The lecturer wishes to acknowledge his indebtedness to a number of members of Bell Telephone Laboratories' staff. In particular, he wishes to thank Messrs. H. A. Affel, A. B. Clark and P. C. Jones for their assistance in preparing this material.

Crosstalk Between Coaxial Transmission Lines

By S. A. SCHELKUNOFF and T. M. ODARENKO

The general theory of coaxial pairs was dealt with in an article on "The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields" by S. A. Schelkunoff (*B. S. T. J.*, Oct., 1934). The present paper considers a specific aspect of the general theory, namely, crosstalk.

Formulae for the crosstalk are developed in terms of the distributed mutual impedance, the constants of the transmission lines and the terminal impedances. Some limiting cases are given special consideration. The theory is then applied to a few special types of coaxial structures studied experimentally and a close agreement is shown between the results of calculations and of laboratory measurements.

If the outer members of coaxial pairs are complicated structures rather than solid cylindrical shells, the crosstalk formulae still apply but the mutual impedances and the transmission constants which are involved in these formulae must be determined experimentally since these quantities cannot always be calculated with sufficient accuracy.

The crosstalk between coaxial pairs with solid outer conductors rapidly decreases with increasing frequency while the crosstalk between unshielded balanced pairs increases. In the low frequency range there is less crosstalk between such balanced pairs than between coaxial pairs but at high frequencies the reverse is true. The diminution of crosstalk between coaxial pairs with increasing frequency is caused by an ever increasing shielding action furnished by the outer conductors of the pairs.

Finally, crosstalk in long lines using coaxial conductors is discussed and the conclusion is reached that, unlike the case of the balanced structure, the far-end crosstalk imposes a more severe condition than the near-end crosstalk in two-way systems which involve more than two coaxial conductors.

A COAXIAL line consists of an outer conducting tube which envelops a centrally disposed inner conductor. The circuit is formed between the inner surface of the outer conductor and the outer surface of the inner conductor. Since any kind of high-frequency external interference tends to concentrate on the outer surface of the outer conductor and the transmitted current on the inner surface of the outer circuit, the outer conductor serves also as a shield, the shielding effect being more effective the higher the frequency.

Due to this very substantial shielding at high frequencies, this type of circuit has been a matter of increased interest for use as a connector between radio transmitting or receiving apparatus and antennae, as well as a wide frequency band transmitting medium for long distance multiplex telephony or television. It has been a subject of discussion in several articles published in this country and abroad.*

The purpose of this paper is to dwell at some length on the shielding characteristics of a structure exposed to interference from a similar structure placed in close proximity. Such interference is usually referred to as crosstalk between two adjacent circuits, so that the purpose of this paper is a study of crosstalk between two coaxial circuits. In what follows we shall give an account of the theory of crosstalk, the results of experimental studies, and application of these to long lines employing coaxial conductors.

General Considerations

Let us consider a simple case of two transmission lines (Fig. 1) and let us assume that both lines are terminated in their characteristic

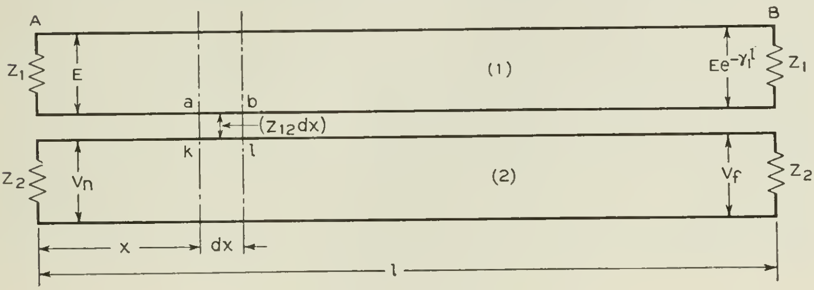


Fig. 1—Direct crosstalk between coaxial pairs.

impedances Z_1 and Z_2, and that their propagation constants per unit length are γ_1 and γ_2 respectively. If the disturbing voltage E is applied to the left end of line (1) and the induced voltage V_n is measured at the corresponding end of line (2), the ratio V_n/E is called the near-end crosstalk ratio from circuit (1) into circuit (2). Similarly, if V_f is the induced voltage as measured at the right end of line (2), when the disturbing voltage E is applied to the left end of circuit (1), we define the ratio $V_f/Ee^{-\gamma_1 l}$ as the far-end crosstalk ratio from circuit (1) into circuit (2). For convenience, we shall speak of the near-end crosstalk and the far-end crosstalk whenever the voltage crosstalk ratios are actually involved. Thus, the magnitude of crosstalk will be given by the absolute value of the corresponding crosstalk ratio. It might be expressed either in decibels as is done in this paper or it might be given

* For references see end of paper.

in terms of crosstalk units, if the absolute value of the crosstalk ratio is multiplied by a factor 10^6.

It is well to observe at this point that, depending upon special conditions, the *significant* crosstalk ratio may be either the voltage ratio or the current ratio or the power ratio. The power ratio, or more commonly the square root of it, is usually the most important but if the outputs are impressed on the grids of vacuum tubes then the voltage ratio becomes the significant measure of crosstalk. However, if one crosstalk ratio is known, any other crosstalk ratio can be readily determined provided that the characteristics of both circuits are known. Thus for the conditions of Fig. 1 the value of far-end crosstalk as given by the ratio $V_f/Ee^{-\gamma_1 l}$ in the voltage ratio system will become $(V_f/Ee^{-\gamma_1 l})(Z_1/Z_2)$ in the current ratio system.

In general, the crosstalk between any two transmission lines depends upon the existence of mutual impedances and mutual admittances between the lines. Generally, then, one can differentiate between two types of crosstalk. The first is produced by an electromotive force in series with the disturbed line in consequence of mutual impedances between the lines, and can be appropriately designated as the "impedance crosstalk." The other is due to an electromotive force in shunt with the disturbed line, induced by virtue of mutual admittances, and can be designated as the "admittance crosstalk." The two types of crosstalk are frequently referred to either as "electromagnetic crosstalk" and "electrostatic crosstalk" or as "magnetic crosstalk" and "electric crosstalk"; the latter terminology is the better of the two.

The Mutual Impedance

Consider the simplest crosstalking system consisting of two circuits only, such as shown schematically in Fig. 1. The mutual impedance between two corresponding short sections of the two lines, between the disturbing section *ab* and the disturbed section *kl*, for instance, will be defined as the ratio of the electromotive force induced in the disturbed section to the current in the disturbing section. In what follows we shall assume that the coupling between the two transmission lines is uniformly distributed; that is, that the mutual impedance between two infinitely small sections, each of length dx is $Z_{12}dx$, where Z_{12} is independent of x. The constant Z_{12} is the mutual impedance *per unit length.*

The mutual impedance between coaxial pairs will be dealt with in a later section. For the present we need only assume that this impedance can be either calculated or measured. We shall find that the crosstalk is proportional to the mutual impedance, the remaining factors depending upon the length of transmission lines and the character of their terminations.

Direct and Indirect Crosstalk

Let us now return to the circuits shown in Fig. 1. Because of the mutual impedance between the two circuits a certain amount of the disturbing energy is transferred from line (1) to line (2), producing voltages at both ends. The voltage at the end A determines the near-end crosstalk. The type of crosstalk present in a simple system of two circuits only in consequence of the direct transmission of energy from one circuit into another we shall call the direct crosstalk. Later on we shall discuss the case where three circuits are involved in such a way that the energy transfer takes place via an intermediate circuit, causing the crosstalk which we call the indirect crosstalk. Both direct and indirect types of crosstalk have a close correspondence to the types of crosstalk used in connection with work on the open-wire lines or the balanced pairs as discussed in the paper on open-wire crosstalk.[1] The direct crosstalk of the present paper is the direct transverse crosstalk; our indirect crosstalk is the total crosstalk due to the presence of the third circuit and as such is the resultant of the indirect transverse crosstalk and the interaction crosstalk of the above paper. Following the general method outlined in the present paper one can easily subdivide the indirect crosstalk into its components. Since only simple crosstalk systems consisting of two coaxial conductors are considered in our paper, the work has not been carried through.

Direct Near-End Crosstalk

We proceed now to develop the formula for the direct near-end crosstalk. The line (1) being terminated in its characteristic impedance Z_1 the current through the generator is E/Z_1 and therefore the current in the section ab is

$$i_{ab} = \frac{Ee^{-\gamma_1 x}}{Z_1}. \tag{1}$$

Hence, by definition of the mutual impedance, the electromotive force induced in the section kl is

$$e_{kl} = i_{ab} Z_{12} dx = \frac{Ee^{-\gamma_1 x}}{Z_1} Z_{12} dx, \tag{2}$$

and the current in the section kl

$$i_{kl} = \frac{e_{kl}}{2Z_2} = \frac{Ee^{-\gamma_1 x}}{2Z_1 Z_2} Z_{12} dx. \tag{3}$$

Therefore the current at the left end of line (2) due to the electromotive force e_{kl} is given by the expression

$$(i_{kl})_n = i_{kl} e^{-\gamma_2 x} = \frac{EZ_{12}}{2Z_1 Z_2} e^{-(\gamma_1+\gamma_2)x} dx. \tag{4}$$

The contribution dV_n to the potential across the left end of line (2) due to crosstalk in the section dx, x cm. away from the left end of the line, is

$$dV_n = (i_{kl})_n Z_2 = \frac{E}{2Z_1} Z_{12}\, e^{-(\gamma_1+\gamma_2)x} dx. \tag{5}$$

Hence the total induced voltage at the near end is

$$V_n = \int_0^l dV_n = \int_0^l \frac{E}{2Z_1} Z_{12}\, e^{-(\gamma_1+\gamma_2)x} dx. \tag{6}$$

Integrating, we obtain

$$V_n = E \frac{Z_{12}}{2Z_1} \frac{1 - e^{-(\gamma_1+\gamma_2)l}}{\gamma_1 + \gamma_2}. \tag{7}$$

The near-end crosstalk is thus given by the expression

$$N_{12} = \left(\frac{V_n}{E}\right)_{12} = \frac{Z_{12}}{2Z_1} \frac{1 - e^{-(\gamma_1+\gamma_2)l}}{\gamma_1 + \gamma_2}. \tag{8}$$

If we reversed the procedure and considered the crosstalk from circuit (2) into circuit (1), we would similarly obtain

$$N_{21} = \left(\frac{V_n}{E}\right)_{21} = \frac{Z_{21}}{2Z_2} \frac{1 - e^{-(\gamma_1+\gamma_2)l}}{\gamma_1 + \gamma_2}. \tag{9}$$

By the reciprocity theorem, $Z_{21} = Z_{12}$. Incidentally, if instead of adopting as the definition of crosstalk the ratio of two voltages we regarded it as the ratio of the induced voltage to the current through the disturbing generator, we should have obtained $N_{21} = N_{12}$.

Finally, if the circuits are alike $Z_1 = Z_2 = Z_0$, $\gamma_1 = \gamma_2 = \gamma$ and the near-end crosstalk is given by the expression

$$N = \frac{V_n}{E} = \frac{Z_{12}}{2Z_0} \frac{1 - e^{-2\gamma l}}{2\gamma}. \tag{10}$$

We observe that the near-end crosstalk depends on length l. Two limiting cases are of importance here. For a length l so small, that for a given frequency $2\gamma^2 l^2$ is negligible when compared with $2\gamma l$, we have

$$\begin{aligned} e^{-2\gamma l} &= 1 - 2\gamma l + 2\gamma^2 l^2 \\ &= 1 - 2\gamma l, \end{aligned} \tag{11}$$

and the expression (10) becomes

$$N = \frac{V_n}{E} = \frac{Z_{12}}{2Z_0} l. \tag{12}$$

The near-end crosstalk is therefore proportional to l.

For very large values of γl, that is, a very high frequency or extreme length or both, where the exponential expression is negligible as compared to unity, the expression (10) becomes

$$N = \frac{V_n}{E} = \frac{Z_{12}}{2Z_0} \frac{1}{2\gamma}, \tag{13}$$

which is independent of length.

The variation of the near-end crosstalk with length for intermediate values of γl can be best followed if instead of the expression (10) we use its absolute value

$$|N_{12}| = \left|\frac{V_n}{E}\right| = \frac{|Z_{12}|}{2|Z_1|} \frac{\sqrt{1 - 2e^{-2\alpha l} \cos (2\beta l) + e^{-4\alpha l}}}{\sqrt{\alpha^2 + \beta^2}}. \tag{14}$$

Here

$$\gamma = \alpha + i\beta, \tag{15}$$

α is the attenuation constant in nepers per unit length and β is the phase constant in radians per unit length.

We observe that for a given value of l one of the factors in (14) is oscillating with frequency. Thus, if we plot the crosstalk against frequency, the resulting curve is a wavy line superimposed upon a smooth curve, with the successive minimum points corresponding to the frequencies for which the given line is practically a multiple of half wave-lengths. The smooth curve is of course given by the magnitude of the expression (13). The curves on Fig. 2 illustrate the change of the near-end crosstalk with frequency for different lengths of a triple coaxial line made of copper conductors.

Direct Far-End Crosstalk

In order to determine the far-end crosstalk, we have to compute the induced voltage arriving at the far end of the system. Proceeding in a way similar to the derivation of the near-end crosstalk, we obtain the contribution dV_f to the potential across the right end of circuit (2), due

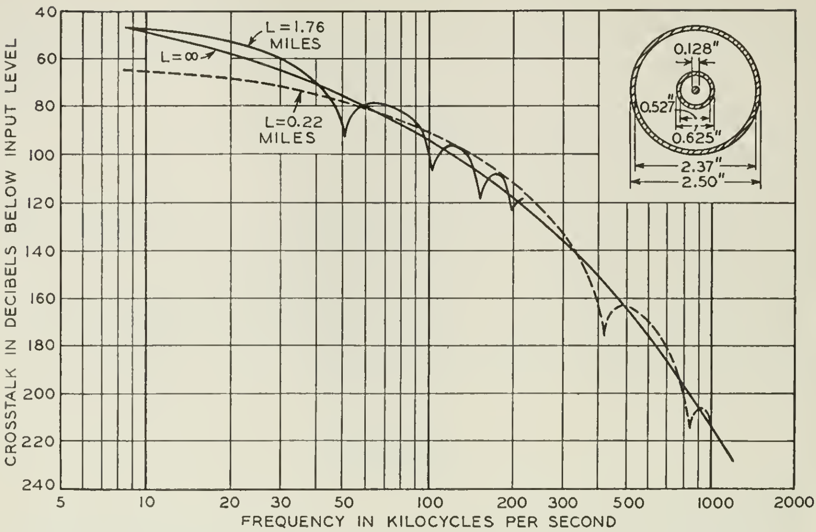


Fig. 2—Direct near-end crosstalk in a system of three coaxial conductors.

to the electromotive force in the section kl, to be given by the expression

$$dV_f = \frac{E}{2Z_1} Z_{12} e^{-\gamma_2 l} e^{(\gamma_2 - \gamma_1)x} dx. \tag{16}$$

Integrating this over the total length l we obtain the total voltage induced at the far end

$$V_f = E \frac{Z_{12}}{2Z_1} \frac{e^{-\gamma_2 l} - e^{-\gamma_1 l}}{\gamma_1 - \gamma_2}, \tag{17}$$

and the far-end crosstalk from circuit (1) into circuit (2) is

$$F_{12} = \frac{V_f}{Ee^{-\gamma_1 l}} = \frac{Z_{12}}{2Z_1} \frac{1 - e^{(\gamma_1 - \gamma_2)l}}{\gamma_2 - \gamma_1}. \tag{18}$$

If two similar lines are considered, with equal propagation constants and the characteristic impedances, equation (18) becomes

$$F = \frac{V_f}{Ee^{-\gamma l}} = \frac{Z_{12}}{2Z_0} l. \tag{19}$$

The far-end crosstalk is proportional to the length of the line at all frequencies.

Inasmuch as we have ignored the reaction of the induced currents upon the disturbing line, the foregoing equations must be regarded as approximations. Under practical conditions these approximations are very good. Only equation (19) must not be pushed to its absurd implication, that for long enough transmission lines most energy will eventually travel via the disturbed line. The true limiting condition is that the energy will ultimately be divided equally between the two lines.

Crosstalk Via an Intermediate Circuit

The simplest case of the coaxial conductor system where the only crosstalk present is of the direct crosstalk type, as considered in the previous section, is the triple coaxial conductor. The mutual coupling in this case is due only to the transfer impedance between two circuits, as there are no other physical circuits involved. The case of two single coaxial conductors, the outer shells of which are in continuous electrical contact or strapped at frequent intervals, approximates the condition for the direct crosstalk if the system is sufficiently removed from any conducting matter. When two single parallel conductors in free space do not touch, an extra transmission line, an "intermediate circuit," is present consisting of the two outer shells of the coaxial conductors. Even two conductors, the shells of which are electrically connected, will form an intermediate circuit consisting of the outer shells and the other parallel conductors.

The voltage impressed on the disturbing coaxial circuit induces currents and voltages in the intermediate circuit, which now acts as a disturbing circuit for the second coaxial circuit, thus causing crosstalk. We shall now consider the near-end and far-end components of this indirect type of crosstalk.

Indirect Near-End Crosstalk

Let us consider a system shown in Fig. 3. The circuit (3) is the

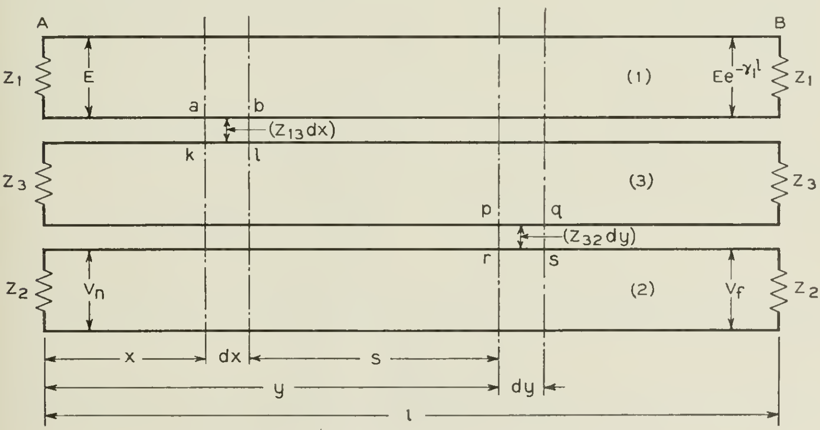


Fig. 3—Indirect crosstalk between two coaxial pairs.

intermediate circuit with an impedance Z_3 and propagation constant per unit length γ_3. Let the disturbing voltage E be applied at the A end of circuit (1). Then the current in the section kl is given by the expression similar to (3)

$$i_{kl} = \frac{EZ_{13}}{2Z_1Z_3} e^{-\gamma_1 x} dx. \tag{20}$$

The current in the section pq is

$$i_{pq} = i_{kl} e^{-\gamma_3 s}, \tag{21}$$

where

$$s = |y - x|. \tag{22}$$

The total current in the generic element of the intermediate transmission line due to coupling with circuit (1) is, then,

$$I_y = \int_0^l i_{pq} = \frac{EZ_{13}}{2Z_1Z_3} \int_0^l e^{-\gamma_3 s} e^{-\gamma_1 x} dx. \tag{23}$$

In carrying out the process of integration, we must keep in mind that from 0 to y, $s = y - x$ and from y to l, $s = x - y$.

Hence, we have

$$\int_0^l e^{-\gamma_3 s} e^{-\gamma_1 x} dx = e^{-\gamma_3 y} \int_0^y e^{(\gamma_3 - \gamma_1) x} dx + e^{\gamma_3 y} \int_y^l e^{-(\gamma_3 + \gamma_1) x} dx,$$

and

$$I_y = \frac{EZ_{13}}{2Z_1Z_3} \left[\frac{e^{-\gamma_1 y} - e^{-\gamma_3 y}}{\gamma_3 - \gamma_1} + \frac{e^{-\gamma_1 y} - e^{\gamma_3 y} e^{-(\gamma_3 + \gamma_1) l}}{\gamma_3 + \gamma_1} \right]. \tag{24}$$

The elementary electromotive force induced in the second coaxial conductor by the current I_y is

$$E_y = Z_{32} I_y dy. \tag{25}$$

The contribution of this electromotive force to the voltage across the near-end of the second coaxial pair will be then

$$dV_n = \frac{1}{2} E_y e^{-\gamma_2 y} = \frac{Z_{32}}{2} I_y e^{-\gamma_2 y} dy. \tag{26}$$

The total induced near-end voltage will be given by the expression

$$V_n = \frac{1}{2} \int_0^l Z_{32} I_y e^{-\gamma_2 y} dy. \tag{27}$$

Using the expressions (23) and (24), we obtain

$$V_n = E\frac{Z_{13}Z_{32}}{4Z_1Z_3}S_n, \tag{28}$$

where

$$S_n = \int_0^l e^{-\gamma_2 y}\left[\frac{e^{-\gamma_1 y} - e^{-\gamma_3 y}}{\gamma_3 - \gamma_1} + \frac{e^{-\gamma_1 y} - e^{\gamma_3 y}e^{-(\gamma_3+\gamma_1)l}}{\gamma_3 + \gamma_1}\right]dy. \tag{29}$$

Integrating (29) we have

$$S_n = \frac{2\gamma_3}{\gamma_1 + \gamma_2}\frac{1 - e^{-(\gamma_2+\gamma_1)l}}{\gamma_3^2 - \gamma_1^2} - \frac{1 - e^{-(\gamma_3+\gamma_2)l}}{(\gamma_3 - \gamma_1)(\gamma_3 + \gamma_2)} + \frac{1 - e^{(\gamma_3-\gamma_2)l}}{(\gamma_3 + \gamma_1)(\gamma_3 - \gamma_2)}e^{-(\gamma_3+\gamma_1)l}. \tag{30}$$

Thus, the near-end crosstalk from circuit (1) into circuit (2) via the intermediate circuit (3) is given by the expression

$$N_{12}' = \frac{V_n}{E} = \frac{Z_{13}Z_{23}}{4Z_1Z_3}S_n. \tag{31a}$$

In a similar manner we can derive the following expression for the near-end crosstalk from circuit (2) into circuit (1) via the intermediate circuit (3):

$$N_{21}' = \frac{Z_{13}Z_{23}}{4Z_2Z_3}S_n. \tag{31b}$$

The factor S_n present in (31*b*) is the same as in (31*a*), being symmetrical with respect to the subscripts 1 and 2 as a close inspection of the formula (30) would prove. $Z_{13} = Z_{31}$ and $Z_{23} = Z_{32}$ by the reciprocity theorem.

For the case of two similar coaxial pairs with equal characteristic impedances Z_0 and propagation constants γ, and symmetrically placed with respect to the intermediate line, so that $Z_{13} = Z_{32}$, we have

$$N' = \frac{(Z_{13})^2}{4Z_0Z_3}\left[\frac{\gamma_3}{\gamma}\frac{1 - e^{-2\gamma l}}{\gamma_3^2 - \gamma^2} - \frac{1 - 2e^{-(\gamma_3+\gamma)l} + e^{-2\gamma l}}{\gamma_3^2 - \gamma^2}\right]. \tag{32}$$

Now for short lengths we may use again the approximation

$$e^{-a} = 1 - a + \tfrac{1}{2}a^2. \tag{33}$$

The expression in the brackets of (32) then becomes equal to l^2 and the near-end crosstalk is given by the expression

$$N' = \frac{(Z_{13})^2}{4Z_0Z_3} l^2, \tag{34}$$

which is proportional to the second power of length.

For γl very large we can rewrite expression (32) as follows:

$$N' = \frac{(Z_{13})^2}{4Z_0Z_3} \frac{1}{\gamma(\gamma_3 + \gamma)} \cdot \tag{35}$$

Thus, for a system sufficiently long the near-end crosstalk via an intermediate line is independent of length.

If the intermediate transmission line is short-circuited a large number of times per wave-length, its propagation constant γ_3 becomes very large on the average and we have approximately

$$S_n = \frac{2[1 - e^{-(\gamma_2+\gamma_1)l}]}{(\gamma_1 + \gamma_2)\gamma_3}, \tag{36}$$

and

$$N_{12}' = \frac{V_n}{E} = \frac{Z_{13}Z_{23}}{2Z_1Z_3\gamma_3} \frac{1 - e^{-(\gamma_1+\gamma_2)l}}{\gamma_1 + \gamma_2} \cdot \tag{37}$$

But $Z_3\gamma_3 = Z$, the distributed series impedance of the intermediate transmission line. Hence the "indirect" cross-talk becomes direct with the mutual impedance given by

$$Z_{12} = \frac{Z_{13}Z_{23}}{Z} \cdot$$

Indirect Far-End Crosstalk

Using the method outlined in the previous section we arrive at the following expression for the far-end crosstalk from circuit (1) into circuit (2) via the intermediate circuit (3); see Fig. 3.

$$F_{12}' = \frac{V_f}{Ee^{-\gamma_1 l}} = \frac{Z_{13}Z_{32}}{4Z_1Z_3} S_f. \tag{38a}$$

The crosstalk from circuit (2) into circuit (1) will be given by a similar expression with Z_2 replacing Z_1 in the denominator, namely

$$F_{21}' = \frac{Z_{13}Z_{32}}{4Z_2Z_3} S_f. \tag{38b}$$

The factor S_f used in the above formulae is given by the expression

$$S_f = e^{-(\gamma_2-\gamma_1)l}\left[\frac{2\gamma_3}{\gamma_1-\gamma_2}\,\frac{1-e^{-(\gamma_1-\gamma_2)l}}{\gamma_3^2-\gamma_1^2} - \frac{1-e^{-(\gamma_3-\gamma_2)l}}{(\gamma_3-\gamma_1)(\gamma_3-\gamma_2)} + \frac{1-e^{(\gamma_3+\gamma_2)l}}{(\gamma_3+\gamma_1)(\gamma_3+\gamma_2)}\,e^{-(\gamma_3+\gamma_1)l}\right]. \quad (39)$$

When both coaxial pairs are similar and placed symmetrically with respect to the intermediate conductors we obtain the following expression for the far-end crosstalk between two coaxial conductors via an intermediate circuit:

$$F' = \frac{(Z_{13})^2}{4Z_0Z_3}\left[\frac{2\gamma_3 l}{\gamma_3^2-\gamma^2} - \frac{1-e^{-(\gamma_3-\gamma)l}}{(\gamma_3-\gamma)^2} - \frac{1-e^{-(\gamma_3+\gamma)l}}{(\gamma_3+\gamma)^2}\right]. \quad (40)$$

For small l the expression for the far-end crosstalk becomes

$$F' = \frac{(Z_{13})^2}{4Z_0Z_3}\,l^2, \quad (41)$$

which is the same as (34) for the near-end crosstalk.

For large l and provided the attenuation of the intermediate circuit is greater than that of the coaxial circuit we have

$$F' = \frac{(Z_{13})^2}{4Z_0Z_3}\left[\frac{2\gamma_3 l}{\gamma_3^2-\gamma^2} - \frac{2(\gamma_3^2+\gamma^2)}{(\gamma_3^2-\gamma^2)^2}\right]. \quad (42)$$

Finally, letting γ_3 approach γ and considering a limiting case when attenuation of the intermediate circuit is equal to attenuation of either of the coaxial conductors we obtain

$$F' = \frac{(Z_{13})^2}{4Z_0Z_3}\left[\frac{l}{2\gamma} + \frac{1}{2}l^2 - \frac{1-e^{-2\gamma l}}{4\gamma^2}\right]. \quad (43)$$

If the intermediate transmission line is short-circuited a large number of times per wave-length its propagation constant γ_3 becomes very large on the average. The equation (37) becomes, then,

$$S_f = \frac{2[1-e^{-(\gamma_2-\gamma_1)l}]}{(\gamma_2-\gamma_1)\gamma_3}, \quad (44)$$

and

$$F_{12}' = \frac{Z_{13}Z_{32}}{2Z_1Z_3\gamma_3}\,\frac{1-e^{(\gamma_1-\gamma_2)l}}{\gamma_2-\gamma_1}. \quad (45)$$

The indirect crosstalk becomes direct with the mutual impedance given by the expression

$$Z_{12} = \frac{Z_{13}Z_{23}}{Z_3\gamma_3} = \frac{Z_{13}Z_{23}}{Z}, \quad (46)$$

where $Z = Z_3\gamma_3$ is the distributed series impedance of the intermediate transmission line.

Comparison between Direct Crosstalk and Crosstalk via Intermediate Circuit for Two Parallel Coaxial Conductors

We have already seen that two parallel coaxial conductors in free space form actually three transmission circuits, the third circuit being formed by two outer shells of the coaxial conductors. When this third line is shorted by direct electrical contact or by frequent straps only direct crosstalk is present. When the third circuit is terminated in its characteristic impedance we have crosstalk via the third circuit. In this last case, however, the crosstalk via the third circuit is also the total crosstalk, since the only available path for the transfer of interfering energy is via the third circuit. Thus, we can directly compare the values of crosstalk for the system for both conditions.

We have shown that for sufficiently short lengths of the crosstalk exposure the direct type of crosstalk is given by (12) or (19), namely,

$$F = N = \frac{Z_{12}}{2Z_0} l. \tag{47}$$

We have also found that the crosstalk via an intermediate circuit is given by (34) or (41) provided that the length of conductors is small enough. Thus

$$F' = N' = \frac{Z_{13}^2}{4Z_0Z_3} l^2. \tag{48}$$

Consequently

$$\frac{F'}{F} = \frac{N'}{N} = \frac{Z_{13}^2}{2Z_{12}Z_3} l. \tag{49}$$

In seeking an experimental verification of equation (49) a series of measurements were taken on a pair of coaxial conductors of varying lengths, separations, and different terminating conditions of the third circuit. The results agreed fully with the theory.

Mutual Impedance

Like the other constants of transmission lines the distributed mutual impedance can be measured. In certain cases, however, it is possible to obtain simple formulae for this impedance. For details of such calculations the reader is referred to a paper by one of the authors.[2]

In this paper the mutual impedance is expressed in terms of *surface* *transfer impedances.* Consider a coaxial pair whose outer conductor is either a homogeneous cylindrical shell or a shell consisting of coaxial homogeneous cylindrical layers of different conducting substances. The transfer impedance from the inner to the outer surface of the outer conductor is then defined as the voltage gradient on the outer surface per unit current in the conductor. In a triple coaxial conductor system this transfer impedance is evidently the mutual impedance between two transmission lines, one comprised of the two inner conductors and the other of the two outer conductors. On the other hand the mutual impedance has a quite different value if one line consists of the two inner conductors while the other is comprised of the innermost and the outermost conductors.

The surface transfer impedance of a homogeneous cylindrical shell is given by the following expression, good to a fraction of a per cent for all frequencies up to the optical range if the thickness t is smaller than 20 per cent of the average radius

$$Z_{ab} = \frac{\eta}{2\pi\sqrt{ab}} \operatorname{csch}(\sigma t). \tag{50}$$

In this equation:

a is the inner radius of the middle shell in cm.
b is the outer radius of the middle shell in cm.
t is the thickness of the middle shell in cm.
$\sigma = \sqrt{2\pi g\mu f i}$ nepers per cm.
$\eta = \frac{\sigma}{g} = \sqrt{\frac{2\pi\mu f i}{g}}$ ohms
g is the conductivity in mhos per cm.*
μ is the permeability in henries per cm.*
f is the frequency in cycles per second.

If the ratio of the diameters of the shell is not greater than 4/3 the following formula correct to 1 per cent at any frequency will hold for the absolute value of the transfer impedance

$$|Z_{ab}| = R_{DC} \frac{u}{\sqrt{\cosh u - \cos u}}, \tag{51}$$

where

$$R_{DC} = \text{the dc resistance of the shell,}$$
$$u = t\sqrt{4\pi\mu g f}.$$

* As in the previous paper by Schelkunoff we adhere throughout this article to the practical system of units based on the c.g.s. system. For copper of 100 per cent conductivity

$$g = 5.8005 \times 10^5 \text{ mhos/cm. and } \mu = 4\pi 10^{-9} \text{ henries/cm.}$$

The expression (51) is plotted in Fig. 3, p. 559 of Schelkunoff's paper.[2]

As it has been already mentioned, (50) and (51) represent the mutual impedance in a triple conductor *coaxial* system. One might anticipate that if the arrangement is not coaxial the mutual impedance has a different value. This is indeed the case if all three conductors have different axes. But if one transmission line is a strictly coaxial pair, then its own current remains substantially uniform around its axis and from equation (81) of Schelkunoff's paper we immediately conclude that the mutual impedance will be the same as if *all three* conductors were coaxial. *Both* transmission lines must be eccentric before their mutual impedance becomes affected by their eccentricities. Thus the mutual impedance Z_{13} between a coaxial circuit and the circuit consisting of its outer shell and a cylindrical shell parallel to it is given very accurately by (50) and (51).

The surface transfer impedance across a shell consisting of two coaxial homogeneous layers is given by

$$Z_{12} = \frac{(Z_{ab})_1 (Z_{ab})_2}{Z}, \tag{52}$$

where Z_{ab} is the transfer impedance for each layer and Z is the series impedance per unit length of the circuit consisting of the two layers insulated from each other by an infinitely thin film, when one layer is used as the return conductor for the other.

The mutual impedance between two coaxial pairs the outer conductors of which are short-circuited at frequent intervals is also given by (52) provided Z is interpreted as the distributed series impedance of the intermediate transmission line comprised of the outer shells of the given coaxial pairs. This Z is the sum of the internal impedances of the two shells $(Z_{bb})_1$ and $(Z_{bb})_2$ and of the external inductive reactance ωL_e due to the magnetic flux between the shells. If the proximity effect is disregarded, the internal impedance of a single cylindrical shell is the same as that with a coaxial return and various expressions for it are given in equations (75) and (82) in the previous paper.[2] The inclusion of the proximity effect does not complicate the formulae if the separation between the shells is fairly large by comparison with their radii, but in this case the proximity effect is not very large either. The more accurate determination of Z leads to complicated formulae; for these the reader is referred to a paper by Mrs. S. P. Mead.[6] However, at high frequencies the important factors in the mutual impedance are the transfer impedances in the numerator of (52).

Under certain conditions it is easy to obtain approximate values of the denominator of (52) and use them for gauging the limits between which the mutual impedance must lie. If the frequency is so high that the proximity effect has almost reached its ultimate value the external inductance and the internal impedance of the intermediate line are approximately

$$L_e = \frac{\mu}{2\pi}\cosh^{-1}\frac{l^2 - b_1{}^2 - b_2{}^2}{2b_1b_2},$$

$$(Z_{bb})_1 + (Z_{bb})_2 = \frac{1}{2\pi}\sqrt{\frac{i\omega\mu}{g}}\,\frac{\left(\dfrac{1}{b_1}+\dfrac{1}{b_2}\right)+\dfrac{b_1{}^2 - b_2{}^2}{l^2}\left(\dfrac{1}{b_1}-\dfrac{1}{b_2}\right)}{\sqrt{\left[1-\dfrac{(b_1+b_2)^2}{l}\right]\left[1-\dfrac{(b_1-b_2)^2}{l}\right]}}, \tag{53}$$

where b_1 and b_2 are the external radii and l is the interaxial separation. Usually $b_2 = b_1 = b$ and consequently

$$\begin{aligned} L_e &= \frac{\mu}{2\pi}\cosh^{-1}\left(\frac{l^2}{2b^2} - 1\right), \\ (Z_{bb})_1 + (Z_{bb})_2 &= \frac{1}{\pi b}\sqrt{\frac{i\omega\mu}{g}}\left[1 - 4\frac{b^2}{l^2}\right]^{-\frac{1}{2}}. \end{aligned} \tag{54}$$

If the proximity effect is disregarded then the external inductance is simply

$$L_e = \frac{\mu}{\pi}\log_e\frac{l}{\sqrt{b_1b_2}}\cdot \tag{55}$$

For this case, then, the mutual impedance is given by the expression

$$Z_{12} = \frac{(Z_{ab})_1(Z_{ab})_2}{(Z_{bb})_1 + (Z_{bb})_2 + \dfrac{i\omega\mu}{\pi}\log_e\dfrac{l}{\sqrt{b_1b_2}}}\cdot \tag{56}$$

For two identical coaxial conductors the expression is further simplified to

$$Z_{12} = \frac{(Z_{ab})^2}{2Z_{bb} + \dfrac{i\omega\mu}{\pi}\log_e\dfrac{l}{b}}\cdot \tag{57}$$

Measuring Method

As defined above, crosstalk between two transmission lines terminated in their characteristic impedances is given by the ratios of the induced and disturbing voltages. Consequently, if the input voltage into the disturbing circuit is known and the induced voltage at one of the ends of the disturbed pair is measured, the far-end or near-end crosstalk values are obtained readily. In fact, the magnitude of the near-end crosstalk is given by the expression

$$|N| = \left|\frac{V_n}{E}\right| \tag{58}$$

and the magnitude of the far-end crosstalk is given by the expression

$$|F| = \frac{|V_f|}{|E|e^{-\alpha l}} . \tag{59}$$

Taking $20 \log_{10} \frac{1}{|N|}$ and $20 \log_{10} \frac{1}{|F|}$, we obtain an equivalent loss in db between the disturbing and disturbed levels of the two crosstalking circuits. This consideration determined the method of measurements used in our experimental studies.

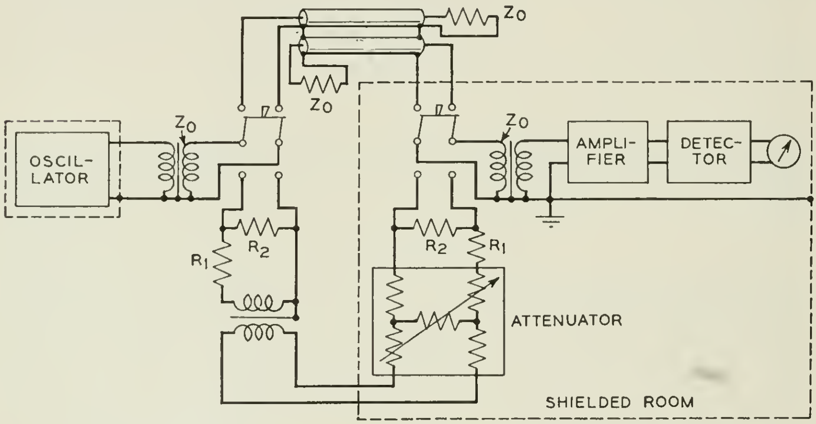


Fig. 4—Crosstalk measuring circuit arranged for far-end measurements.

The circuit used is given in Fig. 4. The two branches of the measuring set are the comparison circuits, the upper containing the crosstalking system and the lower including adjustable attenuators. The input and output impedances of both branches are kept alike by adjusting the resistances R_1 and R_2. Thus, when the lower branch of the circuit is adjusted to produce the same input into the detector as through the crosstalking branch the loss in the calibrating branch gives an equivalent crosstalk loss in db. These values of crosstalk in db below the input level in the disturbing circuit are plotted on all our sketches.

Both coaxial circuits were terminated in resistances closely equal to the absolute values of their characteristic impedances. The terminations were carefully shielded to prevent any crosstalk at these points. Careful shielding and grounding were found necessary to reduce errors due to longitudinal currents, unbalances, and interference between different parts of the measuring circuit. The overall accuracy of the measuring circuit attained was better than .5 db when the difference in input to output levels amounted to 150 db.

Agreement between Theory and Experiments

The general agreement between the theory and the experiments is indicated by the curves in Fig. 5 and Fig. 6, which give the crosstalk values for cases of two small coaxial pairs with solid outer shells in

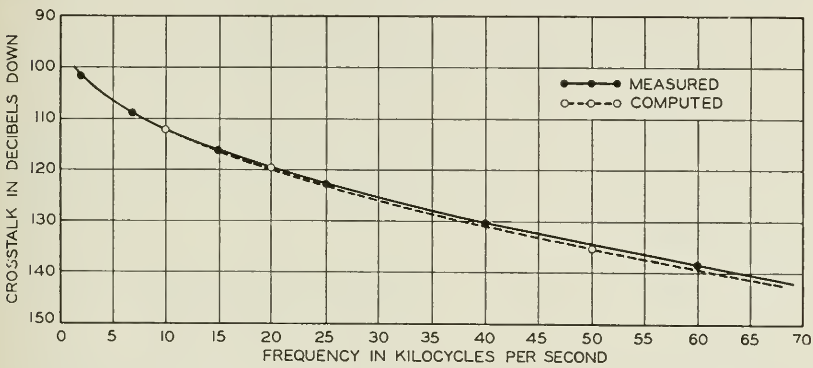


Fig. 5—Crosstalk between two coaxial pairs 20 ft. long using refrigerator pipe .032 inch thick for outer conductors. Both coaxial pairs terminated in 70 ohms. Outer conductors in contact.

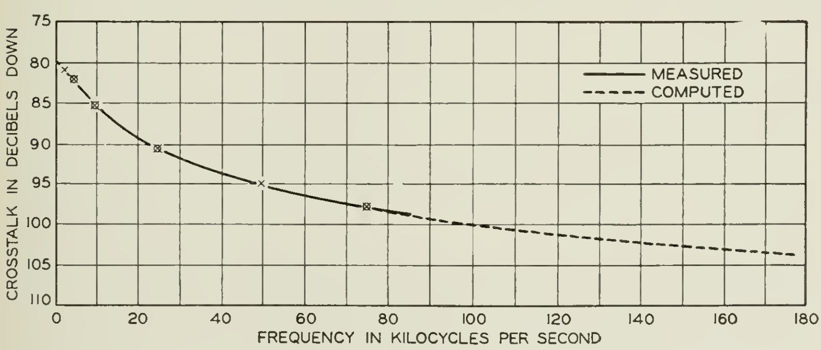


Fig. 6—Crosstalk between two coaxial pairs 25 ft. long. Outer conductor made of copper .008 inch thick, .232 inch inner diameter. Both coaxial pairs terminated in 40 ohms. Outer conductors in contact.

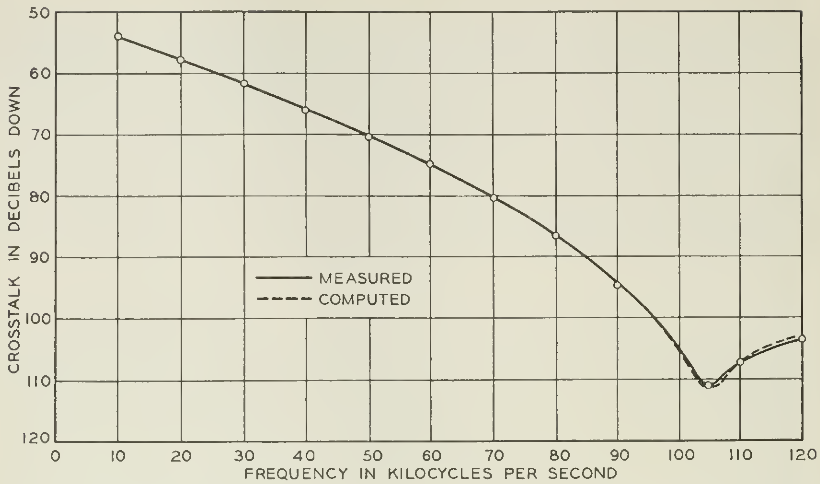


Fig. 7—Near-end crosstalk on a triple coaxial system of conductors at Phoenixville, Pa. Outer to inner circuits. Length .088 mi.

continuous contact. The curves in Fig. 7 show a comparison between measured and computed values of near-end crosstalk for a system of three coaxial conductors .88 mile long as installed at Phoenixville, Pennsylvania.

Also, as was already stated above, full agreement between theory and experiments was established as to validity of equation (49).

Crosstalk in Long Lines Employing Coaxial Conductors

In a system consisting of two coaxial pairs, where two outer conductors are in contact, essentially only one kind of crosstalk is present depending on the direction of transmission on both pairs. It is near-end crosstalk when transmitting in opposite directions and far-end crosstalk for transmission in the same direction. Where more than two coaxial conductors are grouped together and transmission is in both directions both types of crosstalk are present.

Although for a sufficiently short length of crosstalk exposure near-end and far-end crosstalk are identical, in a sufficiently long system the transmission characteristics of the line and associated repeaters will make a marked difference between them. It has been a common experience that in a long system using unshielded balanced structures near-end crosstalk imposes more severe requirements on balance between crosstalking circuits than far-end crosstalk.

We shall now consider a coaxial pair. Here, the magnitude of the far-end crosstalk was found to be given by expression (19). The magnitude of the near-end crosstalk is given by expression (14), which for equal level points becomes

$$|N| = \left|\frac{Z_{12}}{2Z_0}\right| \frac{e^{\alpha l}\sqrt{1 - 2e^{-2\alpha l}\cos(2\beta l) + e^{-4\alpha l}}}{2\sqrt{\alpha^2 + \beta^2}}. \tag{60}$$

Thus, the ratio of the corrected near-end to the far-end crosstalk is obtained by combining equations (60) and (19):

$$\left|\frac{N}{F}\right| = \frac{e^{\alpha l}\sqrt{1 - 2e^{-2\alpha l}\cos(2\beta l) + e^{-4\alpha l}}}{2\sqrt{(\alpha l)^2 + (\beta l)^2}}. \tag{61}$$

The curve in Fig. 8 gives the db difference between near-end and far-end crosstalk for different frequencies on a 10-mile length of two

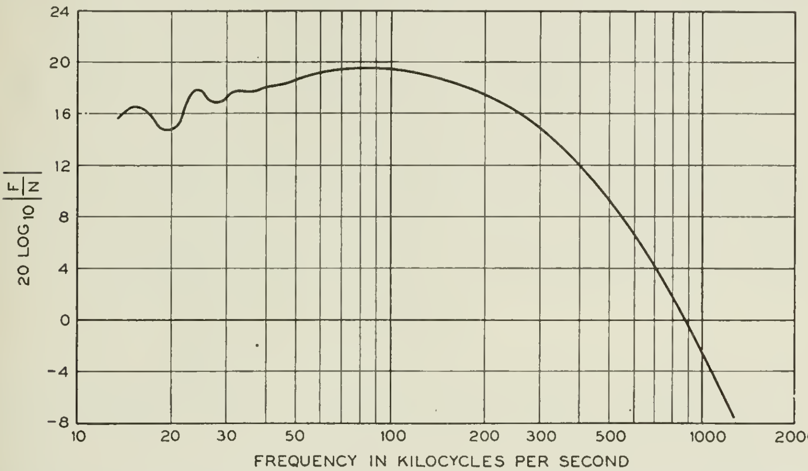


Fig. 8—Values of 20 $\log_{10} |F/N|$ for a 10 mi. repeater section of two parallel coaxial pairs in continuous contact. Coaxial pairs consist of No. 13 AWG solid copper wire, .267 in. inner diameter copper outer conductor .020 in. thick. and rubber disc insulation.

parallel coaxial pairs with hard rubber disc insulation. Each pair consists of a copper outer conductor of .267″ inner diameter and .020″ thick, and a .072″ solid copper inner conductor. It is evident that in a single repeater section far-end crosstalk is higher than near-end crosstalk up to about 900 kc.

When a number of repeater sections are connected in tandem the near-end crosstalk contribution from a single repeater section will reach the terminal of the system modified both in magnitude and in phase due to transmission through intervening sections of crosstalking circuits. At the terminal the phase changes will distribute the crosstalk from all sections in a random manner, which, in accord with both the theory and experimental evidences, will result in a root-mean-square law of addition. Thus, the overall near-end crosstalk from m sections will be equal to the crosstalk from a single section multiplied by the square root of m.

On the contrary, in a system using similar coaxial pairs transmitting in the same direction and employing repeaters at the same points, the far-end crosstalk is affected mostly by the phase differences of the repeaters. If these do not vary from the average by more than a few degrees, the far-end crosstalk in a system involving even a comparatively large number of repeaters will change proportionally to the first power of the number of repeater sections m. Only with a very large number of repeater sections (perhaps 500 or more) and random phase differences of repeaters and line of perhaps 5°–10° will the far-end crosstalk from single sections tend to approach random distribution. In this case the root-mean-square law will hold reasonably well.

Thus, far-end crosstalk will grow faster than near-end crosstalk as the number of repeater sections increases. This, combined with the relationship between the far-end and the near-end crosstalk in a single repeater section as given by equation (61) and Fig. 8, leads us to conclude that in long systems with both near- and far-end crosstalk present the limiting factor will be the far-end crosstalk. This is contrary to the experience with balanced structures stated above.

References

1. A. G. Chapman, "Open-Wire Crosstalk," *Bell System Technical Journal*, Vol. XIII, January 1934, pp. 19–58, and April 1934, pp. 195–236.
2. S. A. Schelkunoff, "The Electromagnetic Theory of Coaxial Transmission Lines and Cylindrical Shields," *Bell System Technical Journal*, Vol. XIII, October 1934, pp. 532–579.
3. L. Espenschied and M. E. Strieby, "Systems for Wide-Band Transmission Over Coaxial Lines," *Bell System Technical Journal*, Vol. XIII, October 1934, pp. 654–769.
4. E. J. Sterba and C. B. Feldman, "Transmission Lines for Short-Wave Radio Systems," *Bell System Technical Journal*, Vol. II, July 1932, pp. 411–450.
5. H. F. Mayer and E. Fisher, "Breitband Kabel mit Neuartige Isolation," *E.T.Z.* No. 46, Nov. 14, 1935.
6. S. P. Mead, "Wave Propagation Over Parallel Tubular Conductors: The Alternating Current Resistance," *Bell System Technical Journal*, pp. 327–338, April 1925.
7. John R. Carson and J. J. Gilbert, "Transmission Characteristics of the Submarine Cable," *Journal Franklin Institute*, December 1921.
8. John R. Carson and Ray S. Hoyt, "Propagation of Periodic Currents over a System of Parallel Wires," *Bell System Technical Journal*, July 1927.
9. H. Kaden, "Das Nebensprechen Zwischen Parallelen Koaxialen Leitungen," *Electrische Nachrichten Technik*, Band 13, Heft 11 (1936), pp. 389–397.
10. M. E. Strieby, "A Million-Cycle Telephone System," *Electrical Engineering*, January 1937; *Bell System Technical Journal*, January 1937.

Sound Recording on Magnetic Tape

By C. N. HICKMAN

This paper describes an improved method of recording sound magnetically on a steel tape, similar in principle to that of the Poulsen telegraphone. In the latter a longitudinal magnetic pattern of the voice current is imprinted on a steel wire by drawing it rapidly past recording pole pieces. The high speed used by Poulsen and subsequent investigators has been directly and indirectly a limiting factor in the application of magnetic recording to commercial uses. The system here described makes use of perpendicular magnetization. This method makes it possible, with suitable equalization, to obtain a substantially uniform frequency-response characteristic up to 8000 cycles per second with a tape speed of only 16 inches per second. In many cases a speed of 8 inches per second is adequate for recording speech. At the same time the ratio of signal to background noise has been substantially increased.

The decrease in efficiency resulting from the use of perpendicular instead of longitudinal magnetization is offset to a great extent by the use of a better design and construction of the pole-pieces and a more suitable recording medium. The recording medium is a steel tape having a thickness of about 1.0 to 2.0 mils (0.025 to 0.051 mm.) and a width of about 50 mils (1.3 mm.).

Introduction

A SYSTEM of recording speech magnetically on a steel wire was invented by Poulsen almost forty years ago. The wire was drawn past a pair of pole-pieces surrounded by coils carrying a speech current. A magnetic pattern corresponding to the current was thus impressed on the wire. When the wire thus magnetically treated was again drawn past the pole-pieces a current corresponding to the recording current was induced in the surrounding coils. It was common practice to place the pole-pieces on opposite sides of the wire and offset with respect to each other. The magnetic pattern in the wire thus consisted mainly of a variation in the intensity of magnetization, the direction of the magnetization being substantially parallel to the axis of the wire. This method of putting the record on the wire is known as longitudinal magnetization. With such a system the wire must travel at a very high speed if high frequencies are to be recorded and reproduced. It was customary to use speeds of from six to ten feet per second. By using tape instead of wire, the recording and reproducing pole-pieces may be placed directly opposite each other so that the magnetic pattern consists of variations in the intensity of magnetization, the direction of the magnetization being substantially perpendicular to the surface of the tape. This type of magnetization will be called perpendicular magnetization. There is another method of recording in which the magnetization is in a direction perpendicular to an edge and parallel to the surface of the tape which has been called cross or transverse magnetization.

In spite of the fact that the principle of magnetic recording has been known for a long time, there has been very little literature on the subject until recently. Several papers [1] which deal almost entirely with the longitudinal method of magnetization have been published abroad during the past two years. Cross magnetization is discussed briefly in one of the papers. Apparently, perpendicular magnetization has not been seriously considered. This paper will treat mainly the perpendicular method of magnetization with which a good frequency-response characteristic may be obtained with a tape speed of only 16 inches per second.

Forms of Recording Media

Steel wire has been used as a recording medium in most of the telegraphones. This was probably because it was easier to obtain. When wire is used it is necessary to make the longitudinal separation of the pole-pieces rather large. This is done in order to minimize the distortion caused by the continual rotation of the wire about its axis. Such rotations change the relation of the magnetic patterns in the wire with respect to the reproducing pole-pieces from that which existed at the time the record was made.

When the pole-pieces have a wide separation, high linear speed must be used in order to record and reproduce high frequencies. The high speed required in this method of recording gives rise to a number of mechanical difficulties. The contacting pole-pieces wear away rapidly and it is difficult if not impossible to construct and hold them so that they will ride smoothly against the wire. These variations in contact with the wire change the magnetic reluctance of the flux path so that the signal strength varies and an excessive amount of noise is introduced.

Recording on steel discs has been investigated from time to time but no practical results have yet been reported.

[1] See list at end of this article of recently published papers dealing with magnetic recording.

Steel tape as a recording medium was suggested by V. Poulsen in his U. S. patent No. 661,619–1900. Its use eliminates many of the objectionable features of the wire recording system. The magnetic patterns in the tape pass the pole-pieces during reproducing in the same relative positions as at the time they were made. It is practical to wind the tape on reels of pancake shape. Snarling difficulties encountered when using wire are thereby avoided. Thin tape permits the use of smaller pulleys without exceeding the bending fatigue limit of the metal. The use of tape permits the perpendicular method of magnetization to be employed. High frequencies may therefore be recorded and reproduced with a relatively low linear tape speed.

Methods of Magnetization

There are two methods of longitudinal magnetization in use, one and two pole-piece recording. A detailed description of these methods is given in two of the papers which have been mentioned. It will be sufficient here to consider them only briefly.

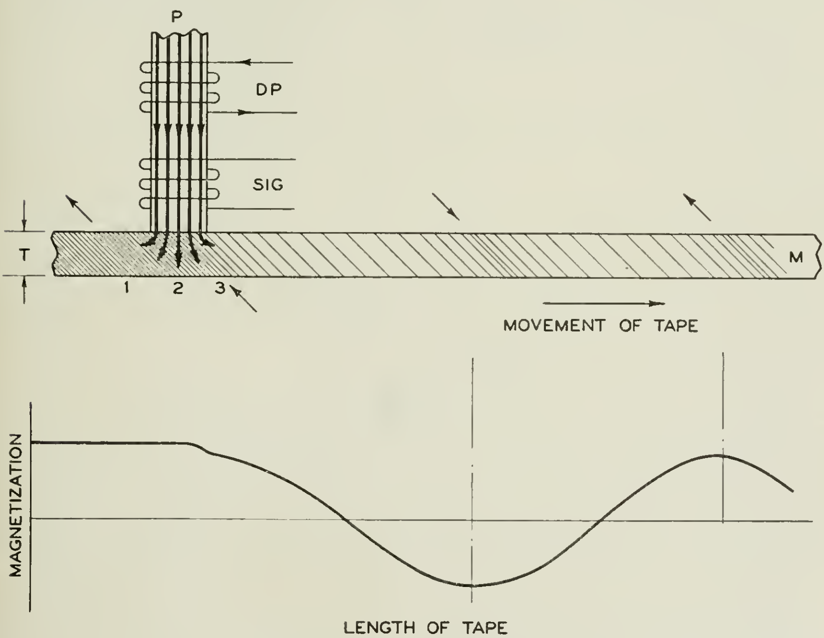


Fig. 1—Longitudinal magnetization of a recording medium by a single pole piece. SIG = Signal coil; DP = Depolarizing coil.

Figure 1 shows the action taking place in recording with one pole-piece. M is the recording medium and P is the recording pole-piece.

It will be assumed that the recording medium has been previously magnetized by drawing it past a pole-piece so that the residual magnetization in it has a direction as indicated by the upper arrow at the left.[2] In this method of recording the magnetization is principally parallel to the axis of the medium but in order to simplify the drawing, the direction of magnetization in Fig. 1 is shown at a considerable angle. If the pole-piece *P* carries a steady flux in the direction indicated by the heavy lines, this flux will spread in the medium. At the point 2 in the middle of the pole-face, the flux will be substantially perpendicular to the axis of the medium. On either side it will be approximately parallel to the axis of the medium but of opposite directions. As the elements of the recording medium approach the pole-piece *P*, they will first be subjected to the flux 1 which is in approximately the same direction as the residual magnetization in the medium so that no appreciable change will take place. When the elements are directly opposite the face of the pole-piece they will be acted on by a flux 2 which is nearly perpendicular to the residual magnetization of the medium. When the elements reach the position 3, the flux will be in opposition to the original magnetization within the medium. If it were not for these changes in the direction of the flux while the elements are passing from the position 1 to the position 3 a signal record without appreciable distortion could be left on the medium at the point 3 by superimposing a signal flux on the steady flux in the pole-piece *P*. It will be realized that the positions 1, 2 and 3 are not discrete points but that they cover an appreciable distance. The spreading of the flux at 3 will be considerable so that it will be necessary for the medium to travel at high speed in order to get the recorded signals away from the recording flux before the record is distorted by subsequent signals.

Figure 2 shows a similar diagram for two pole-piece recording. Where two pole-pieces are relatively close to each other, the flux will not spread so much in the medium and the direction of magnetization will be approximately the same as that of the recording flux. It is again assumed that the residual magnetization within the medium is mainly in the opposite direction to the motion of the medium as indicated by the upper arrow at the left. The flux 1 will have no appreciable effect on the residual magnetization. The flux 3 is in the opposite direction to the residual magnetization and were it not for

[2] In Figs. 1, 2, 3, 4 and 6, the heavy lines passing through the pole-pieces represent the instantaneous recording flux. The density of the fine lines in the recording medium represents the intensity of magnetization. The arrows above and below the medium show the direction of this magnetization. The curve below represents the nature of the signal that has been recorded on the tape.

these changes in the direction of the flux while passing from 1 to 3, a modulation of the flux 3 might be expected to leave an undistorted record on the medium as was the case with one pole piece recording. However, in the case of two pole-pieces, the elements still must pass the pole-piece P_2 where the flux 4 is approximately perpendicular to the medium. After passing the pole-piece P_2 the record is subjected to the flux 5 which is in the opposite direction to the recorded flux. The record which was made by the flux 3 is therefore distorted by fluxes 4 and 5. This distortion is greater than it is for one pole-piece

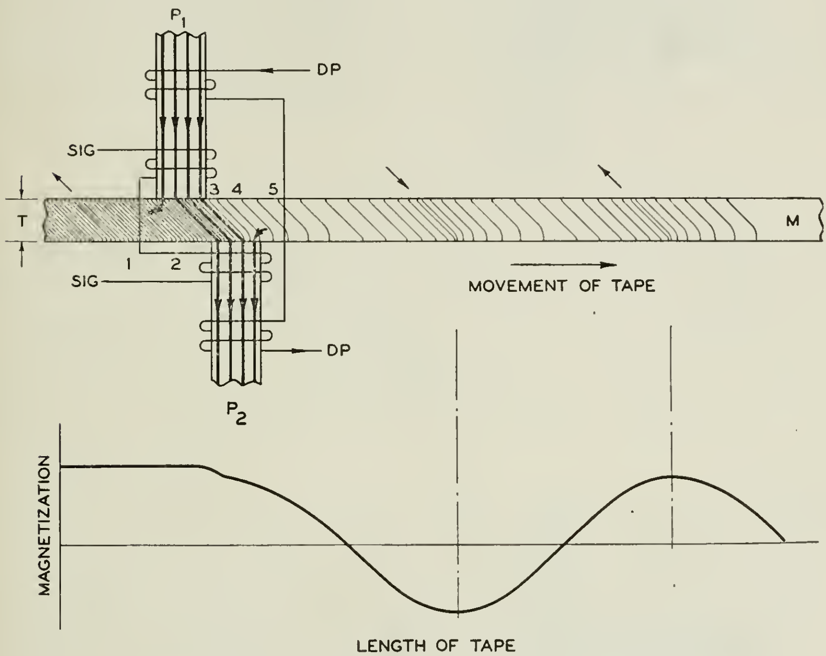


Fig. 2—Longitudinal magnetization of a recording medium by two pole pieces.

recording. In practice the stray fluxes 4 and 5 are sufficiently small so that the distortion introduced may be tolerated in exchange for the improved frequency response which is obtained with two pole-piece recording.

Figure 3 shows the action taking place where cross-magnetization is used. It is here assumed that the recording medium has been previously magnetized so that the residual magnetization is in the direction indicated by the upper arrows at the left. W represents the width of the recording medium which in this case is a steel tape. It will readily be seen that if W is very large there will be considerable spread-

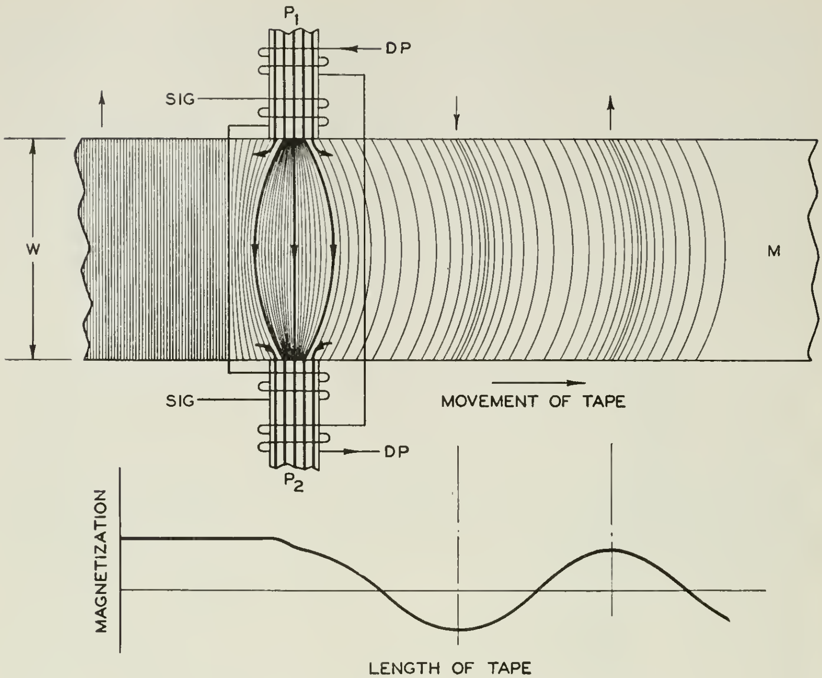


Fig. 3—Transverse magnetization of a steel tape.

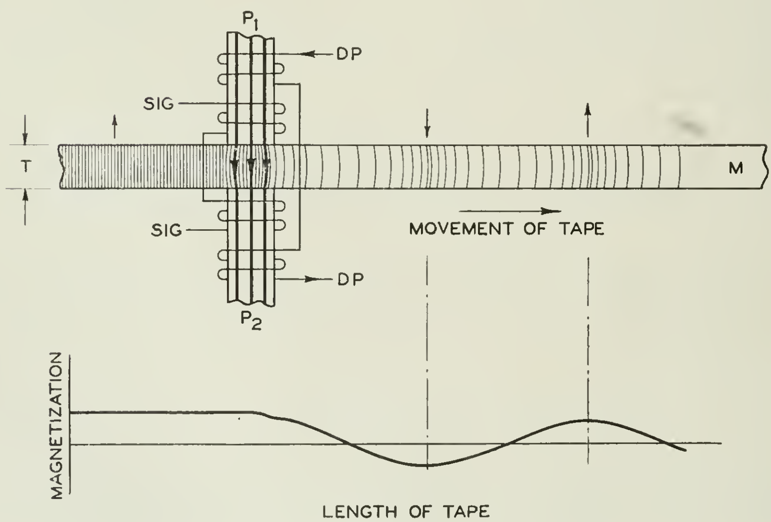


Fig. 4—Perpendicular magnetization of a steel tape.

ing of the recording flux within the tape. The recording flux is always at substantially right angles to the axis of the tape and parallel to its surface and is in the opposite direction to the residual magnetization.

If W is made quite small or in other words if the pole-pieces P_1 and P_2 are directly opposite each other with the thin dimension of the tape between them, we have the conditions shown in Fig. 4. The tape is so thin that there is very little spreading of the flux so that the width of the flux path is not appreciably dependent on the strength of the signal. This type of recording is called perpendicular magnetization in order to distinguish it from cross-magnetization, where the width of the tape instead of the thickness determines the pole-piece separation. The perpendicular method of magnetization permits a relatively low tape speed. The thickness of the pole-piece tips determines the frequency response for a given tape speed.

Method of Recording with Perpendicular Magnetization

If the tape is first subjected to a saturation flux which is at right angles to the surface of the tape, it will be left with one side of north and the other of south polarity. If the tape in this condition is passed between recording pole-pieces carrying only AC flux, it is obvious that only half cycles will be recorded. The record is therefore much distorted. The current reproduced from such a record is similar to the alternating current which may be obtained from a single wave rectifier.

If on the other hand the tape is passed through an alternating high-frequency field which is strong enough to erase the record, it is left in a substantially neutral condition. If it is then passed between the recording pole-pieces, both half cycles will be recorded but there will be amplitude distortion. Figure 5 shows a magnetization curve for iron which has previously been demagnetized with alternating current. The slope of the first part of the curve is small in either direction of magnetization and then increases with increase in the flux and finally becomes smaller again. Small signals will therefore be recorded weakly and strong signals will be recorded relatively higher. Both will have wave form distortion. The same effects would be obtained with longitudinal or cross magnetization. In the past, investigators have often utilized only one side of the magnetization curve. A direct current was used as a bias to bring the recording flux to the most suitable part of the curve such as at n, Fig. 5.

The method employed here will be made clear from Figs. 6 and 7. As the tape elements enter the field of the polarizing pole-pieces P_1, P_2 (Fig. 6) they are subjected to an increasing magnetizing force. The

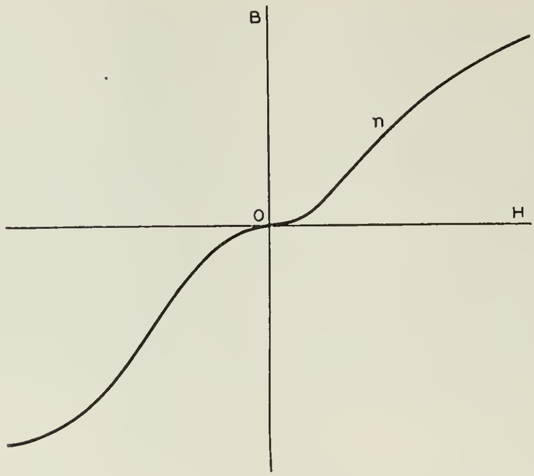


Fig. 5—Typical magnetization curve for steel.

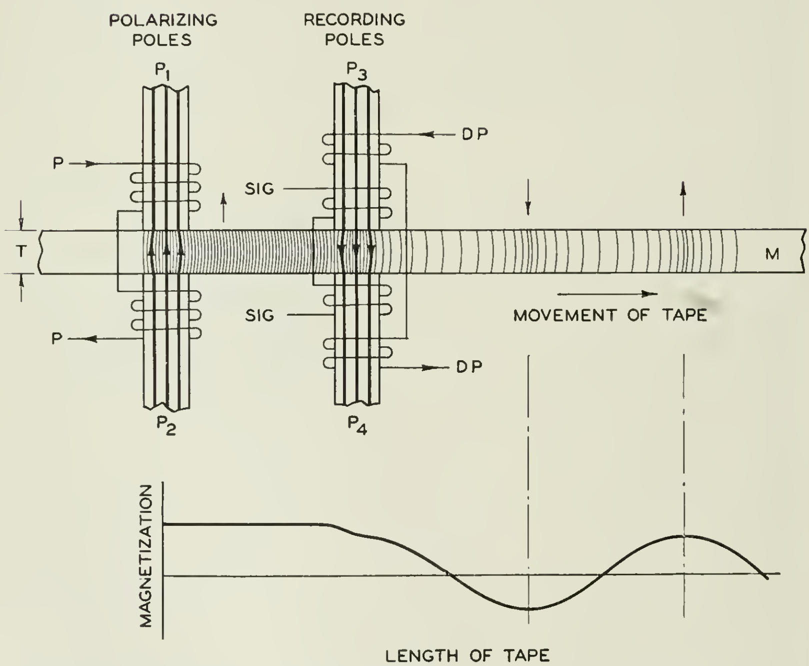


Fig. 6—Perpendicular magnetization method of recording on steel tape. SIG = Signal coil; P = Polarizing coil; DP = Depolarizing coil.

elements are magnetized to the saturation point P as shown by curve a, Fig. 7. As the elements leave the polarizing field they are subjected to a field of decreasing strength so that the magnetic induction drops along the curve **b** to R, this point being reached when the applied field is zero. In Fig. 7, the magnetizing force H refers to the externally applied field. The tape elements then pass between the recording pole-pieces which carry a flux in opposite direction to that of the polarizing pole-pieces. If there is no signal current present, the magnetic induction will be brought down to the point N by the biasing field. As the elements pass out from between the pole-pieces, the field will decrease to zero and the magnetic induction will change from

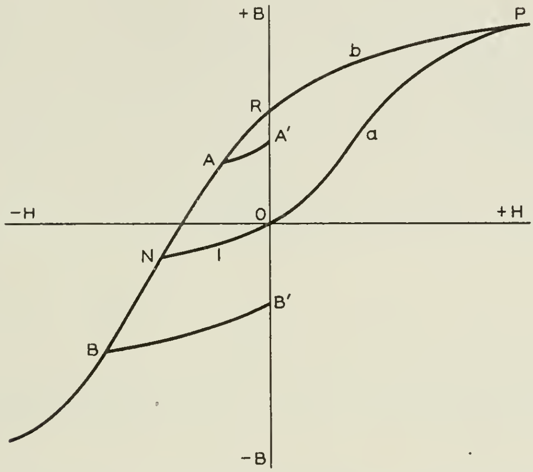


Fig. 7—Diagram showing the cycles of magnetization through which the elements of a steel tape may pass during the process of recording.

N to O, which is a substantially neutral condition. However, if there is a signal current present at the time the tape elements are passing between the recording pole-pieces, the magnetization will be reduced to a point A higher than N if the cycle is in opposition to the bias flux or to the point B lower than N if the signal flux is in the same direction as the bias flux. In either case the elements will retain a magnetization value corresponding to A' or B' respectively. This system makes it possible to record over a longer portion of the magnetization curve without appreciable distortion.

Unless the proper value of biasing field is used to bring the magnetization approximately to the point N when no voice current is present, the maximum recording range cannot be obtained without excessive amplitude distortion. For example if no bias is used, it has been found that the output plotted against input will be too steep as shown in the curve 1, Fig. 8. If the bias is too great, the output input curve will be inclined too much as shown in curve 2 of the same figure. When the

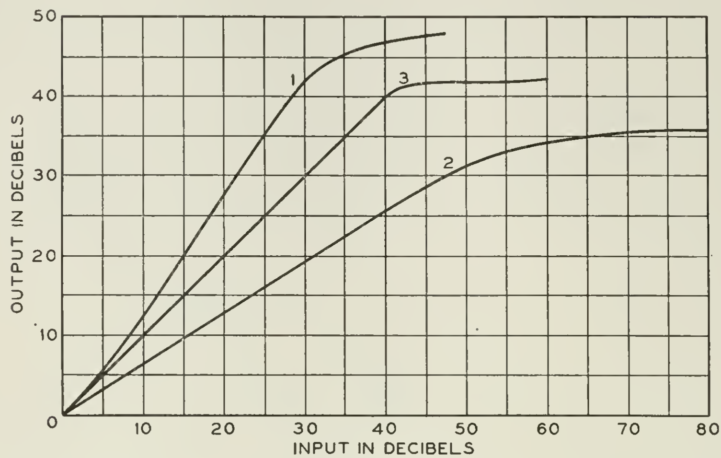


Fig. 8—The effect of the biasing current on the slope of the output-input level curves.

proper value of bias is used the curve 3 is obtained which is inclined at 45 degrees. Measurements of output versus input may therefore be used to determine the proper bias current.

Another method of determining the proper amount of bias is to plot the output obtained from records of very weak signals as a function of

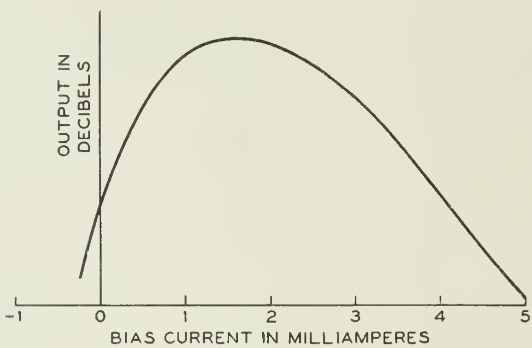


Fig. 9—Effect of the biasing current on the intensity of weak signals.

the bias current used during recording. Figure 9 shows such a curve for a 1000-cycle record. In selecting the proper bias current from the curve, the point at the crest gives the most efficient value but it is better to favor some point slightly to the right of the crest in order to avoid the danger of very strong signals operating too far down on the left side of the curve. The recording current for the 1000 cycles had a value 30 db below the overload so that only a small amplitude was used in obtaining the data for this curve.

If the same set of pole-pieces is used both for recording and reproducing, there is no question of getting the reproducing pole-pieces in the same alignment as the recording pole-pieces. If different sets are used care must be taken to get the same alignment.

In order to keep the signal high above the tape noise, it is desirable to record so that the flux or current amplitude is independent of the frequency. Since the impedance of the recording coils rises rather sharply with frequency, it is necessary either to place a high resistance in series with the recording coils or to connect them to a high impedance. The later method is of course the more efficient. Since the energy present in the higher frequencies of voice and music is usually less than in the 1000-cycle region, an amplifier having a rising characteristic may be used in order to record these frequencies at a higher level. A corrective network may be used in reproducing to obtain the desired frequency response. Such a procedure increases the apparent ratio of signal to the tape noise.

Reproducing

If the thickness of the pole-piece tips is small with respect to the wave length of the signal on the tape, the voltage generated in the reproducing coils is proportional to frequency, so that the coils may be matched to favor the lower frequencies. If a straight line frequency characteristic is desired, a corrective network may be used.

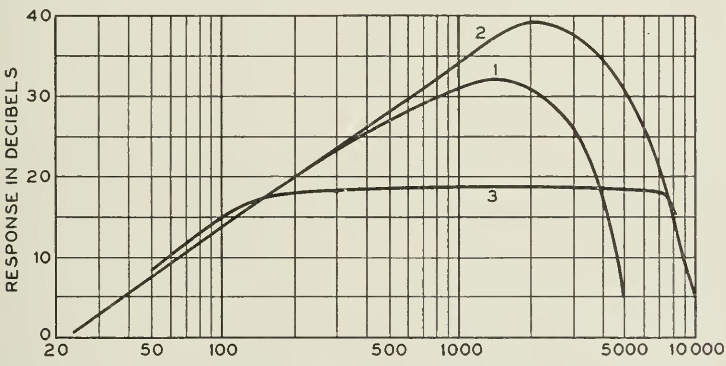


Fig. 10—Frequency response curves obtained with the perpendicular method of magnetization.

Figure 10 shows several frequency response curves. Curve 1 shows a response without the use of equalization for a tape speed of 8 inches per second. Curve 2 shows the response under the same conditions for a tape speed of 16 inches per second. Curve 3 shows the same signals of curve 2 reproduced through a suitable equalizer.

The ratio of the maximum reasonably undistorted 1000-cycle signal to the noise with typical good tape is about 38 db. The transfer loss is approximately 60 to 70 db. The maximum power required in recording is about 0.3 milliwatt.

Characteristics of Magnetic Recording

Magnetic recording differs from other methods in several respects. Since no processing is required, the record may be reproduced without a long delay. The recording medium may be used over and over again for new records. It is only necessary to subject the tape to a strong magnetic field in order to obliterate a record. The obliteration is conveniently done at the same time that the new record is being made. Where temporary records are desired, magnetic recording therefore has some advantages over other methods. On the other hand it should be fully appreciated that the records may be kept, filed away, or reproduced thousands of times with no appreciable deterioration in the quality.

The magnetic system is very convenient for use where short delays are desired. A short loop of tape in conjunction with recording, reproducing, and obliterating pole-pieces is all that is required. Instead of a loop of tape, a disc or cylinder rotating at high speed may be used to carry the recording medium. The latter method makes it possible to obtain very short delays. Where perpendicular magnetization is used, very long records may be obtained from a medium which occupies a relatively small amount of space. For example, a thin coil of 2 mil tape 9 inches in diameter will give a playing time of 1/2 hour with a tape speed of 16″ per second.

There are no moving parts in the modulating unit. The difficulties of obtaining high frequencies due to the inertia of the cutting stylus in mechanical recording are therefore not present. The system is subject to the same difficulties of eliminating flutter that we find in other methods of recording; however, mechanical vibrations due to the motor and other moving parts of the recording system do not have to be filtered out as is the case with mechanical recording. There are of course no shavings. The recording medium cannot be easily scratched and may be handled in any kind of light and subjected to large variations in temperature. When properly wound on reels it is not liable to breakage or damage during transportation.

LITERATURE

Stille, K., "Die Electromagnetische Schallaufzeichnung," *E. T. Z.*, 1930, p. 449.
Hormann, E., "Zur Theorie der Magnetischen Tonaufzeichnung," "On The Theory of Magnetic Sound Recording," *E. N. T.*, Vol. 9, pp. 388–403, Oct., 1932.
Meyer, E. and Schuller, E., "Magnetische Schallaufzeichnung auf Stählbander," "Magnetic Sound Recording on Steel Ribbons," *Zeits. f. Techn. Physik*, Vol. 13, No. 12, pp. 593–599, 1932.
Begun, S., "Die Diktiermaschine im Grossbetrieb," *E. T. Z.*, 1932, p. 204.
Hickman, C. N., "Delayed Speech," *Bell Labs. Record*, Vol. 11, pp. 308–310, June, 1933.
"Magnetic Recording and Reproducing," The Marconi-Stille apparatus described, *Wireless World*, Vol. 34, pp. 8–10, Jan. 5, 1934.
Rust, N. M., "Marconi-Stille Recording and Reproducing Equipment," *Marconi Rev.*, No. 46, pp. 1–11, Jan.–Feb., 1934.
Frederick, H. A., "Recording and Reproducing Sound," *Rev. Sci. Instruments*, Vol. 5, pp. 177–182, May, 1934.
Braunmühl, H. J. von, "Magnetische Schallaufzeichnung im Rundfunkbetrieb," "Magnetic Sound Recording in Broadcasting Service," *Funktechnische Monatshefte*, No. 12, pp. 483–486, Dec., 1934.
Mallina, R. F., "A Mirror for the Voice," *Bell Labs. Record*, Vol. 13, pp. 200–202, Mar., 1935.
Volk, T., "Magnetophon, ein Neues Tonaufzeichnungsgerät," "Magnetophone, a New Sound Recording Apparatus," *Filmtechnik*, Vol. 11, pp. 229–231, Oct. 26, 1935.
Hansen, W. H., "Das Magnetophon," "The Magnetophone," *E. T. Z.*, Vol. 56, p. 1232, Nov. 7, 1935.
Schuller, E., "Magnetische Schallaufzeichnung," "Magnetic Sound Recording," *E. T. Z.*, Vol. 56, pp. 1219–1221, Nov. 7, 1935.
Schrage, W. F., "Sound Recording on Magnetic Materials," *Radio Craft*, Vol. 7, pp. 537–562, March, 1936.

Constant Resistance Networks with Applications to Filter Groups

By E. L. NORTON

The problem investigated is the determination of two finite networks such that, when connected in parallel, they will have a constant resistance at all frequencies. The admittance of any network may be written as the ratio of two polynomials in frequency. A network to be one of a constant resistance pair must have certain restrictions imposed on its admittance. In case the two networks are both filters of negligible dissipation, the expression for the input conductance of each may be written from a knowledge of the required loss characteristic.

The poles of the expression for the conductance are then found. They will be identical for the two networks. The networks are then built up by synthesis from those poles of the conductance which have negative real parts, these corresponding to real network elements.

The methods which have been developed for this last process are described in detail.

ONE of the most useful principles available to the network design engineer is that of constant resistance networks. The use of these networks is widespread in the telephone system for purposes of loss equalization and distortion correction, where they have the advantage of providing a means for altering the transmission properties

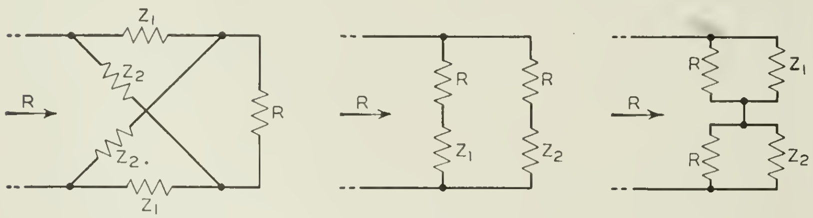


Fig. 1—The three fundamental forms of constant resistance networks.

of a circuit without affecting its impedance.[1] The three usual types of constant resistance networks are shown in Fig. 1, where, in all cases, $Z_1Z_2 = R^2$, a relationship which is always possible to fulfill if

[1] "Distortion Correction in Electrical Circuits with Constant Resistance Recurrent Networks," Otto J. Zobel, *Bell Sys. Tech. Jour.*, July 1928.

Z_1 and Z_2 are built up of resistive and reactive elements in the usual way.

The lattice type will not be considered here. The first step in extending the other two is shown in Fig. 2, where the networks shown

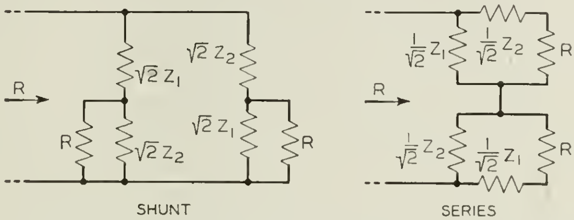


Fig. 2—The first step in extending the fundamental forms of the constant resistance networks.

have a constant resistance if $Z_1Z_2 = R^2$. The networks have now taken on the form of two half-section filters in parallel or series, provided that Z_1 and Z_2 are purely reactive. This suggests the possibility of an extension to more complicated configurations having the general properties of wave filters with constant resistance. Since the shunt and series types are analytically the same, only the former will be considered in detail.

Use will be made of the following theorem:

Any finite network of linear elements having a constant conductance at all frequencies, and no purely reactive shunt across its terminals, has zero susceptance.

The admittance may be written [2]

$$Y(\lambda) = \frac{A_0 + A_1\lambda + \cdots + A_m\lambda^m}{B_0 + B_1\lambda + \cdots + B_n\lambda^n},$$

where $\lambda = i(\omega/\omega_0)$ and m is equal to or one greater or one less than n. ω_0 is a constant which fixes the frequency scale. If the real part of Y is to be a constant other than zero, A_0 cannot be zero and m must be equal to or greater than n. If there is no purely reactive shunt across the terminals, B_0 cannot be zero and m cannot be greater than n. The expression for the admittance may then be written

$$Y(\lambda) = G\frac{1 + A_1\lambda + \cdots + A_n\lambda^n}{1 + B_1\lambda + \cdots + B_n\lambda^n}.$$

[2] See "Synthesis of a Finite Two Terminal Network Whose Driving Point Impedance is a Prescribed Function of Frequency," Otto Brune, *M. I. T. Journal of Mathematics and Physics*, vol. 10, 1931.

By elementary methods it may be shown that if the real part of this expression is constant for all real frequencies then $A_1 = B_1$, $\cdots A_n = B_n$, and the imaginary part is zero. All other possibilities involve special relations between the B's, which correspond to a $Y(\lambda)$ with poles on the imaginary axis. This has been excluded by the condition of no purely reactive shunt across the terminals. The study of networks of constant admittance may then be restricted to the study of the conditions for constant conductance.

We will consider, then, the problem of designing two passive networks of linear elements such that, when connected in parallel, they will have constant conductance. The value of the constant conductance may be taken as unity without loss of generality.

The conductance of a finite network may be written as a ratio of two polynomials in frequency. Its value must always be positive for real frequencies, and for the case under consideration it may never exceed unity, since otherwise the conductance of the second network to make up the constant resistance pair would be required to be negative. The expression for the conductance of the first network may be written in the form

$$G_1 = \frac{1}{1 + F(\lambda)}, \tag{1}$$

where λ may be $i(\omega/\omega_0)$ as above, or it may be taken as any imaginary function of frequency which may be realized by the impedance of a reactive network, and $F(\lambda)$ is the ratio of two polynomials in even powers of λ. By subtracting G_1 from unity the required expression for G_2 may be obtained:

$$G_2 = \frac{1}{1 + \dfrac{1}{F(\lambda)}}. \tag{2}$$

An investigation of general networks of an arbitrary number of resistance and reactance elements fulfilling the relations (1) and (2) would take the present investigation too far from its main objective. If the networks are to have the general properties of wave filters with a minimum of loss in a band, they may be restricted to reactive networks having a single resistance. Furthermore, both resistances may be taken as unity, for, in cases where this is not necessary, a transformation to some other value may be made after the design is completed on the unit resistance basis. We assume, too, that when $\lambda = 0$, $F(\lambda) = 0$ and $G_1 = 1$, $G_2 = 0$. This implies the proper choice of the expression for λ.

With a voltage E_0 applied to the common terminals the power absorbed by the first network is $E_0^2G_1$ and by the second is $E_0^2G_2$. Since in both cases the power delivered to the network must be absorbed in the single resistance, the two insertion losses are given by

$$e^{-2\alpha_1} = G_1 = \frac{1}{1 + F(\lambda)}, \tag{3}$$

$$e^{-2\alpha_2} = G_2 = \frac{1}{1 + \dfrac{1}{F(\lambda)}}. \tag{4}$$

Since $F(\lambda)$ must be an even function of λ, the poles of (3) may be written $\pm c_m \pm id_n$ and (3) is

$$e^{-2\alpha_1} = -D^2 \frac{1}{\lambda + c_0} \frac{1}{\lambda - c_0} \prod_{1}^{m=(n-1)/2} \frac{1}{\lambda + c_m \pm id_m} \frac{1}{\lambda - c_m \pm id_m} \tag{5}$$

if the degree of $F(\lambda)$ is $2n$. If n is even, the terms $\lambda + c_0$ and $\lambda - c_0$ are omitted and the product taken from $m = 1$ to $m = n/2$. The quantity D^2 is the denominator of $F(\lambda)$, a polynomial in λ^2.

Let β_1 be the phase angle between E_0 and the voltage E_1 across the resistance in the first network. The left side of equation (3) may then be factored in the form $e^{-2\alpha_1} = e^{-(\alpha_1+i\beta_1)}e^{-(\alpha_1-i\beta_1)}$. Similarly half of the factors on the right belong with $e^{-(\alpha_1+i\beta_1)}$ and half with $e^{-(\alpha_1-i\beta_1)}$. Now the terms with poles having a negative real part[3] must belong with $e^{-(\alpha_1+i\beta_1)}$ so that:

$$\begin{aligned} e^{-(\alpha_1+i\beta_1)} &= D \frac{1}{\lambda + c_0} \Pi \frac{1}{\lambda + c_m \pm id_m} \\ &= D \frac{1}{\lambda + c_0} \Pi \frac{1}{c_m^2 + d_m^2 + \lambda^2 + 2c_m\lambda}. \end{aligned} \tag{6}$$

Since λ is an imaginary function of frequency, say $\lambda = ix$, and D is real if λ or x is real, the phase β_1 is given by

$$\beta_1 = \tan^{-1}\frac{x}{c_0} + \sum_{m=1}^{m=(n-1)/2} \tan^{-1}\frac{2c_m x}{c_m^2 + d_m^2 - x^2}. \tag{7}$$

If n is even the expression is

$$\beta_1 = \sum_{m=1}^{m=n/2} \tan^{-1}\frac{2c_m x}{c_m^2 + d_m^2 - x^2}. \tag{7a}$$

[3] These being the factors that correspond to physically realizable network elements, they belong with the physically realizable factor of the exponent.

The network can be designed from equation (6) or by making use of both (3) and (7). Both methods will be illustrated in two types of networks giving filter characteristics.

Filters with Characteristics Similar to the "Constant K" Type of Filter

As the simplest form of $F(\lambda)$ take $F(\lambda) = [(\lambda)/(i)]^{2n}$. The poles of (3) are then simply the $2n$ roots of $(-1)^{n-1}$, which may be written $\pm \cos (m\pi/n) \pm i \sin (m\pi/n)$ if n is odd, and $\pm \cos [(2m-1)\pi/2n] \pm i \sin [(2m-1)\pi/2n]$ when n is even. In the first case m varies between zero and $(n-1)/2$ and in the second case between unity and $n/2$. For the case of n being odd, equation (6) may then be written

$$e^{-(\alpha_1+i\beta_1)} = \frac{1}{1+\lambda} \prod_{m=1}^{m=(n-1)/2} \frac{1}{1+\lambda^2+2\cos\dfrac{m\pi}{n}\lambda} \tag{8}$$

where the polynomial D is unity in this case. The last equation expanded is in the form

$$e^{\alpha_1+i\beta_1} = 1 + A_1\lambda + A_2\lambda^2 + \cdots + A_n\lambda^n, \tag{9}$$

which is the form for the ratio E_0/E_1 for the network shown in Fig. 3. By writing out the ratio E_0/E_1 for this network and comparing terms with equation (8) expanded in the form of (9) the values of the a's may be found to be [4]

$$\begin{aligned}
a_1 &= \sin\frac{\pi}{2n}, \\
a_2 &= \frac{\sin\dfrac{3\pi}{2n}\sin\dfrac{\pi}{2n}}{a_1\cos^2\dfrac{\pi}{2n}}, \\
&\cdots\cdots\cdots\cdots \\
a_m &= \frac{\sin\dfrac{2m-1}{2n}\pi\sin\dfrac{2m-3}{2n}\pi}{a_{m-1}\cos^2\dfrac{m-1}{2n}\pi}, \\
a_n &= n\sin\frac{\pi}{2n}\cdot
\end{aligned} \tag{10}$$

[4] By the evaluation of the finite sums and products of the trigonometric terms. No short method has been found for obtaining the results.

The second network, which when connected in parallel with Fig. 3 will give a constant resistance, is obtained from the first by replacing λ by $1/\lambda$. It is shown in Fig. 4.

These structures have been designed on the basis of λ being a pure imaginary. Note, however, that the two structures will have a constant resistance provided that λ is any function realizable by a combination of resistances and reactances. Equations (8) and (9) will still hold but (3) and (7) will no longer be true. Note, too, that for the simplest case of $n = 1$ the structures reduce to the usual form for constant resistance networks as shown in Fig. 1.

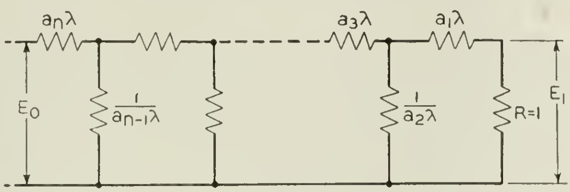


Fig. 3

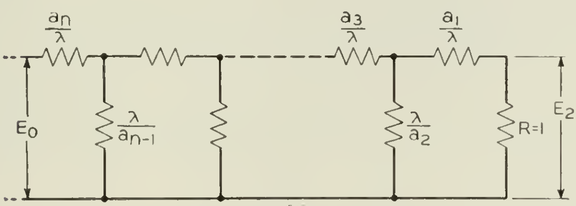


Fig. 4

Figs. 3–4—A pair of constant resistance networks of "constant K" configuration.

If λ is taken of the form $i(f/f_0)$, the structure of Fig. 3 will be made up of series coils and shunt condensers in the form of a low-pass filter. The structure of Fig. 4 will be of the form of a high-pass filter with series condensers and shunt inductances. The loss of the first network is

$$e^{2\alpha_1} = 1 + \left(\frac{f}{f_0}\right)^{2n}$$

and of the second

$$e^{2\alpha_2} = 1 + \left(\frac{f_0}{f}\right)^{2n}.$$

With $f < f_0$ the loss of the first network will be small and the loss of the second network large. With $f > f_0$ the reverse is true. At $f = f_0$ each of the networks takes half of the available power, illustrating a necessary property of constant resistance networks of this type, of a three db loss at the cross-over frequency.

If λ is taken of the form $i\dfrac{\left(\dfrac{f}{f_m}-\dfrac{f_m}{f}\right)}{\left(\dfrac{f_2}{f_m}-\dfrac{f_m}{f_2}\right)}$ the networks become band-pass and band-elimination filters, respectively. By taking other functions for λ multiple band structures may be designed, subject always to the limitation that the combined bands of both filters must extend over the whole frequency range, with a three db loss at each cross-over point.

The evaluation of the elements is easily done from equations (10). The impedance denoted by $a_1\lambda$, for example, in the low-pass filter would have the value $i(f/f_0)\sin(\pi/2n)$, which is an inductance of a value $(1/2\pi f_0)[\sin(\pi/2n)]$. For a terminating resistance different from unity the value of the first inductance is $L_1 = (R_0/2\pi f_0)[\sin(\pi/2n)]$ or in general any inductance is $L_m = a_m L_0$ where $L_0 = R_0/2\pi f_0$. Similarly, any capacitance is $C_m = a_m C_0$ where $C_0 = 1/2\pi f_0 R_0$. The corresponding formulas for the second network are $C_m = C_0/a_m$ and $L_m = L_0/a_m$. The same formulas hold for n even; in that case the networks of Fig. 3 and Fig. 4 would terminate on the right in a shunt arm with impedances of $1/a_1\lambda$ and λ/a_1, respectively. This is illustrated by Fig. 2 for $n = 2$.

Filters with Characteristics Similar to Those of the "M-derived" Type

The networks shown in Fig. 3 and Fig. 4 have the same configuration and similar characteristics to constant K filters. They are subject to the same objection of a relatively slow rate of cut-off and an excessive loss at frequencies remote from the cut-off. A type of characteristic similar to that obtained with M-derived filters, with points of infinite loss at finite frequencies, is necessary for an economical design in the majority of cases.

The loss characteristic of the network is of course fixed by the function $F(\lambda)$, a ratio of two polynomials in λ. It may be written

$$F(\lambda) = A_0\lambda^2\frac{1 + A_1\lambda^2 + \cdots + A_n\lambda^{2n-2}}{1 + B_1\lambda^2 + \cdots + B_n\lambda^{2n-2}}.$$

Now the first filter will have infinite loss points when the denominator is zero, and the second filter when the numerator is zero. If these peaks are to occur at real frequencies, $F(\lambda)$ must have poles at $\lambda^2 = -1/S_m^2$ and zeros at $\lambda^2 = -P_m^2$. Moreover, since $1/[1 + F(\lambda)]$ and $1/[1 + (1/F(\lambda)]$ must always be positive for real frequencies, the expression for $F(\lambda)$ when all its zeros and poles occur at real frequencies must be a perfect square. It may then be written

$$F(\lambda) = A_0\lambda^2 \frac{(P_1^2 + \lambda^2)^2 \cdots (P_{n-1}^2 + \lambda^2)^2}{(1 + S_1^2\lambda^2)^2 \cdots (1 + S^2_{(n-1)/2}\lambda^2)^2}.$$

In order to get an idea of the significance of the expression, let $\lambda = i(f/f_0)$ and restrict the P's and the S's to values less than unity. The first network will then have zero loss points at $f = 0$ and $f = P_m f_c$ and infinite loss points at $f = f_0/S_m$ and $f = \infty$. The second network will have infinite loss points when the loss of the first is zero, and zero loss points when the loss of the first is infinite. The first network is therefore a low-pass filter and the second a high-pass filter.

The following work is considerably simplified if $S_m = P_m$. This implies that the characteristic of the second filter is the same function of $1/\lambda$ that the first is of λ. If the cross-over point is fixed at $\lambda^2 = -1$, the value of A_0 is -1 and in order to write equation (6) or (7), it is necessary to find those zeros with a negative real part of

$$1 - \lambda^2 \frac{(P_1^2 + \lambda^2)^2 \cdots (P^2_{(n-1)/2} + \lambda^2)^2}{(1 + P_1^2\lambda^2)^2 \cdots (1 + P^2_{(n-1)/2}\lambda^2)^2}$$
$$= \left[1 + \lambda \frac{(P_1^2 + \lambda^2) \cdots}{(1 + P_1^2\lambda^2) \cdots}\right]\left[1 - \lambda \frac{(P_1^2 + \lambda^2) \cdots}{(1 + P_1^2\lambda^2) \cdots}\right].$$

Now since the zeros of the second factor on the right are the negatives of the zeros of the first factor, it will be sufficient to find all of the zeros of the first factor and reverse the signs when necessary to secure negative real parts. Consider, then, the equation

$$1 + \lambda \frac{(P_1^2 + \lambda^2) \cdots (P^2_{(n-1)/2} + \lambda^2)}{(1 + P_1^2\lambda^2) \cdots (1 + P^2_{(n-1)/2}\lambda^2)} = 0.$$

One root is $\lambda = -1$. It may be shown further that the magnitude of all of the roots is unity. Writing $\lambda = \rho e^{i\theta}$ as a root, the magnitude of the typical product term $(P^2 + \lambda^2)/(1 + P^2\lambda^2)$ may be written

$$\left|\frac{P^2 + \lambda^2}{1 + P^2\lambda^2}\right|^2 = 1 + \frac{\left(\frac{1}{P^2} - P^2\right)\left(\rho^2 - \frac{1}{\rho^2}\right)}{\left(P\rho + \frac{1}{P\rho}\right)^2 - 4\sin^2\theta}.$$

Now since the denominator of the expression on the right is always positive, and all of the P's are less than unity, the magnitude of each of the product terms is greater than unity if ρ is greater than unity and less than unity if ρ is less than unity. Since, however, the magnitude of the complete product must be unity, the value of ρ must be unity.

After dividing through by the factor $1 + \lambda$, the remaining function is a reciprocal equation in λ and may be written as an equation in $p = \lambda + (1/\lambda)$. Since the magnitudes of the roots in λ are all unity, the roots in p must all be real and be in the region $-2, +2$.

The degree of the polynomial in p is $(n - 1)/2$. It may be shown further that if $(n - 1)/2$ is even there are an equal number of positive and negative real roots, if the degree is odd there is one more positive than negative root.

The equations in p for various values of $(n - 1)/2$ are

$$\begin{aligned}
\frac{n-1}{2} &= 1, \qquad p - (1 - \Sigma_1) = 0 \\
&= 2, \qquad p^2 - (1 - \Sigma_2)p - (1 - \Sigma_1 + \Sigma_2) = 0 \\
&= 3, \qquad p^3 - (1 - \Sigma_3)p^2 - (2 - \Sigma_1 + \Sigma_3)p \\
&\qquad\qquad + (1 - \Sigma_1 + \Sigma_2 - \Sigma_3) = 0 \\
&= 4, \qquad p^4 - (1 - \Sigma_4)p^3 - (3 - \Sigma_1 + \Sigma_4)p^2 \\
&\qquad\qquad + (2 - \Sigma_1 + \Sigma_3 - 2\Sigma_4)p \\
&\qquad\qquad + (1 - \Sigma_1 + \Sigma_2 - \Sigma_3 + \Sigma_4) = 0,
\end{aligned}$$

where the Σ's are the symmetric functions of the P^2's, that is,

$$\begin{aligned}
\Sigma_1 &= P_1^2 + P_2^2 + \cdots P^2_{(n-1)/2}, \\
\Sigma_2 &= P_1^2 P_2^2 + \cdots + P^2_{(n-3)/2} P^2_{(n-1)/2}.
\end{aligned}$$

The equations in p may also be written in trigonometric form as follows:

$$\begin{aligned}
\frac{n-1}{2} &= 1, \qquad \cos\frac{3}{2}\theta + \Sigma_1 \cos\frac{\theta}{2} = 0 \\
&= 2, \qquad \cos\frac{5}{2}\theta + \Sigma_2 \cos\frac{3}{2}\theta + \Sigma_1 \cos\frac{\theta}{2} = 0 \\
&= 3, \qquad \cos\frac{7}{2}\theta + \Sigma_3 \cos\frac{5}{2}\theta + \Sigma_1 \cos\frac{3}{2}\theta + \Sigma_2 \cos\frac{\theta}{2} = 0 \\
&= 4, \qquad \cos\frac{9}{2}\theta + \Sigma_4 \cos\frac{7}{2}\theta + \Sigma_1 \cos\frac{5}{2}\theta + \Sigma_3 \cos\frac{3}{2}\theta \\
&\qquad\qquad + \Sigma_2 \cos\frac{\theta}{2} = 0.
\end{aligned}$$

These equations include the root at $\lambda = -1$ corresponding to $\theta = \pi$. Excluding this they will each have $(n-1)/2$ roots between $\theta = 0$ and $\theta = \pi$. The roots in p will then be given by $p = 2\cos\theta$.

Equation (7) becomes

$$\beta_1 = \tan^{-1} x + \sum_{1}^{(n-1)/2} \tan^{-1} \frac{p_m x}{1 - x^2}, \tag{11}$$

where the quantities p_m are the roots of the above equations, without regard to sign.

We require also the value of $d\beta_1/dx$, which may be written

$$\frac{d\beta_1}{dx} = \frac{1}{1 + x^2}\left[1 + \sum_{1}^{(n-1)/2} \frac{p_m}{1 - \dfrac{(4 - p_m{}^2)x^2}{(1 + x^2)^2}}\right]. \tag{12}$$

A possible configuration for the first network is shown in Fig. 5 and for the second in Fig. 6.

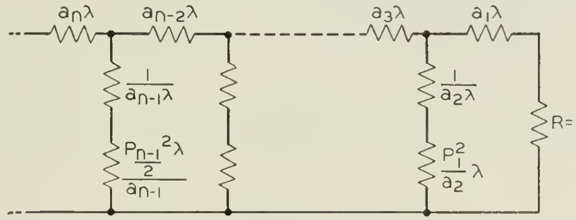


Fig. 5

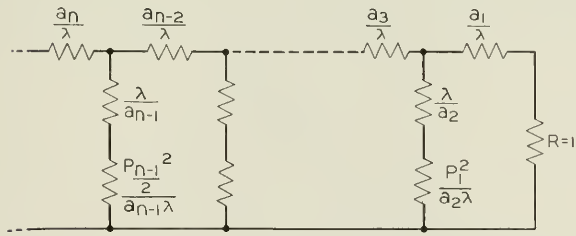


Fig. 6

Figs. 5–6—A pair of constant resistance networks of the "M-derived" configuration.

To find the elements it would be possible to expand the voltage ratio and solve for the a's as was done in the constant K illustration. Another method would be to find the input admittance of the network from the known input conductance, and find the a's from this expression. A simpler method, however, takes advantage of the fact that each structure is a purely reactive network with the exception of the terminating resistance and finds the network elements in terms of the short circuit reactance as measured from the resistance end of the network.

Use may be made of the following theorem:

With any four-terminal reactive network the reactance measured at terminals 3–4 with terminals 1–2 short-circuited is equal to the tangent of the phase shift between a voltage E_0 applied to terminals 1–2 and the resultant voltage E_1 across a unit resistance connected to terminals 3–4.

The open-circuit voltage across 3–4 due to E_0 would be $\pm kE_0$, where k is a real quantity, if the network contains only reactances. By Thévenin's Theorem, then,

$$E_1 = \frac{\pm kE_0}{1 + iX},$$

where X is the reactance of the network from terminals 3–4. If β is the phase shift between E_0 and E_1, $X = \tan \beta$.

Since this phase shift is given by (11) the short-circuit reactance is known. At a value of $\lambda = i(1/P_1)$ or $x = 1/P_1$, the impedance of the first shunt arm from the right of Fig. 5 is zero, so that the reactance of the filter is simply the reactance of the arm $a_1\lambda$, which gives the value of a_1 directly as

$$a_1 = P_1(\tan \beta_1)_1,$$

where $(\tan \beta_1)_1$ denotes the value of $\tan \beta_1$ when $x = 1/P_1$. The reactance of the network after subtracting a_1x is $\tan \beta_1 - P_1(\tan \beta_1)_1x$. At values of x very close to $1/P_1$ this is the reactance of the first shunt arm, or

$$a_2 = \frac{\dfrac{d}{dx}\left(P_1^2x - \dfrac{1}{x}\right)}{\dfrac{d}{dx}(\tan \beta_1 - a_1x)},$$

where, after differentiation, $x = 1/P_1$. Carrying through the differentiation,

$$a_2 = \frac{2P_1^2}{(1 + \tan^2 \beta_1)_1 \left(\dfrac{d\beta_1}{dx}\right)_1 - a_1}.$$

Similar formulas may be found for the rest of the elements. If X_m denotes the reactance starting with the series arm $a_m\lambda$ or with the shunt arm $(P_{m/2}^2\lambda^2 + 1)/a_m\lambda$, then for m odd, that is, for a series element,

$$a_m = P_{(m+1)/2}X_m \qquad \left(x = \frac{1}{P_{(m+1)/2}}\right)$$

and for a shunt arm, m even,

$$\frac{1}{a_m} = \frac{1}{2P_{m/2}^2}\frac{dX_m}{dx} \qquad \left(x = \frac{1}{P_{m/2}}\right).$$

When $m = n$, or for the last series arm, a special relation is necessary, readily obtained by the limiting value of reactance as x approaches zero. This gives

$$(a_1 + a_3 + \cdots + a_n)x = (1 + \Sigma p_m)x$$

or

$$a_n = 1 + \Sigma p_m - (a_1 + a_3 + \cdots + a_{n-2}).$$

To use these relations it is necessary to know the expression for X_m, the reactance to the left from the successive points in the network. To determine this in terms of the elements already known use may be made of the following theorem:

If the impedance looking to the left into a network is Z, the impedance to the left from A, any point within the network is the negative of the impedance to the right from A when the network is terminated on the right by an impedance $-Z$.

For example, referring to Fig. 5, to determine a_3 it is necessary to know the reactance to the left starting with a_3x. By the theorem this is

$$X_3 = \frac{1}{\dfrac{a_2x}{1 - P_1^2x^2} - \dfrac{1}{a_1x - \tan\beta_1}}$$

and when $x = 1/P_2$ we have for a_3

$$\frac{1}{a_3} = \frac{a_2}{P_2^2 - P_1^2} - \frac{1}{a_1 - P_2(\tan\beta_1)_2}.$$

The impedance at that end of the filter terminated by the resistance is of interest. Its value of course depends upon the terminating impedance at the junction of the two filters, but assuming that this impedance and the separate terminating resistances are all R_0, the impedance from the load of the first filter is $R_0 \tanh(\alpha_2 + i\beta_2)$ if terminated in a series arm and $R_0 \coth(\alpha_2 + i\beta_2)$ if terminated in a shunt arm. Note that the impedance of the first filter depends upon the transfer constant of the second. The impedance from the load of the second filter depends in the same way upon the transfer constant of the first. The proof of these relations is based upon both networks being purely reactive.

Applications

The use of the constant resistance pairs of filters is indicated wherever the impedance at the junction of two filters is of major importance. Another application which is of some importance is that of separating the energy in a band of frequencies into two or more channels, delivering all of the energy into one or the other of the loads.

The method may be extended to more than two networks in parallel or series to give a constant resistance. For example, the combination

$$G_1 = \frac{1}{[1 + F_1(\lambda)]\left[1 + \dfrac{1}{F_2(\lambda)}\right]},$$

$$G_2 = \frac{1}{\left[1 + \dfrac{1}{F_1(\lambda)}\right]\left[1 + \dfrac{1}{F_2(\lambda)}\right]},$$

$$G_3 = \frac{1}{1 + F_2(\lambda)},$$

will give a constant resistance for the three networks. Designs have been carried through on this basis where the networks are low-pass, high-pass and band-pass, respectively. This is one method of avoiding the limit of three db in the loss of the low-pass and the high-pass filters at their cross-over point, since in this case the band-pass filter will take up the power. A second method is to use a pair of low and high-pass filters, each terminated in another pair with different cross-over points. This method requires the use of both a low-pass and a high-pass filter as power absorbing networks but they would be simple structures and together would require no more elements than the single band-pass filter in a three-filter combination.

The two methods are illustrated in Fig. 7 and Fig. 8, respectively. The structures for the second type are given by Fig. 9. Note that the filters designated L.P. II, L.P. III, H.P. II and H.P. III have one series arm missing and are apparently terminated at a shunt point at the load end of the filter. This is a consequence of selecting the two P's in such a way that the coefficient a_1 becomes zero, a matter of no particular difficulty in the case of a two-section filter.

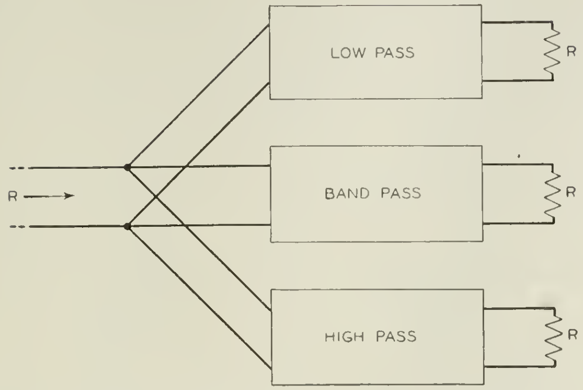


Fig. 7—A three-filter constant resistance combination.

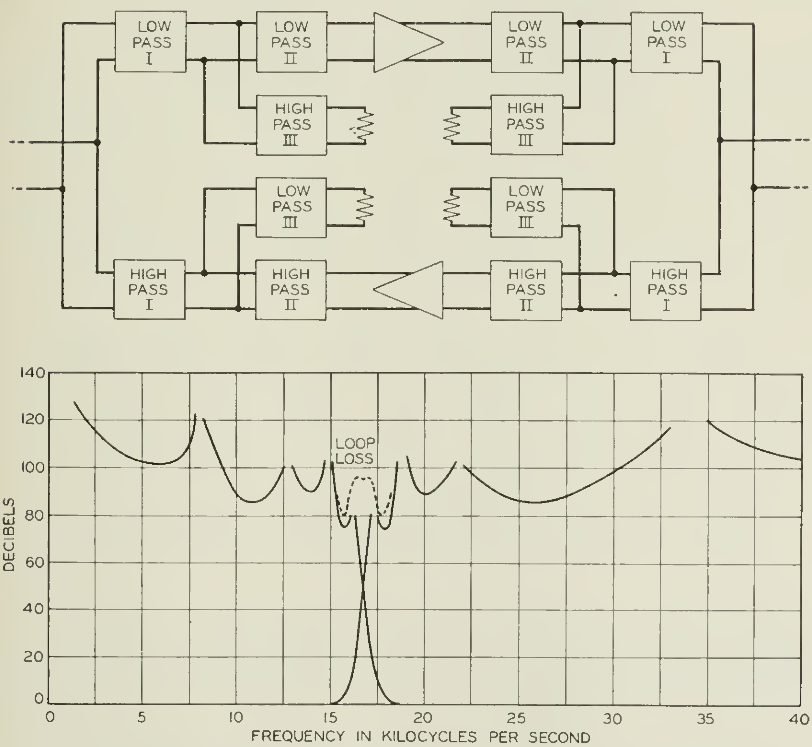


Fig. 8—Constant resistance networks used as directional filters.

It will be found that a filter of several sections of the type described in this paper will have somewhat less loss in the attenuated band than the usual type of design. On the other hand the loss in the band will, in general, be less unless additional elements are used in the standard type of filter to reduce reflection losses. A design for a pair of constant

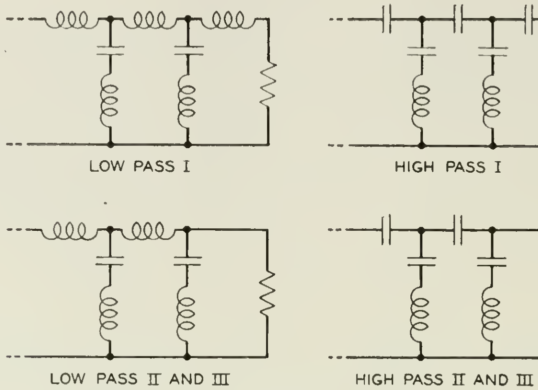


Fig. 9—The configuration of the filters of Fig. 8.

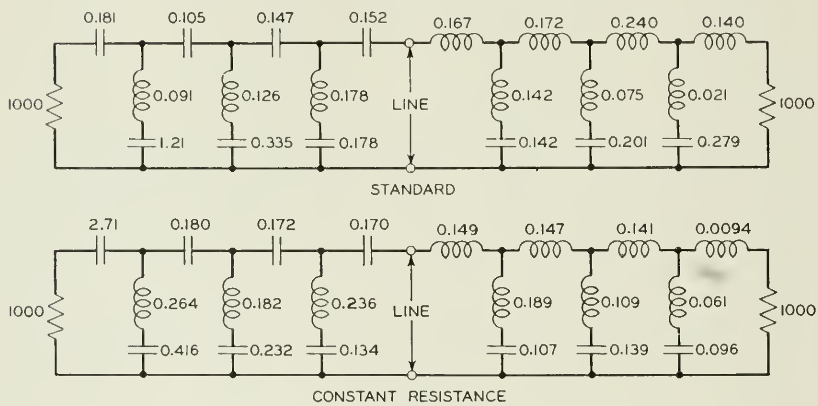


Fig. 10—Comparison of the elements of typical standard and constant resistance filters.

resistance filters having a cross-over frequency of 1000 cycles is compared with a design for a pair of standard filters in Fig. 10. No additional elements have been added to the standard type to improve the impedance.[5] The loss characteristics for the two low-pass filters

[5] "Impedance Correction of Wave Filters," E. B. Payne, and "A Method of Impedance Correction," H. W. Bode, *Bell Sys. Tech. Jour.*, October 1930.
"Extensions to the Theory and Design of Electric Wave Filters," Otto J. Zobel, *Bell Sys. Tech. Jour.*, April 1931.

are compared in Fig. 11. Note that in this case the constant resistance filters have only about sixty per cent of the loss of the standard

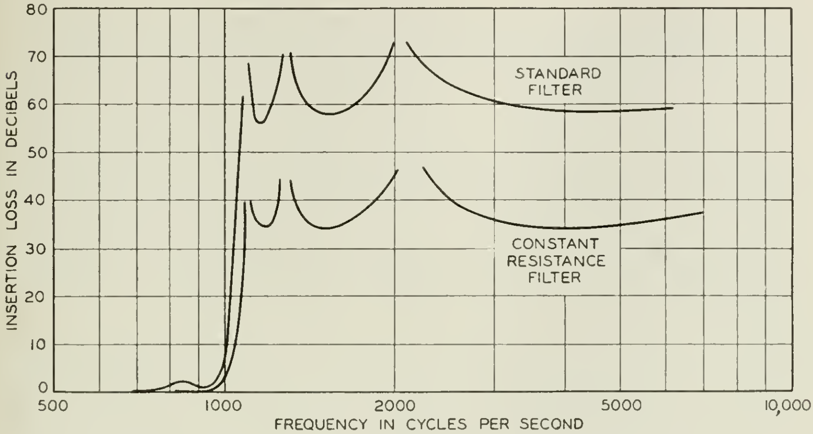


Fig. 11—Loss characteristics obtained by the filters of Fig. 10.

filters. The difference would not be as great for filters of less sharp discrimination.

A Laboratory Evaluation of Wood Preservatives

By R. E. WATERMAN, JOHN LEUTRITZ and CALEB M. HILL

Evolution of a simple laboratory technique for the assay of materials proposed for use in the preservation of wood is reported in this paper. This test involves a measurement of the actual decay resistance of the treated wood. Included are a resumé of the limitations imposed by current test-methods and a discussion of the adaptations of this new technique to the numerous variables inherent in laboratory simulations of outdoor exposure.

FUNDAMENTAL scientific discoveries in the biological sciences during the latter part of the nineteenth century slowly brought organized knowledge out of the chaos of conflicting theories as to the character of many natural phenomena. This was especially true in the field of fermentation where these accumulated findings and observations finally served as the basis for the proof that the filamentous fungi were the causal agents in the decay of wood. This knowledge of the decay mechanism, together with increased demand for wood products due to industrial expansion and the concomitant depletion of our best stands of naturally rot-resistant species of timber, served as a stimulus towards organized studies of the physiology of decay organisms and possible means of prophylaxis.

While Nature has been lavish in the supply of fast-growing species, she has also been provident in making such timber more vulnerable to attack by the micro-flora and fauna which act as scavengers for the forests and as conservators for vast quantities of materials which trees take from the soil during their growth period. The necessity of preserving this more easily decayed wood accelerated the search for satisfactory means of protection. This has been especially true in the Bell System where fast-growing but easily rotted southern pine has, to a large extent, been supplanting chestnut and cedar for poles.

In the past the use of certain materials for the preservation of wood was based entirely on their availability or the personal prejudice of proponents for them. This method of selection could result only in widespread waste and oftentimes disastrous consequences, but a wood-preserving industry utilizing certain materials such as coal-tar creosote gradually evolved. The controversies as to what properties of creosote make it an effective preservative still rage, and the problem of choosing and specifying the type of creosote best fitted for the preservation of timber is still urgent. This is particularly vital in that creosote is a loose term covering a congregation of compounds rarely twice the same in quality or proportion.

When in 1927, research on the development of a rapid means of evaluating wood preservatives was initiated in the Chemical Department of these Laboratories, primary consideration was accorded the selection of the best available method for measuring the toxicity of proposed preservatives against wood-destroying fungi. The technique used was one which had been developed and extended to a considerable degree by the workers at the Forest Products Laboratory in Madison, Wisconsin. This petri dish method, described at some length by Richards in 1923,[1] was standardized in 1929 [2] at a conference of American workers in St. Louis. Briefly, the method consists in adding various amounts of the toxic agent under test to a nutrient medium in the form of a hot malt-agar solution which is poured into a petri dish, cooled, and the resulting gel inoculated with small sections of the hyphæ of a wood-destroying fungus (Fig. 1). The organism usually used is culture no. 517 from the Forest Products Laboratory but others may be chosen. The excellence of the preservative is based on the lowest concentration which is able to kill or totally inhibit growth of the test organism.

While the petri dish method can be brought to a high degree of efficiency, accuracy and precision by suitable precautions, it is definitely limited in its practical application. It tells nothing of the permanency of the material under test from the standpoint of leaching, evaporation or chemical instability. Nothing is learned of the possible reaction of the preservative with wood, and the dispersion in warm liquid agar which later gels is a far cry from that obtained in wood. There is an axiom of biological assay, that the substratum for *in vitro* tests be as similar as possible to that encountered in nature. Neglect of this principle in the field of antiseptics and germicides has been responsible for many outstanding failures *in vivo* of materials which had given brilliant promise in the culture tube. Doubt concerning the validity of the petri dish test was substantiated when several preservatives highly toxic according to this method failed in outdoor exposure tests. In many cases such failures could not be ascribed to obvious conditions such as high volatility or solubility. Instances were also met wherein materials of little value according to the petri dish method were able to prevent decay in the field. There is no disposition to advise against all use of this method as it is a valuable tool in making initial judgments on a new material; but it should be verified by other means before the expense of a field trial can be justified, and a

[1] Numbers refer to bibliography at end.

proper degree of skepticism should be exercised before condemning a preservative on the basis of this test alone.

Parallel with the use of nutrient substrata of the malt-agar type in this country, there grew up in Europe a technique which utilized the wood itself as a medium for dispersion of the toxic agent. This kolle flask method (Fig. 2) was standardized and accepted by a conference

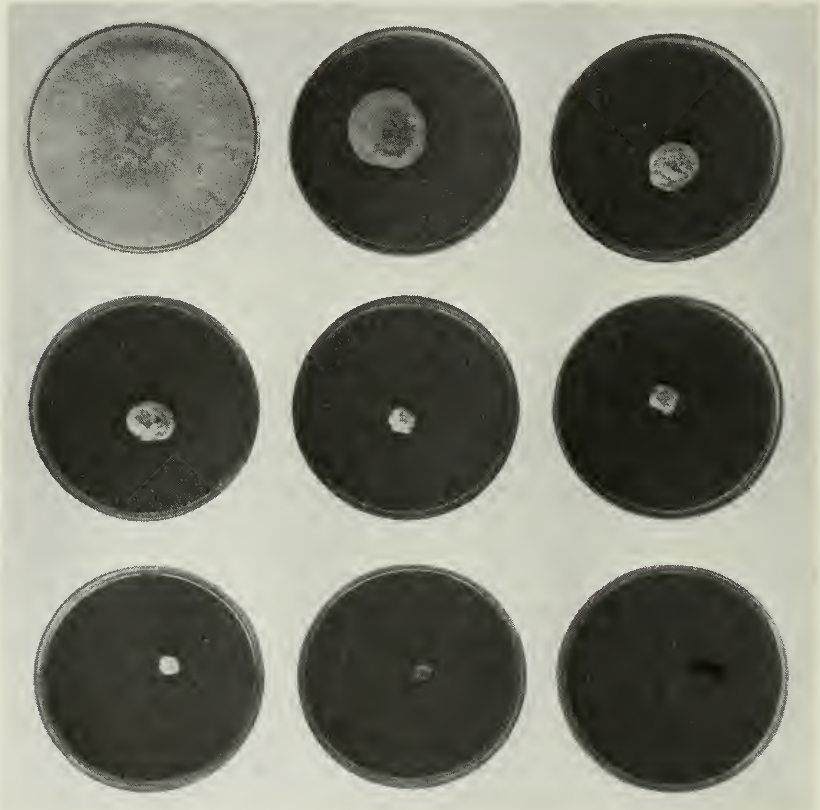


Fig. 1—Assay by petri dish method. Test-fungus No. 517 on increasing amounts of coal-tar creosote.

of European workers at Berlin in 1930.[3] An outline of the method follows: Wood blocks of a convenient size are impregnated with the toxic agent, usually in solution, and after evaporation of the solvent the blocks are placed in kolle flasks and supported on glass rods set in malt-agar covered with the actively growing mycelia of the test fungus. The conference advised the use of *Coniophora cerebella* as the test fungus, but suggested that at least two species should be used in each test. After three or four months' exposure to the wood-destroying fungi, the blocks are removed from the flasks, freed from adhering mycelium, and the weights taken before and after the test period used as a measure of the amount of decay.

The kolle flask method has much to recommend it, overcoming as it does many of the difficulties inherent in the petri dish technique. However, the test as standardized at Berlin presents serious drawbacks. The kolle flasks are expensive, comparatively fragile, difficult to

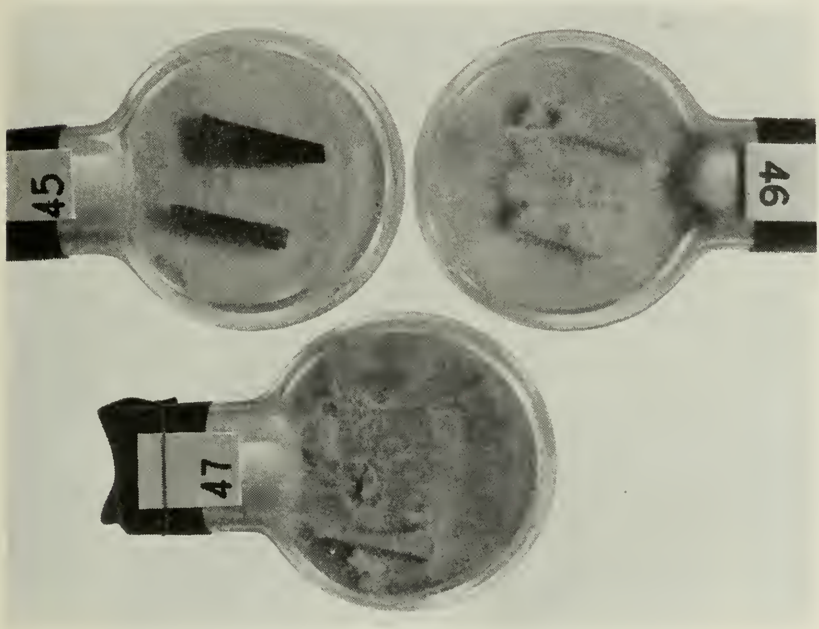


Fig. 2—Assay by kolle flask method. *Poria incrassata* used in comparison of southern pine heartwood versus sapwood.

handle, inconvenient to store, and maintenance of proper moisture conditions is particularly difficult. For really satisfactory results the flasks during the tests should be kept at a constant humidity and temperature. Despite all precautions there is the ever-present danger of excess moisture and resultant lack of decay should the test block touch the agar or any condensed moisture on the flasks. Another unusual problem arose when certain over-ambitious fungi rotted the cotton plugs used to stopper the flasks and even continued to grow into other flasks where they did not belong.

A New Assay Method

Both the petri dish and kolle flask methods had shown definite limitations, and it became apparent that further experimentation on a laboratory assay-method should be directed along somewhat different lines. By chance a few treated pieces which had been removed unscathed after a routine exposure in the kolle flask, were dropped on a beakerful of moist wood heavily infected with a wood-destroying fungus. The beaker was merely covered with a watch-glass and set aside. Growth progressed over the treated blocks with unexpected rapidity and vigor, and when removed at the end of three months, the specimens were found to be severely decayed. Occasional results of this character were so encouraging that efforts were renewed to

Fig. 3—Apparatus required for modified wood-block method of assay.

develop a technique which would incorporate the use of wood as a secondary substrate together with more favorable moisture control. Experimentation had demonstrated that the amount of moisture in the wood block should be slightly above fibre saturation for optimum growth of the fungus. Inoculated wood placed in air at 100 per cent relative humidity will rot but slowly while too much moisture decelerates and even inactivates the fungal metabolism. The problem therefore was to bring about these optimum decay conditions with low-priced, easily handled equipment. The test in its present state of evolution is inexpensive, easy to manipulate and capable of increased uniformity due to better regulated moisture conditions. The only equipment necessary consists of a straight-sided screw-capped bottle about 5 inches high and 2 inches in diameter, a smaller bottle 2.5 inches high and 1 inch in diameter, a wad of cotton, a small flat piece of untreated wood and an applicator such as is used by the medical profession for swabs (Fig. 3).

The treated blocks, previously brought to moisture equilibrium, are supported by means of a thin slab of untreated wood on the top of the small bottle which is placed inside the larger screw-topped bottle.

Fig. 4—Assembly of apparatus required for modified wood-block method of assay.

Through holes bored in the test piece and the thin slab of supporting wood are passed the pieces of wooden applicator, which act as a means of anchorage and as wicks for conduction of water to the wood under test. Although not absolutely necessary, cotton is usually wrapped around the small bottle to reduce shock during handling. Water is placed in both bottles and after sterilization of the complete set-up (Fig. 4) the thin slab of wood is inoculated with a portion of hyphæ of the test-fungus which has been growing on a malt-agar substratum. The bottles are then placed in an incubation room (Fig. 5) at 26–28° C., customarily for a period of 24 weeks. At the end

Fig. 5—Incubator with tests in progress.

of the test period the blocks, freed of adhering mycelium, are again brought to equilibrium at a specified humidity, reweighed and the pieces finally dissected to determine the loss of strength occasioned by the attack of the fungus.

Materials to be tested as possible preservatives are injected in serial concentrations into the blocks of sapwood, commonly southern pine, under conditions simulating as nearly as possible those which would

Fig. 6—Constant humidity chamber filled with test blocks.

be used in practice. Due to the variance in the moisture pick-up of wood at different relative humidities, it is necessary to bring the blocks to equilibrium under standard conditions both before and after the test in order to determine actual weight losses. Since oven-drying is obviously a poor reference standard when volatile materials are under consideration and may also bring about serious changes in the wood, a constant humidity chamber is used for this purpose. An ordinary bacteriological incubator kept at 30° C., fitted with slow-moving fans and a shallow pan containing a saturated solution of common table salt (Fig. 6), has proved to be completely satisfactory in this respect, maintaining a relative humidity of 76 per cent with very little deviation. The test pieces after treatment are placed on racks (Fig. 7) and only a few days in the chamber are necessary for equilibration.

Such a test method allows of three criteria as the basis for judging the degree of attack. First, there is the amount and vigor of the growth of the test-fungus on the wood block, readings of which are made every four weeks. As this is difficult of expression, recourse is had to the classical method of the serologists, wherein plus four denotes the maximum. An attempt is made to evaluate both vigor and extent

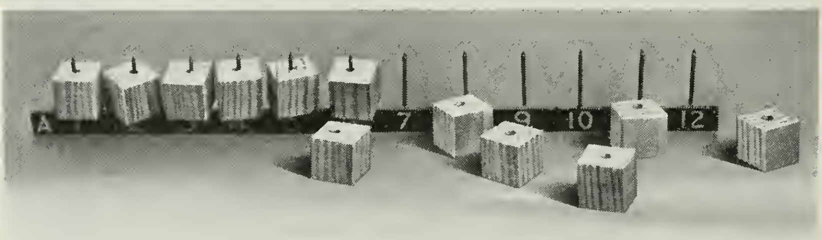


Fig. 7—Test blocks and rack.

of the fungal growth; the notation "2–4" would mean that the test block was partly covered with a heavy mycelial mat, whereas "3–2" would mean almost covered with relatively weak growth (Figs. 8 and 9). Often no growth occurs on the test piece and sometimes the mycelial inoculum is actually killed. The second measure of extent of decay is based on the loss of weight, with corrections for the effect of leaching and evaporation during the test period, computed from data obtained on treated controls put through the entire cycle without inoculation. These controls are also of value in the empirical strength rating made at the end of the test when the pieces are dissected in an effort to judge the remaining strength—a rating of ten indicates no detectable loss in strength as compared to the control and zero denotes complete disintegration.

Long experience with the petri dish method had emphasized the high degree of specificity of the fungi to various toxics. No single fungus is equally resistant to all preservatives and gross errors are inevitable unless cognizance is taken of this situation. Since it is impossible to use all the organisms which destroy wood, a choice has been made to include genera which are known to be of considerable

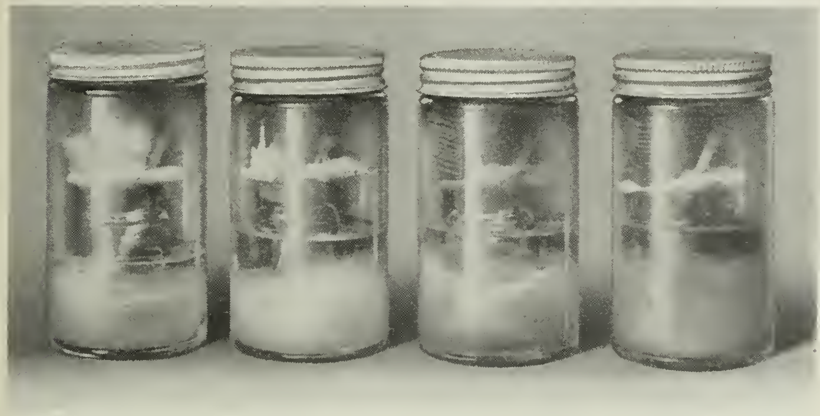


Fig. 8—Assay of a polychlor phenol showing effect of increasing concentration. The test organism *Lentinus lepideus*. The growth ratings from left to right are 4–4, 3–3, 1–2, and √ = no growth on specimen.

Fig. 9—Same range of concentration as in Figure 8 with U-10 as test fungus. The growth ratings from left to right are 4–4, 3–4, 1–3 and 1–2.

economic importance in the decay of timber or which have been encountered in the actual decay of telephone poles, being sure to include in any given test the fungi which past experience has shown to be resistant to the type of preservative under consideration. Four organisms in duplicate are used in each test (Fig. 10). *Lentinus lepideus*, cited by Buller,[4] Snell [5] and Humphrey [6] and isolated several times from posts in the Gulfport, Mississippi, test plot,[7] as well as from poles in service, is used in all cases of organic preservatives, but is seldom used against metallic salts, to which it is extremely sensitive. *Lenzites trabea*, another species of great economic importance, Hubert,[8] and also isolated several times from rotted southern pine poles, is somewhat parallel in resistance to *Lentinus lepideus*, but produces a markedly different type of decay. *Polyporus vaporarius*, *Poria incrassata*, and *Coniophora cerebella*, the common "dry rots," although easily killed by many hydrocarbons, are resistant to most inorganic compounds, and at least one of these organisms is included in each test on such materials. *Fomes roseus*, another fungus of wide distribution, reacts in a most inconsistent manner, but its occasional specific virulence is sufficient to warrant its inclusion in all assays of

Fig. 10.—Assay of worthless preservative at maximum concentration. The fungi in duplicate from left to right are *Lenzites trabea*, U-10, *Fomes roseus* and *Lentinus lepideus*.

new and unusual preservatives. Unfortunately the fastest and most versatile decay organism used has no name and masquerades under the designation U (unknown)-10. Isolated several years ago from a decayed pine pole, the identity of U-10 is still a mystery, despite the efforts of many mycological authorities. U-10 is included in every test and is especially valuable when a quick indication of the value of a new preservative is needed, as it is capable of producing an appreciable weight loss in about three months. In addition to the above fungi occasional use is made of such common wood-destroyers as *Trametes serialis*, *Lenzites sepiaria*, *Polystictus versicolor*, *Polyporus sulphureus*, and *Fomes pinicola*.

At the present stage of development this wood block method tells nothing directly about the ability of a wood preservative to resist the action of termites. Most materials which inhibit decay also prevent termite attack. In addition all promising leads are verified by means of the sapling test [9] at Gulfport, Mississippi, and here Nature has provided a bountiful supply of these industrious insects.

Fig. 11—Growth on southern pine sapwood controls. From left to right; *Polyporus sulphureus*, *Polyporus vaporarius*, *Polystictus hirsutus* and U-10.

Fig. 12—Growth on southern pine sapwood controls. *Lenzites trabea*, *Fomes roseus*, U-10 and *Lentinus lepideus*.

Experimental Results

The above-mentioned fungi have been tested on untreated southern pine sapwood in order to set up standards of comparison (Figs. 11 and 12). Table I shows individual weight and strength losses effected by our most commonly used organisms in the regular 24-week period. Considerable decay with many of these fungi occurs in a somewhat

TABLE I

Weight Losses and Dissection Ratings on Untreated Southern Pine Blocks Exposed to Most Common Test Fungi for Twenty-Four Weeks

Organism	Initial Weight in Grams		Final Weight in Grams		Loss in Per Cent	Empirical Rating Based on Dissection	Description of Decay
	76% Relative Humidity	Oven Dry	76% Relative Humidity	Oven Dry	Based on Oven-Dry Weights		
Lentinus lepideus	2.34	2.05	1.87	1.64	20.0	4	Rather advanced decay throughout
	2.35	2.06	1.87	1.64	20.4	4	–ditto–
	2.25	1.97	1.85	1.62	17.8	5	Moderately advanced decay throughout
	2.26	1.98	1.77	1.55	21.7	4	Rather advanced decay throughout
Fomes roseus	2.25	1.97	1.84	1.61	18.3	4	Rather advanced decay throughout
	2.27	1.99	1.84	1.61	19.1	4	–ditto–
	2.22	1.95	1.78	1.56	20.0	4	–ditto–
	2.21	1.94	1.56	1.37	29.4	1	Thoroughly rotted
U–10	2.33	2.04	1.13	0.99	51.5	0	Complete disintegration
	2.33	2.04	1.37	1.20	41.2	0	–ditto–
	2.20	1.93	1.08	0.95	50.8	0	–ditto–
	2.18	1.91	1.19	1.04	45.5	0	–ditto–
Lenzites trabea	2.32	2.03	1.73	1.52	25.1	2	Deep surface disintegration, advanced decay elsewhere
	2.33	2.04	1.76	1.54	24.5	3	Deep surface disintegration, rather advanced decay elsewhere
	2.07	1.81	1.51	1.32	27.1	2	Advanced decay
	2.09	1.83	1.29	1.13	38.3	1	Thoroughly rotted
Polyporus vaporarius	2.12	1.86	1.69	1.48	20.4	4	Rather advanced decay throughout
	2.01	1.76	1.74	1.53	13.1	6	Mild decay throughout
	2.18	1.91	1.78	1.56	18.3	4	Rather advanced decay throughout
	2.26	1.98	1.82	1.60	19.2	4	–ditto–

Growth rating in all cases was 4–4, signifying that the test blocks were covered with heavy normal growth shorter time, but experience has shown that more consistent results are obtained with the longer period. This is especially true in the case of materials of moderate toxicity, wherein often little growth is seen for two or three months, after which the fungus may become established and rot the test piece.

Using the routine technique, volatile compounds are often found to be practically worthless. This is true of naphthalene, for instance, but when special precautions are taken to insure its presence during the exposure to the fungus, the effective toxicity of this hydrocarbon cannot be questioned. Analysis of control blocks proved that the naphthalene had evaporated quite completely from the test piece even before sterilization. This difficulty can be surmounted satisfactorily in the case of a single compound by injecting a generous quantity and determining the actual amounts of material present from equilibrium weights of the test pieces before treatment and just prior to inoculation. Steam sterilization, of course, would introduce errors under such circumstances, and while the risk of contamination with foreign organisms is high, satisfactory results have been obtained with unsterilized blocks. In the case of volatile mixtures a similar procedure permits the knowledge of the total evaporation before inoculation, but determination of the loss of the individual constituents is practically impossible.

For all relatively volatile preservatives such as creosotes, the regular method including sterilization can in a way be considered a permanency test. Fortunately this evaporative loss is in the same order of magnitude as that encountered in the field after an exposure of several years, and correlation with outdoor tests is unexpectedly good. Closer control of the amount evaporated would be desirable, but experience has shown this to be difficult of consistent attainment. Artificial weathering machines such as that described by Gillander, Rhodes, King and Roche [10] constitute a reasonably successful attempt to reproduce natural conditions. Leaching is more easily controlled and duplicated than evaporation, and preservatives which by their nature might be considered to be water soluble are subjected to a standard leaching cycle if the initial test has shown them to be promising. Again the close correlation of field results with the laboratory is gratifying.

Of necessity, details of technique have been only sketchily reviewed in this paper; information as to the exact procedure will soon be available in the chemical press together with complete results on several preservatives of interest. Table II contains the results of an assay on a supposedly permanent inorganic preservative before and after

TABLE II

Assay by Wood-Block Method of "Water Insoluble" Preservative Before and After Leaching

Concentration in Pounds per Cubic Foot	*Poria incrassata*			*Coniophora cerebella*			U–10			*Polyporus vaporarius*			Uninoculated
	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Control Weight Loss in Per Cent
2.24 Original.....	√ √	3.8 7.3	10 10	√ √	5.3 4.1	10 10	√ √	3.4 8.3	10 10	√ √	5.5 4.6	10 10	3.9 5.4
Leached.....	4–3 4–4	4.0 1.4	9 10	3–2 3–2	1.8 1.7	10 10	2–2 2–2	2.0 0.8	9 10	2–3 3–2	2.1 1.2	9 10	0.8 1.4
1.24 Original.....	√ √	2.1 2.4	10 10	√ √	1.3 1.9	10 10	√ √	2.2 2.0	10 10	√ √	0.9 3.1	10 10	2.2 1.9
Leached.....	3–3 3–3	1.2 7.5	9 8	3–3 3–2	9.7 1.9	7 9	3–3 3–3	15.9 6.1	5 7	4–4 3–4	4.7 7.0	8 8	0.4 0.3
0.60 Original.....	√ √	0.4 1.3	10 10	√ √	0.4 0.4	10 10	√ √	0.4 0.9	10 10	√ √	0.9 1.2	10 10	0.9 0.4
Leached.....	4–2 4–2	5.1 10.4	8 6	1–1 4–2	0.8 0.8	10 10	3–4 3–4	16.7 15.8	4 4	3–3 3–3	3.2 11.9	9 6	0.0 0.0
0.28 Original.....	√ √	0.0 0.4	10 10	√ √	0.0 1.0	10 10	√ √	0.0 0.0	10 10	√ √	0.5 0.0	10 10	0.3 0.1
Leached.....	4–3 4–4	13.6 30.9	6 2	4–3 4–4	17.6 23.5	4 2	4–4 4–4	50.2 12.4	0 6	4–4 4–4	19.8 15.7	4 4	0.0 0.0
0.14 Original.....	√ 1–1	0.0 0.0	10 10	√ 1–1	0.0 0.0	10 10	1–2 √	0.0 0.0	10 10	√ √	0.0 0.0	10 10	0.0 0.0
Leached.....	4–4 4–4	22.9 32.8	3 1	4–3 4–3	12.8 26.5	6 2	4–4 4–4	46.4 68.2	0 0	4–4 4–4	26.6 32.8	2 1	0.0 0.0

Ratings under each heading represent duplicate determinations
√ = No growth on specimen

TABLE III

ASSAY BY WOOD BLOCK METHOD OF PROPRIETARY PRESERVATIVE

Concentration in Pounds per Cubic Foot	*Lentinus lepideus*			*Fomes roseus*			U-10			*Lenzites trabea*			Uninoculated
	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Control Weight Loss in Per Cent
12.53	4-2 4-1	3.5 1.7	9 10	2-1 √	0.4 0.4	10 10	1-1 √	0.9 0.4	10 10	√ √	0.5 0.0	10 10	1.7 0.8
10.04	4-4 4-4	9.9 7.1	6 7	4-4 4-4	9.3 7.7	5 6	4-3 4-4	24.5 10.4	3 5	4-2 2-1	11.9 3.2	6 8	0.6 0.3
8.06	4-4 4-4	11.1 8.0	5 6	4-4 4-3	10.6 7.7	5 7	4-3 4-3	5.0 4.3	9 9	4-2 4-2	12.6 10.9	6 5	0.2 0.5
6.64	4-4 4-4	13.4 9.6	5 6	4-4 3-1	7.7 0.5	7 10	4-4 4-4	10.1 9.2	6 6	2-2 4-1	4.5 3.7	8 9	0.1 0.0
4.25	3-4 4-3	6.4 6.2	9 9	4-4 4-4	21.1 19.1	4 4	4-4 4-4	27.6 24.2	3 3	4-4 4-3	30.9 12.9	3 6	0.0 0.0
2.60	4-4 4-4	15.6 8.4	4 7	4-4 4-4	12.9 8.7	4 7	4-4 4-4	38.5 30.6	1 1	4-3 4-4	14.7 10.7	5 6	0.0 0.0

Ratings under each heading represent duplicate determinations
√ = no growth on specimen

TABLE IV

ASSAY BY WOOD BLOCK METHOD OF TYPICAL CREOSOTE

Concentration in Pounds per Cubic Foot	*Lentinus lepideus*			*Fomes roseus*			U–10			*Trametes serialis*			Uninoculated
	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Control Weight Loss in Per Cent
8.82	2–2 2–2	2.7 1.9	10 10	√ √	1.3 0.9	10 10	1–1 √	0.9 0.9	10 10	√ √	1.3 0.9	10 10	1.4 0.8
4.29	3–1 3–2	2.6 1.8	9 9	3–1 2–1	3.2 1.5	9 10	1–1 √	1.0 0.5	10 10	3–2 2–1	0.0 0.0	10 10	0.6 0.3
2.24	4–2 4–2	2.7 2.7	8 8	2–2 3–2	8.8 6.5	6 7	4–4 4–4	38.9 36.1	1 1	3–2 2–1	2.1 0.0	9 10	0.3 0.1
1.09	4–3 4–3	12.6 11.4	5 5	4–4 4–4	19.4 13.3	4 6	4–4 4–4	44.3 32.1	0 1	4–4 4–4	30.8 12.1	1 6	0.1 0.0

Ratings under each heading represent duplicate determinations
√ = no growth on specimen

TABLE V

ASSAY BY WOOD BLOCK METHOD OF COMPOUND INDICATED AS WORTHLESS ACCORDING TO THE PETRI DISH METHOD

Concentration in Pounds per Cubic Foot	*Lentinus lepideus*			*Fomes roseus*			U–10			*Lenzites trabea*			Uninoculated
	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Growth Rating	Weight Loss in Per Cent	Dissection Rating	Control Weight Loss in Per Cent
2.8	√ 1–1	0.0 0.4	10 10	2–1 2–1	0.8 0.0	10 10	3–2 2–2	0.0 0.0	10 10	3–1 3–1	+0.4 +0.4	10 10	+0.4 0.0
2.1	1–1 1–1	+0.4 +0.4	10 10	2–1 3–1	0.0 +0.4	10 10	3–2 3–2	+0.4 +0.4	10 10	3–1 3–1	+0.4 +0.4	10 10	+0.8 +0.4
0.9	1–1 2–1	1.4 1.4	10 10	3–2 4–2	1.4 0.8	10 10	3–1 3–2	0.0 0.9	10 10	3–1 3–1	0.0 0.0	10 10	0.0 0.5

Ratings under each heading represent duplicate determinations
√ = No growth on specimen

leaching. The losses in weight on the unleached specimens are also present in the controls and are presumably due to an extremely soluble non-toxic salt known to be present. Analysis of the leach waters indicates that the toxic substances were also slowly but definitely soluble. Field results on this same preservative were favorable for a year, but considerable decay was found the second year in all but the two highest concentrations. Table III presents the results obtained with a well-known proprietary preservative of the organic type. The concentrations given are for the preservative as purchased, which consists of a 25 per cent solution of solids in a volatile solvent. This solvent was allowed to evaporate completely before exposure of the test blocks to the fungus. For comparative purposes Table IV illustrates a test of a typical coal-tar creosote. Included as a matter of special interest, Table V outlines the wood-block assay on a material which the petri dish method indicated to be worthless.

This adaptation of the kolle flask method has been in constant use more or less in its present form for the past three years. Hundreds of complete assays have been made with results to date in good agreement with the slower and more expensive outdoor tests. By the use of a range of concentrations the relative efficacy of various preservatives can be judged, but definite expressions of the absolute value of any preservative have been avoided. With conditions controlled for maximum decay, this test is admittedly severe. This very severity, however, is probably an asset in the elimination at the outset of the poor and mediocre materials unworthy of further study.

Bibliography

1. Richards, C. A.—"Methods of Testing Relative Toxicity," *Proc. Amer. Wood Pres. Assoc.*, **19** (1923).
2. Schmitz, H.—"Laboratory Methods of Testing the Toxicity of Wood Preservatives," *Indus. & Engg. Chem., Analytical Ed.*, **1**, 76–79 (1929).
3. Liese, J. et al.—"Toximetrische Bestimmung von Holzkonservierungs Mitteln." *Z. f. Angew. Chem.*, **48**, 21 (1935).
4. Buller, A. H. R.—"The Destruction of Paving Blocks by the Fungus *Lentinus Lepideus*," *Jour. Econ. Biol.*, **1**, 101–138 (1905).
5. Snell, W. H.—"Studies of Certain Fungi of Economic Importance in the Decay of Building Timbers," *U. S. Dept. Agr. Bull.*, 1053, 1–47 (1922).
6. Humphrey, C. J.—"Timber Storage Conditions in the Eastern and Southern States with Reference to Decay Problems," *U. S. Dept. Agr. Bull.*, 510, 1–42 (1917).
7. Lumsden, G. Q.—"Proving Grounds for Telephone Poles," *Bell Laboratories Record*, **2**, 9–14 (1932).
8. Hubert, E. E.—"Outline of Forest Pathology," 416 (1931), John Wiley & Sons, Inc., New York.
9. Waterman, R. E. and Williams, R. R.—"Small Sapling Method of Evaluating Wood Preservatives," *Indus. & Engg. Chem., Analytical Ed.*, **6**, 413–19 (1934).
10. Gillander, H. E., King, C. G., Rhodes, E. O., and Roche, J. N.—"The Weathering of Creosote," *Indus. & Engg. Chem.*, **26**, 175–183 (1934).

Study of Magnetic Losses at Low Flux Densities in Permalloy Sheet *

By W. B. ELLWOOD and V. E. LEGG

Energy losses in ferromagnetic materials subject to alternating fields have long been considered as due solely to hysteresis and eddy currents. However, at the low flux densities encountered in certain communication apparatus, a further loss is observed which has been variously termed "residual," "magnetic viscosity," or "square law hysteresis." The search for an explanation of this loss has led to precise measurements of hysteresis loops with a vacuum ballistic galvanometer, and of a.-c. losses with inductance bridges. From these results, it appears that that part of the a.-c. effective resistance of a coil on a ferromagnetic core which is proportional to the coil current is strictly identified with the hysteresis loop area as measured by a ballistic galvanometer, or as indicated by harmonic generation in the coil. The hysteresis loop can now be constructed in detail as to size and skewness on the basis of a.-c. bridge measurements. This conclusion was reached previously on a compressed iron powder core, and is now confirmed on an annealed laminated 35 per cent nickel-iron core. Observed eddy current losses for this core exceed those calculated from classical theory by 20 per cent. This excess is ascribed to the presence of low permeability surface layers on the sheet magnetic material. The a.-c. residual loss per cycle (nominally independent of frequency, like hysteresis) is not observed by ballistic galvanometer measurements, although it indicates an energy loss some eight times the hysteresis loss for the smallest loop measured ($B_m = 1.3$ gauss). Analysis of the residual loss shows that it increases with frequency up to about 500 cycles, and remains constant at higher frequencies (to 10,000 cycles per second). Concurrently with the increase of residual loss, the permeability of the alloy is observed to decline with increasing frequency about 1 per cent below the value predicted from eddy current shielding. This effect is most noticeable at frequencies below 1000 cycles.

THE search for an explanation of the excessive magnetic losses observed at low flux densities by alternating current bridge measurement as compared with theoretical indications based on direct-current measurements has led to a more accurate review of both types of measurement.[1, 2, 3] The a.-c. energy loss per cycle which has re-

* To be published in May 1937 issue of *Jour. of Applied Physics.*

[1] W. B. Ellwood, *Physics*, **6**, 215 (1935).
[2] V. E. Legg, *Bell Syst. Tech. Jour.*, **15**, 39 (1936).
[3] W. B. Ellwood, *Rev. Sci. Inst.*, **5**, 301 (1934).

ceived most study is the "residual" or "viscosity" loss.[4] It appears related to hysteresis loss because it is nominally independent of frequency, but it differs in being proportional to B_m^2 instead of B_m^3 which would be required by Rayleigh's law for hysteresis loops. Any satisfactory investigation of this anomalous loss demands precise determination of its value, and of its variation with frequency. For this purpose, ballistic galvanometer measurements of the hysteresis loop have been made and compared with bridge measurements of a well annealed 35 permalloy laminated core.

In a previous paper, the magnetic properties of a ring of compressed powdered iron were studied at low flux densities using a sensitive vacuum galvanometer[3] and a multiple swing ballistic method.[1] Hysteresis loops were measured at flux densities B_m ranging from 1.8 to 115 gauss, which showed energy losses proportional to B_m^3 in accordance with Rayleigh's law. Alternating-current measurements agreed with the ballistic measurements as to the magnitude of the energy loss and the proportionality to B_m^3, but in addition showed a residual loss proportional to B_m^2 which was of the same order of magnitude as the Rayleigh hysteresis at these low flux densities.

The analysis of measurements made on the compressed dust core was complicated by the inhomogeneous structure, by the variety of particle shapes and thickness of insulation, and by the mechanical stresses incident to forming the core. To eliminate these objections, the present study was undertaken using a core consisting of well annealed sheet material, for which eddy current losses can be calculated by classical formulae.

Selection of Material

Considerable a.-c. data were at hand from which to select material for this experiment. The properties of a few representative materials are given in Table I. The constants are defined by the equation

$$\frac{R_f}{\mu_m f L_f} = aB_m + ef + c = \frac{8\pi W}{B_m^2}, \tag{1}$$

where R_f is the difference between the resistances measured with a.-c. and with d.-c. on a toroidal coil with inductance of L_f due to core material of permeability μ_m, when the maximum flux density is $\pm B_m$ and the frequency is f cycles per second. Here the hysteresis coefficient

[4] H. Jordan, *E. N. T.*, **1**, 7 (1924); H. Wittke, *Ann. d. Phys.* (5) **23**, 442 (1935); F. Preisach, *Zeit. f. Phys.*, **94**, 277 (1935); R. Goldschmidt, *Zeit. f. tech. Phys.*, **13**, 534 (1932).

is a, the eddy current coefficient e, and the residual loss term is c. W is the energy loss per cycle in ergs/cm^3 of core material. The permeability coefficient $\lambda = (\mu_m - \mu_0)/\mu_0 B_m$.

Table I shows that annealed 35 permalloy in sheet form has the most convenient values of B_r and c/a for further study of this effect. This alloy is of the face centered cubic lattice type common to a large class of magnetic alloys. The numerals preceding the various permalloys give the nickel or alloy percentages, as classified by G. W. Elmen.[5]

The measurement of magnetic losses of 35 permalloy involved further refinements in technique. The high initial permeability required the construction of a special air core mutual inductance to simplify the

TABLE I

Material	Initial Permeability	$\lambda \times 10^4$	$c \times 10^6$	$a \times 10^6$	c/a	$B_r^* \times 10^4$.
Compressed Powder Cores						
Grade B Iron	35	7.0	110	50.	2.2	13.
81 Permalloy	75	1.8	40	5.5	7.3	3.1
Laminated Cores						
35 Permalloy, Annealed	1660	30.	60	5.0	12.	62.
38 Permalloy, Hard	100	7.0	118	9.6	12.	7.2
38 Permalloy, 800° Annealed	1330	9.0	27	1.5	18.	15.
40 Permalloy, 1000° Annealed	2060	12.0	20	1.4	14.	22.
45 Permalloy, Annealed	2550	5.4	14	.43	33.	8.2
78.5 Permalloy, Annealed	3900	8.1	0	0.6	0	18.
2.4–78 Cr Permalloy, Annealed	14600	6.4	3	.07	43.	7.6
8–79 Cr Permalloy, Annealed	3025	31.	14	2.6	5.4	60.
45–25 Perminvar, Annealed	450	0.02	0.0	.002	—	0.0

* These values of remanence were computed from Rayleigh's law as $3a\mu_0 B_m^2/16$ for $B_m = 2$.

measuring circuit and increase its stability. The high rate of change of permeability with temperature made it necessary to enclose both the specimen and the air core mutual inductance in a constant temperature box (at 37.1 ± 0.01° C.) throughout the tests.

The Specimen

The material was melted in a high-frequency furnace, cast, and cold-rolled with intermediate annealings, to strip of thickness t = 0.0160 cm. and width 7.62 cm. Analysis showed the following composition: Ni, 35.00 per cent; Fe, 64.25 per cent; Mn, 0.40 per cent; S, 0.030 per cent; Si, 0.02 per cent; C, 0.01 per cent. The resistivity ρ was 82.2 micro-ohm-cms. at 37.1° C. The strip was wound into a tight

[5] *Electrical Engineering*, 54, 1292 (1935).

spiral core with successive turns insulated from each other by painting with a suspension of fine quartz powder in CCl_4 immediately prior to winding. The core had an effective magnetic diameter $d = 11.22$ cm., and cross-sectional area of alloy $A = 3.96$ cm.^{2}. It was annealed in pure hydrogen for one hour at a temperature of 1000° C.

In order to protect the annealed core from mechanical stress during subsequent winding, it was placed on felt in an annular bakelite box which held it without constraint. The box was wound with a 20-turn magnetizing coil using a flat tape composed of 28 parallel strands of insulated wires connected together at the ends. This winding practically covered the box with a single layer of wire, and gave a uniform magnetizing force. It was employed as the magnetizing coil in both the ballistic and the a.-c. bridge measurements. For the ballistic

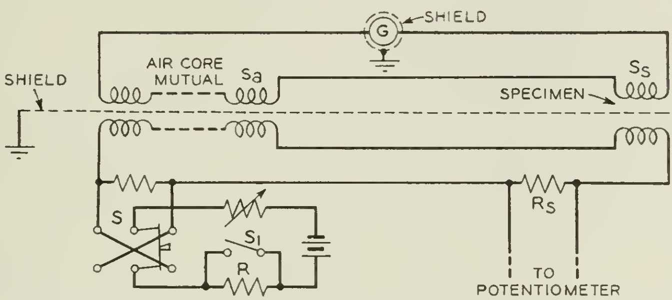


FIG. 1—Ballistic galvanometer circuit, showing adjustable air-core mutual inductance in series-opposition with the test coil.

tests, a layer of insulation was applied over the primary winding before applying the secondary winding. This insulation consisted of two wrappings of silk tape interspersed with a tinfoil sheath which formed a grounded electrostatic shield between the secondary and the primary winding. The foil was cut to avoid a short-circuited turn. The toroidal secondary winding consisted of 5000 turns No. 19 silk-covered enamelled copper wire.

D.-C. APPARATUS

In the former experiment,[1] the specimen was compared with a fixed air core mutual inductance in terms of the galvanometer deflection and the primary currents required to obtain approximate balance. For this experiment, the circuit was modified so that the same current flowed in the primaries of both the specimen and the air core mutual (Fig. 1). Thus variable thermal effects in the primary circuits were eliminated and the measuring technique simplified. This required that the mutual inductance be adjustable so as to bring the unbalance onto the galvanometer scale.

The mutual inductance (approximately 0.260 henrys) was constructed for convenience in four separate sections. Each section had a hard wood toroidal core, a low resistance toroidal secondary winding, an inter-winding shield, and sectionalized primary windings on the outside. The secondary windings were connected in series with the galvanometer by a twisted shielded pair of wires. The primary winding groups were also connected in series, and adjusted so that the combination resulted in a mutual inductance of the right value to obtain balance. To eliminate humidity as a source of error each coil was painted with cellulose acetate, covered with silk tape, painted again, baked 48 hours at 108° C. and finally potted in Superla wax in an earthenware jar with only the tops of the terminals exposed. All connections were made by soldering.

During the measurements, the maximum primary current corresponding to H_m was held constant to 0.01 per cent by comparing the voltage drop across R_s with a battery of Weston standard cells. Switching was automatically performed by a photocell and selector switch mechanism previously described [3, 1] but not shown here. These operated switches S and S_1 at the proper time and in the right order. The difference in flux turns between the air core mutual inductance and the specimen was determined in terms of the ultimate galvanometer deflection as before. From this the difference in B between the side of the hysteresis loop and a straight line drawn through its tips could be computed for a given H. A number of values of this difference ΔB were thus determined for different values of H, and plotted to give the hysteresis loop.

A.-C. Apparatus

In order to compare results obtained by the vacuum ballistic galvanometer with those obtained with alternating currents, bridge measurements of resistance and inductance were made over the same range of flux densities at a number of frequencies ranging from 35 to 10,000 cylces. The secondary winding was removed and the special 20-turn primary winding used for most of the measurements. Later an additional 60-turn winding was used for checking the measurements in the low-frequency range. In either case the inductance was low enough to depress any effect of distributed capacitance far below the precision of the measurements.

Measurements were made on a 10-ohm equal ratio arm inductance comparison bridge,[2] and were verified at low frequencies using a 1-ohm ratio arm bridge. Calibration of the bridge and standard coils was effected by making measurements over the entire frequency range on a calibrated high quality air core coil substituted for the test coil. The maximum correction required on this account was approximately 0.1 per cent of the resistance due to the magnetic core.

The source of alternating current was an oscillator-amplifier supplying approximately 0.4 watt undistorted power, calibrated for these measurements against the Laboratories' standard frequency. The current was adjusted by the insertion of resistance in series with the

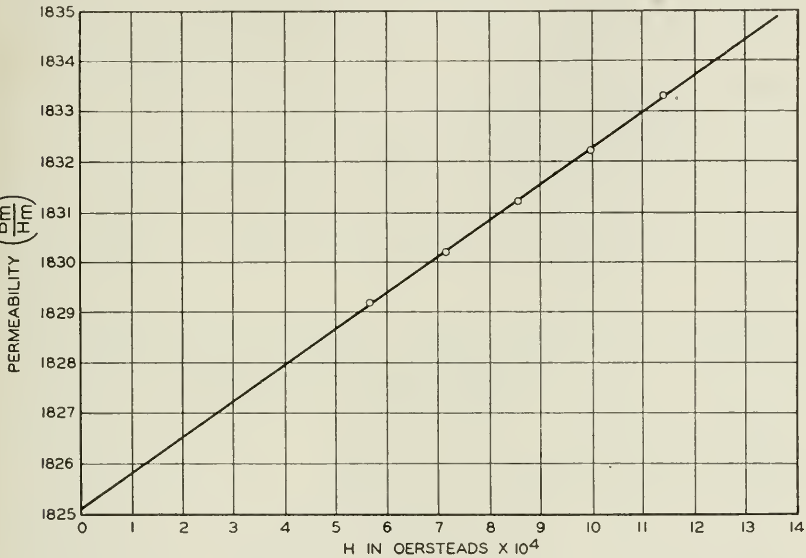


FIG. 2—Core permeability as measured by the ballistic galvanometer.

primary of the bridge input transformer, and was measured by means of a thermocouple between the transformer secondary and the bridge.

The bridge unbalance was amplified by means of an impedance coupled amplifier for the 10-ohm bridge, and by means of a resistance coupled amplifier for the 1-ohm bridge. The amplified unbalance was observed by means of head phones at frequencies above 200 cycles, and by means of a vibration galvanometer at lower frequencies. The d.-c. balance required bridge current of about 3 m.a. in the test coil winding, and had the same precision as the a.-c. balance, viz., $\pm$ 0.0002 ohm. The inductance readings were corrected for the air space within the winding, and had a relative accuracy of about 0.03 per cent, and an absolute accuracy of approximately 0.1 per cent.

D.-C. Results

The permeability $\mu = B_m/H_m$ of the specimen is shown as a function of H_m in Fig. 2, on a greatly enlarged scale in which the zero of permeability is not shown. The slope of this line gives $\lambda = 21.5 \times 10^{-4}$.

Values of ΔB are plotted against H for two different hysteresis loops in Fig. 3, from the area of which the energy loss W is computed. For

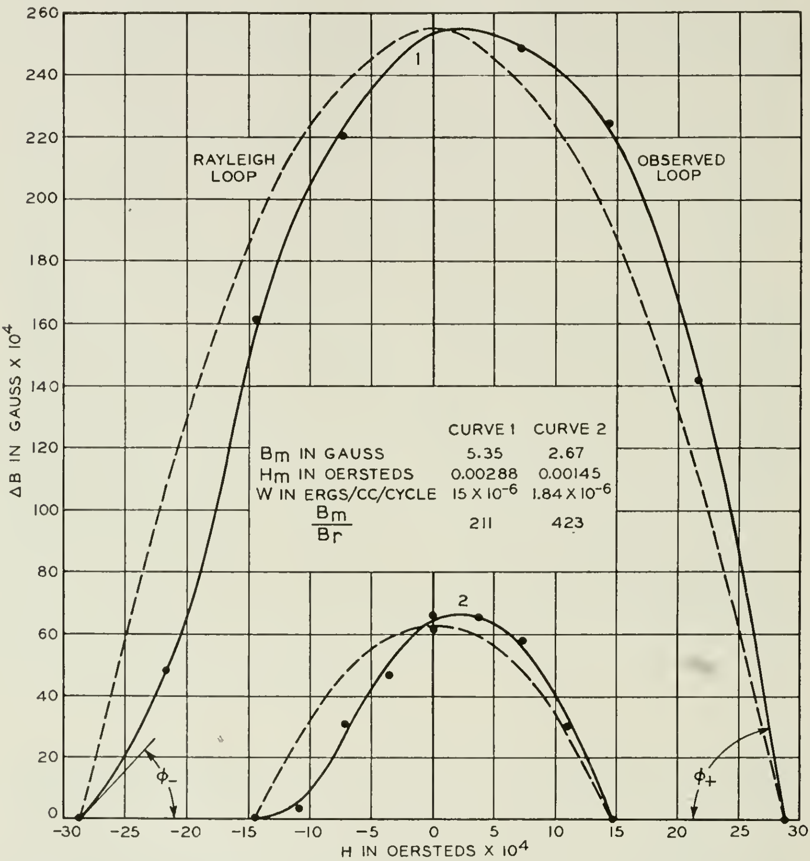


Fig. 3—Two hysteresis (half) loops plotted with reference to a straight line through their tips.

comparison, Rayleigh loops are shown as broken lines. By analysis of the interior angles ϕ_+ and ϕ_- between the ΔB vs. H curve and the H axis at points $+H$ and $-H$ respectively, it is found that ϕ_+ is larger than for the Rayleigh loop, and is given by the relation $\tan \phi_+ = \mu_0 \lambda B_m$, while ϕ_- is a corresponding amount smaller than for the Rayleigh loop. The hysteresis coefficient a and the permeability coefficient, λ are related by the equation $a = 8\lambda/3\mu$ for a loop having parabolic shape, in which case the interior angles are equal, and tan $\phi_\pm = \mu_0\lambda B_m$. Since the latter equation applies for ϕ_+ only, on the observed skewed loop, it follows that the Rayleigh relation between a and λ is more or less inaccurate. In fact, the ratio of $8\lambda/3\mu$ to a is a measure of the skewness of the loop. For the present material, this ratio is about 1.15. This result is in accord with our previous data, but was evidently not noted by Rayleigh because the free poles in his magnetic circuit tended to mask the asymmetry. The fact that these hysteresis loops are slightly skewed shows that those processes which produce the familiar S-shaped loop at high flux densities are already present at these low flux densities.

Despite a skewed shape, the area of the observed loop approximates closely the area of a parabola drawn through the remanence and the tips. Hence, supplementary values of energy loss W were obtained from remanence observations at several values of H_m, using the formula $W = 2B_rH_m/3\pi$. The slenderness factor of the loops may be measured by the ratio B_m/B_r, which varies from 211 to 890 for the different loops studied.

The a.-c. resistance introduced by the hysteresis loss of the core material yields the ratio $8\pi W/B_m^2$, as noted in Eq. (1). Values of this quantity computed from the areas of the loops of Fig. 3. and from remanence determinations are plotted in Fig. 4. They agree closely with the aB_m term of Eq. (1) obtained by a.-c. measurements, as shown by the solid line in Fig. 4. The sum of $c + aB_m$ is shown by the broken line. It is evident that the ballistic galvanometer gives no indication of residual loss.

It is interesting to note that the hysteresis loop at low flux densities can now be constructed in detail using the data obtained from a.-c. measurements. The remanence is $B_r = \frac{3}{16}a\mu_0B_m{}^2$ and tan $\phi_+ = \mu_0\lambda B_m$. The angle included between the upper and lower branches of the loop at the tips is $(\phi_+ + \phi_-)$ and is given by the equation tan $[\frac{1}{2}(\phi_+ + \phi_-)] = \frac{3}{8}\mu^2{}_0aB_m$.

A.-C. Results

Values of R_f and L_f were measured as a function of the current at fixed frequencies. The values of $R_f/\mu_m fL_f$ are plotted in Fig. 5 as a function of current with frequency as a parameter. In order to shorten the vertical scale, the appropriate ordinates are indicated in connection with each line. These form a family of straight lines parallel to one another. This shows that the hysteresis coefficient is practically a constant over the low flux density range for all frequencies. From the slope of these straight lines the hysteresis coefficient a of eq. (1) is calculated to be 2.6×10^{-6} which agrees with the value 2.53×10^{-6}

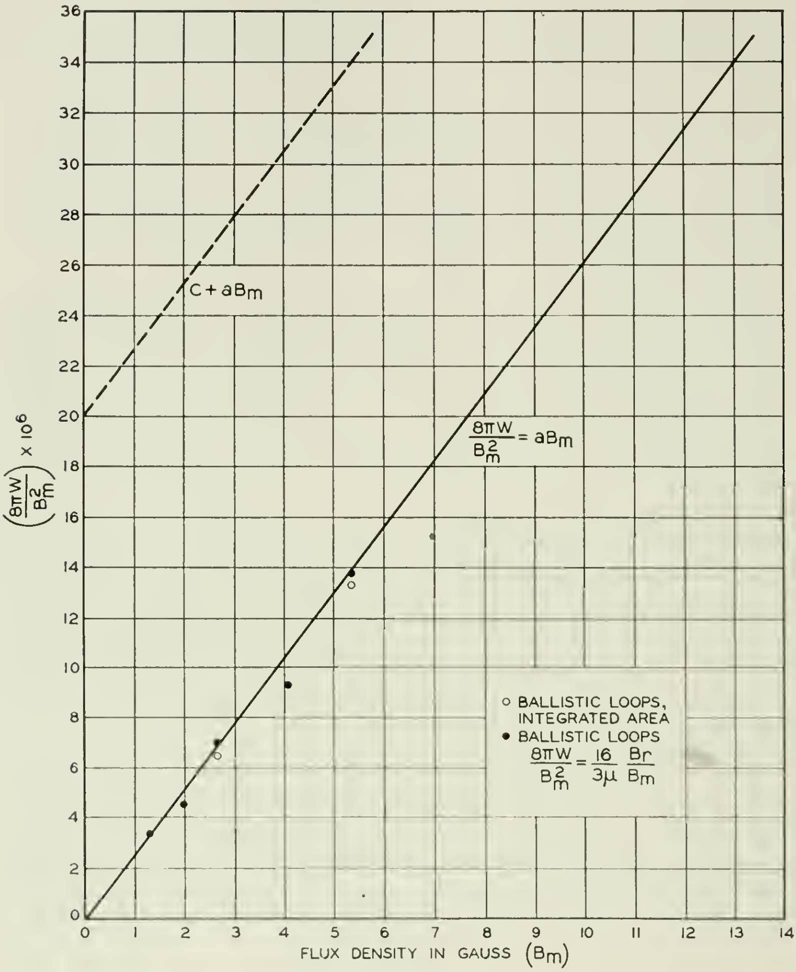


Fig. 4—Comparison of ballistic galvanometer and a.-c. determinations of hysteresis loss. Note absence of residual loss from ballistic observations.

computed from the ballistic tests. The divergence of the data from linearity at higher currents is shown by the dotted curves, which indicate divergence of the hysteresis loop from the Rayleigh form at higher flux densities.

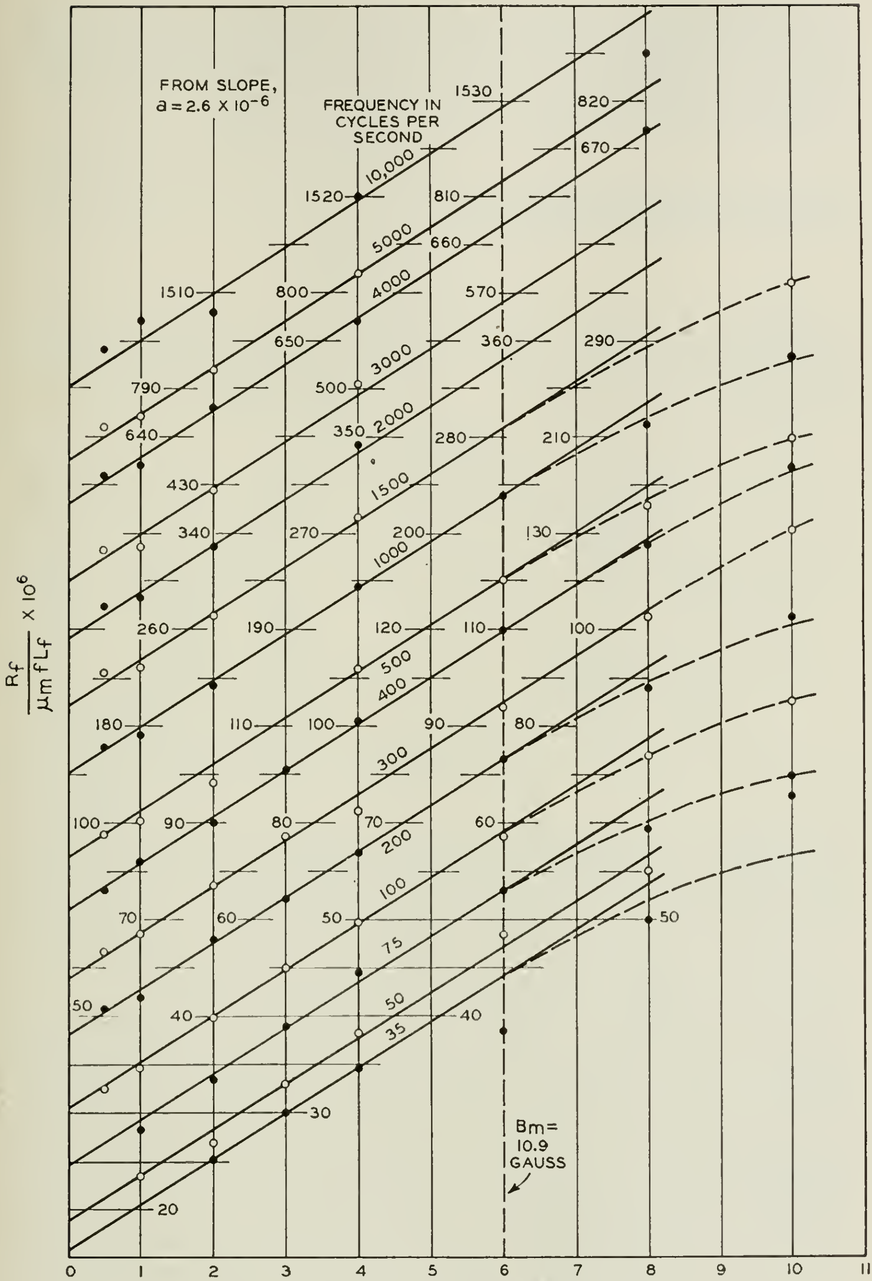


FIG. 5—A.-c. bridge values of $R_f/\mu_m f L_f$ vs. coil current of various frequencies. Slope of the straight parallel lines gives the hysteresis coefficient a.

Further information on hysteresis loss is obtained from measurements of harmonic voltages generated in the coil winding when there is current I_1 of frequency f. It has been shown [6] that the third harmonic voltage for materials with Rayleigh hysteresis loops is

$$E_3 = 0.6\, aB_m\mu_m L_m f I_1,$$

from which the hysteresis coefficient is

$$a = \frac{25\, E_3}{3\, f I_1^2}\sqrt{\frac{Ad \times 10^{-9}}{2\mu_m^3 L_m^3}}.$$

Measurements of third harmonic voltages have been made on the coil described in this paper by P. A. Reiling and the results are shown in Table II.

TABLE II

f	I_1 m.a.	E_3 m.v.	$a \times 10^6$	f	I_1 m.a.	E_3 m.v.	$a \times 10^6$
1000	2.0	.0168	2.2	100	3.0	.00447	2.6
	5.0	.133	2.8		5.0	.0106	2.2
	10.0	.55	2.9		10.0	.0473	2.5
					16.8	.1497	2.8
400	1.41	.00335	2.2				
	2.0	.00709	2.3	75	8.0	.0199	2.2
	3.0	.0158	2.3		10.0	.0335	2.3
	5.0	.0457	2.4		18.2	.112	2.4
	10.0	.195	2.5				
				50	10.0	.0359	3.7

The values of a thus obtained show no consistent variation with current or frequency. They give an average value of 2.5×10^{-6}, which is in close agreement with the ballistic and a.-c. bridge results. It therefore appears that that part of the effective resistance which is proportional to current is identifiable with hysteresis loss as obtained by ballistic means, and with that which generates harmonic voltages.

The intercepts for $I = 0$ in Fig. 5 are therefore due to eddy current and any residual losses. They have been plotted against frequency in Fig. 6. The line through these points is generally assumed to be straight, and the eddy current and residual loss coefficients are derived from its slope and intercept. It appears, however, that this line is not strictly straight, but has a somewhat steeper slope at lower frequencies, so that the ordinary graphical method of loss separation fails.

An analytical separation of losses can be made for any frequency interval by returning to the values of $R_f/\mu_0 f L_f$ as obtained from Fig. 5,

[6] E. Peterson, Bell Syst. Tech. Jour., 7, 762 (1928).

subtracting the value at f_1 from that at f_2 and dividing by the frequency interval $f_2 - f_1$ to give the eddy current coefficient e of equation (1).[7] Figure 6 gives e thus derived, as a function of f, showing a value approximately 20 per cent higher than calculated from the relation $e = 4\pi^3 t^2/3\rho$ at frequencies above 500 cycles, and progressively higher as the frequency approaches zero.

The fact that e is larger than predictable from classical theory has been ascribed to the presence of a low permeability surface skin on

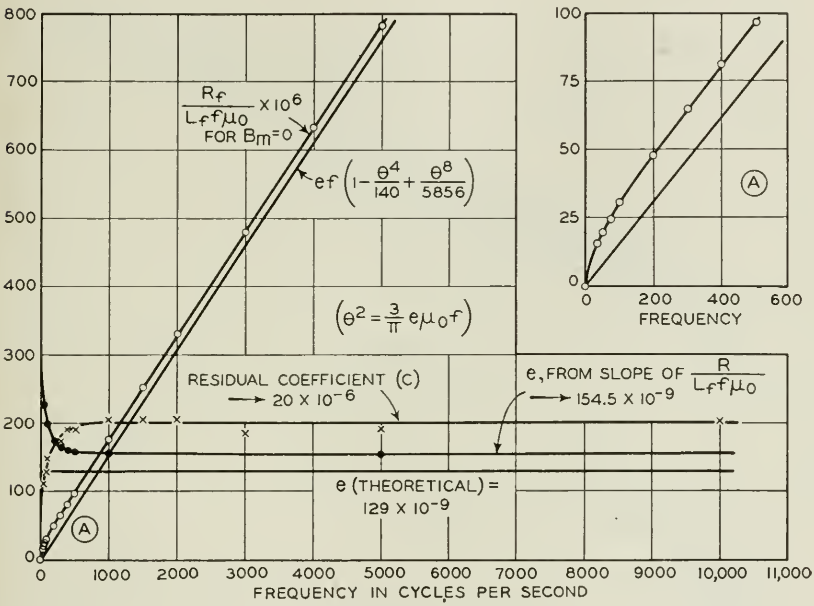


FIG. 6—Intercepts of Fig. 5 vs. frequency. Slope of the curve gives eddy current coefficient e. Residual loss coefficient c varies with frequency near $f = 0$.

practically all materials which have been reduced to sheet or wire form by mechanical deformation.[8] But since the eddy current coefficient depends only on the resistivity and effective thickness of the material, any apparent variation of e with frequency can only be interpreted as an indication that the residual loss is varying with frequency.

Taking the value of $e = 154.5 \times 10^{-9}$ characteristic of the higher frequencies, the eddy current loss per cycle has been calculated, and is indicated in Fig. 6. The amount by which the observed loss exceeds

[7] Correction terms must be included at higher frequencies to take account of eddy current shielding as noted in ref. 2.

[8] E. Peterson and L. R. Wrathall, *I. R. E. Proc.*, 24, 275 (1936).

the calculated eddy current loss gives the residual loss coefficient c. Thus, the value of c is found to be a constant 20×10^{-6} at frequencies above 500 cycles, but to decline toward zero as the frequency approaches zero, in evident accord with the ballistic galvanometer result.

The inductance due to the core shows a similar frequency effect. The observed inductances for the 20-turn winding are given in Fig. 7. The values at each frequency for the various currents fall on a straight

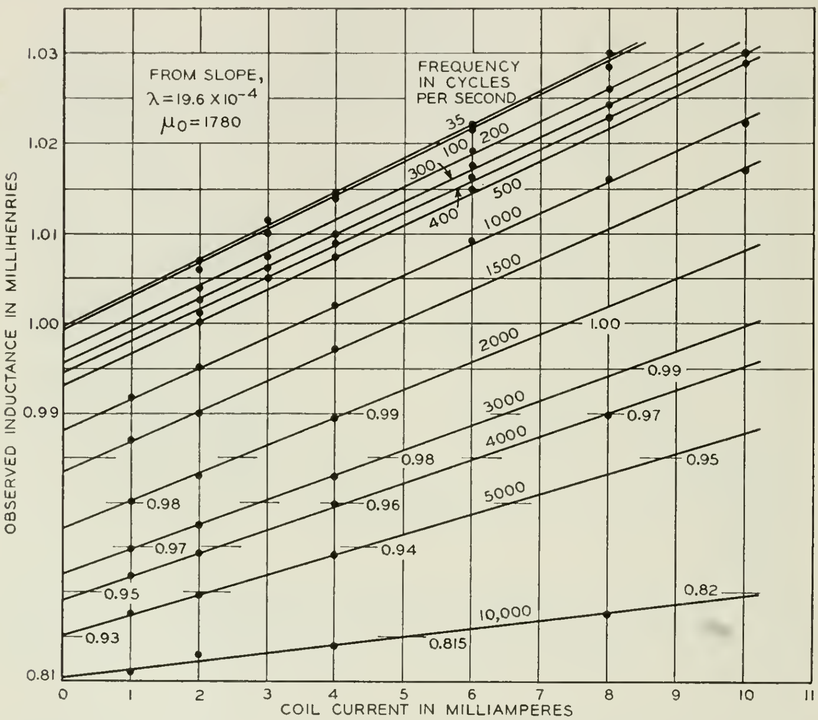


Fig. 7—Inductance observed on 20 turn coil, at various frequencies.

line, the slope of which gives the permeability coefficient, $\lambda = 19.6 \times 10^{-4}$, for the lower frequencies where eddy current shielding can be neglected.

The values of L for $I = 0$, obtained from Fig. 7, are plotted against frequency in Fig. 8. The most remarkable feature of this curve is the decline of inductance (or apparent permeability) of about 1 per cent at low frequencies, where very little decrease on account of eddy current shielding is to be expected. The characteristic shielding curve has been computed using the value of e obtained from the resistance measurements in the relation

$$\frac{L}{L_0} = \frac{1}{\theta}\frac{\sinh\theta + \sin\theta}{\cosh\theta + \cos\theta} = \left(1 - \frac{\theta^4}{30} + \frac{\theta^8}{732} - \cdots\right),^*$$

where $\theta = \sqrt{3e\mu_0 f/\pi}$. Using the values of L/L_0 thus computed, the effect of eddy current shielding was eliminated from the observed values, and the results plotted in the upper curve in Fig. 8, showing a rapid

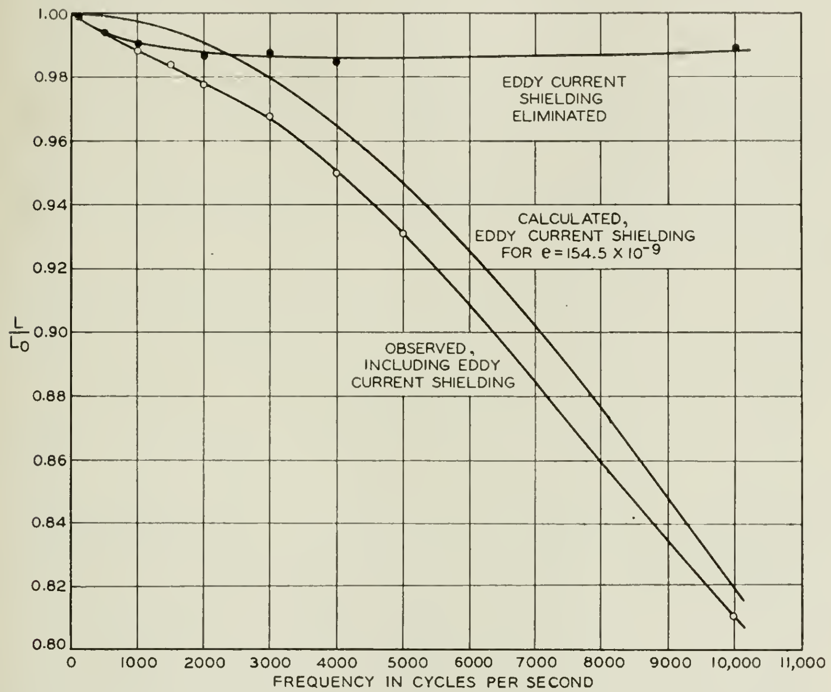


FIG. 8—L/L_0 from intercepts of Fig. 7. Observed values about 1 per cent lower than calculated for eddy current shielding.

decline of inductance or apparent permeability with increasing frequency at low frequencies, and the flattening off to a fairly constant value at higher frequencies.

The initial permeability thus obtained (average of results from the 20-turn and 60-turn windings) is 1780. This is somewhat lower than the value found by the ballistic galvanometer. The difference is due to magnetic aging which was observed to decrease the permeability of

* The series is inaccurate at frequencies above 5000 cycles.

the core at the rate of approximately 1 per cent per month. In contrast, no change of resistivity was detected in an accelerated aging test consisting of a bake at 100° C. for 150 hours.

Discussion

The ballistic data on both the present annealed 35 permalloy sheet core and the previous compressed powdered iron core show that the area of the hysteresis loop varies as B_m^3 in agreement with Rayleigh's law. The magnitude of the loss is not given by the fractional slope λ of the μ,B line as required by Rayleigh's law because the loops are not parabolic in shape. This discrepancy gives a measure of the skewness of the hysteresis loop. The B_m^3 portion of the a.-c. data agrees with the loop areas obtained with the ballistic galvanometer. The threefold agreement between the ballistic data, the harmonic measurements, and the a.-c. resistance measurements indicates that the hysteresis loop is substantially unchanged in shape over a frequency range extending from 0 to 10,000 cycles.

Since the hysteresis loop has a size and shape independent of frequency, and area strictly proportional to B_m^3, it accounts for that part of the effective resistance of a coil on a ferromagnetic core which is proportional to the alternating magnetizing current. The remainder consists of eddy current and residual losses.

The ordinary graphical method of separating these losses is excluded by the obviously non-linear relation between $R_f/\mu_0 f L_f$ and f. Using an analytical method, the eddy current loss is found to be some 20 per cent larger than computed by classical theory, indicating the presence of low permeability surface layers on the sheet material. The residual loss coefficient is found to increase with frequency up to about 500 cycles, and to remain constant at higher frequencies (up to 10,000 cycles).

The observed inductance diminishes with increasing frequency about 1 per cent below the value calculated for eddy current shielding, the most noticeable decrease occurring below 1000 cycles where eddy current shielding is practically absent.

Various theories have been advanced to account for residual loss, as noted in a previous paper.[1] Goldschmidt [9] and Dannatt[10] have attributed the loss to non-homogeneous alloy structure, or preferred axes in such directions as to give a flux component perpendicular to the sheet surface, with accompanying eddy-currents unconstrained by the sheet thickness. This theory fails to account for residual losses in com-

[9] R. Goldschmidt, *Helv. Phys. Act.*, **9**, 33 (1936).
[10] C. Dannatt, *I. E. E. J.*, **79**, 667 (1936).

pressed powder cores, where eddy-currents are confined to single particles and cannot be increased by a modified direction of magnetic flux.

The most notable feature of residual loss is its large value for unannealed materials, and its extremely small values for well annealed alloys, particularly 78.5 permalloy and 45–25 Perminvar. (See Table I.) The permeability is increased by annealing while both c and a are decreased. On the other hand B_r is slightly increased. The decreases in hysteresis and residual loss are attributed to the decrease in work done against internal strains, which also tend to limit initial permeability.[11]

This suggests that residual loss may be due to elastic hysteresis or even simple mechanical friction, with magneto-striction providing the necessary coupling between the elastic or frictional variables and the magnetizing field, as pointed out in our previous paper.[1] Thus, in addition to losses from eddy-currents and magnetic hysteresis, mechanical work is done by the alternately expanding and contracting core—work expended on itself and its supports and insulation. Since the ballistic galvanometer measures only equilibrium values of B and H, this work is not revealed in the area of the ballistic loop. However, in the a.-c. loop the magnetostriction strains produce stresses too rapidly to be relieved, so that B lags behind H with an absorption of energy into the surroundings. This results in an additional effective resistance beyond that due to magnetic hysteresis and eddy-currents. For a sufficiently slow process in well annealed material supported with minimum constraint, the stresses may relieve themselves by thermal agitation and do very little work. But for sufficiently rapid traversals of the loop, all the magnetostrictive stresses will do the same amount of work on the core and its surroundings every cycle. Unannealed materials, or materials rigidly constrained, should continue to show residual loss at very low frequencies. The magnitude of c and its variation with frequency thus should depend on the magnetostrictive constant for the material, and on the types of dissipative constraints.

[11] R. M. Bozorth, *Elec. Eng.*, **54**, 1251 (1935).

Moisture in Textiles

By ALBERT C. WALKER

Evidence is presented that for a cotton hair structure of the specific type described, calculations are in such close agreement with many experimental data as to suggest the following tentative conclusions:

1. The moisture content necessary to form a monomolecular layer on all internal surface of the cotton hair appears to be slightly more than 1 per cent of the hair weight.

2. Less than half the internal surface, that termed fibril surface, appears to be involved in moisture adsorption which causes appreciable transverse swelling of the cotton hair. Upon this surface multimolecular chains of water seem to condense, the length of such chains increasing progressively up to saturation with corresponding increases in hair diameter throughout the whole of this range, each hydroxyl group in the cellulose surface being the base of a water chain, with separations between these chains along the surface corresponding to the arrangement of the hydroxyl groups on the cellulose surface.

3. Moisture adsorbed on surfaces within the cellulose aggregates composing the fibrils does not appear to be involved in transverse swelling, but may be responsible for the slight longitudinal swelling exhibited by cotton. The capacity of the cotton hair for this type of adsorption suggests that its locus is the ends of crystallites and therefore within the body of the fibrils. To account for the slight swelling, it is assumed that only a monomolecular layer can be adsorbed on these surfaces.

4. A theory is proposed to explain the dependence of the electrical properties of textiles upon their moisture adsorbing properties, and upon the surface distribution of moisture within the submicroscopic structure.

1. Introduction

A STUDY of the electrical properties of textiles and their dependence on atmospheric conditions and naturally-occurring impurities in the material has resulted in important economies and improvements in the use of textile insulation in the telephone industry. Recently, calculations have been made as to the moisture content and swelling of cotton at various equilibrium conditions, based on assumptions, first as to the structure of the cotton hair,* then as to the

* In keeping with recognized terminology, the individual cotton fiber is called a hair, suggestive of its morphological origin.

location and distribution of the internal surface upon which moisture might condense, and finally as to the manner in which moisture may be held upon this internal surface.

From this rather specific picture of the cotton hair structure it has been possible to calculate moisture contents and swelling properties

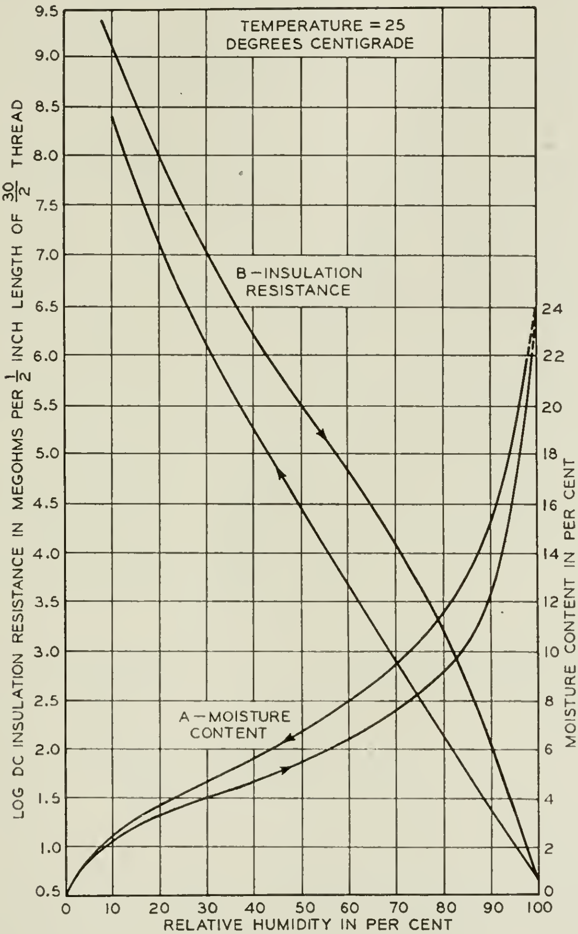


Fig. 1—Moisture sorption and electrical properties of raw cotton.

consistent with experimental data. It has been possible, also, to present a more comprehensive explanation of the change in the relations between electrical resistance and moisture content of cotton over the whole range of atmospheric humidity than that given in previous publications from these Laboratories.* It is therefore considered that such a picture should contribute towards a better understanding of the moisture-sorbing † properties, not only of cotton, but also of other similar fibrous materials, despite the hypothetical nature of some of the assumptions upon which the calculations are based.

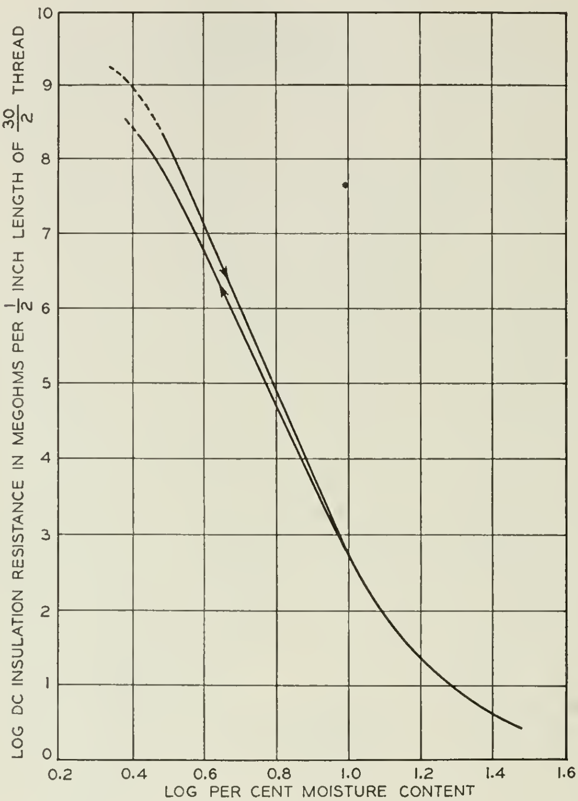


Fig. 2—Effect of previous history on resistance of raw cotton.

The electrical insulation resistance of textiles, when dry, is enormous compared with the resistance observed under atmospheric conditions. A change of but 1 per cent in atmospheric relative humidity (R. H.), equivalent to between 0.1 per cent and 0.2 per cent change in moisture content (M. C.), causes a change of about 25 per cent in resistance of

* See references 1 and 2 of the list with which this article concludes.

† "Adsorption" is here defined as the taking up of a gas or vapor by a solid, "desorption" the giving up of a gas or vapor, and "sorption" the general process without special indication of gain or loss. The use of these terms implies no assumptions with regard to the mechanism of the processes they denote.

cotton. With silk the corresponding change in resistance is somewhat greater, and although silk sorbs materially more moisture than cotton at any equivalent atmospheric condition, it is much the better insulating material. Small amounts of naturally-occurring, water-soluble salts in cotton, such as NaCl and K_2SO_4, seriously impair the resistance of this textile. Traces of acids or alkalies left after degumming have a similar effect on silk. By washing these materials in water, their electrical properties are greatly improved.

Figure 1-*A* shows the familiar equilibrium relation between relative humidity and moisture content for cotton, including the hysteresis loop. A similar hysteresis characteristic, Fig. 1-*B*, in the relative humidity-resistance relation has been discussed in previous publications.[1, 2] Figure 2 shows more clearly, as suggested by a comparison of the two types of curves in Fig. 1, that the resistance of cotton is critically dependent upon its moisture content. The curves in Fig. 2 show another important fact. The resistance of cotton may have, not one, but a *range* of resistance values for a single moisture content, depending upon the previous treatment or "history" of the sample.

This fact, which is one of great practical importance, is illustrated in Table I. Eleven samples of cotton, taken successively from the same spool, were dried to constant weight at 100° F., in a current of dry air, then equilibrated together under very carefully controlled conditions, first at 87.7 per cent R. H., then re-dried as before and re-equilibrated at 84.3 per cent R. H. several days later. The moisture contents of these samples were as follows:

TABLE I.

Sample No.	% Moisture Contents at 87.5% R. H.	% Moisture Contents at 84.3% R. H.
1	10.95	10.1
2	10.8	10.0
3	10.7	9.9
4	10.8	9.8
5	11.1	10.2
6	11.0	9.9
7	11.8	10.8
8	10.7	10.1
9	10.85	9.8
10	11.0	10.0
11	10.7	9.9

The moisture contents of these samples showed small but definite differences, persisting even between tests several days apart. One of the samples had apparently been treated slightly differently from the others in preparation, since it preserved a marked difference in moisture content in both tests. Since a change of only 0.1 per cent in M. C.

may cause a change of about 10 per cent in resistance, these data are considered significant in such testing methods as are used for electrical textiles.

In a previous publication [2] a series of simple equations was formulated to show the quantitative relations between moisture content,

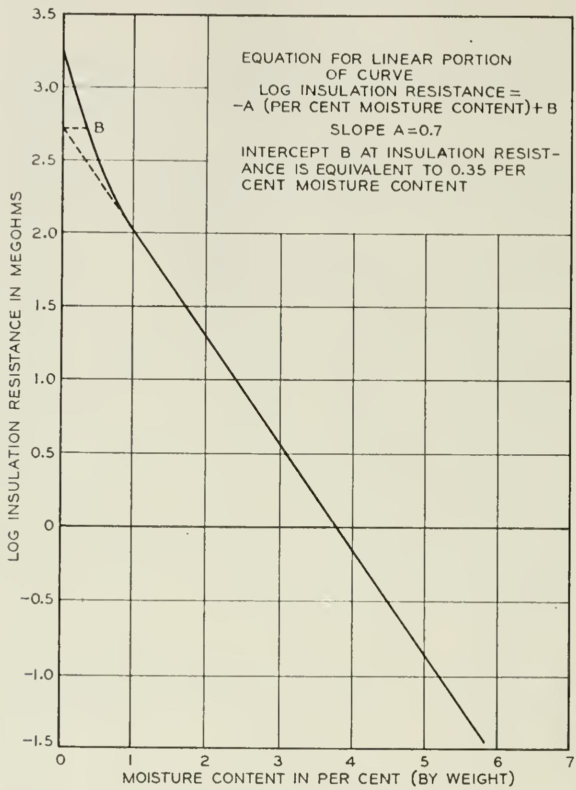


Fig. 3—Moisture content-resistance relation for rag paper.

relative humidity, and the resistance (Ω) of cotton:

$$\log \Omega = -A(\%\ \text{M. C.}) + B, \tag{1}$$

$$\log \Omega = -a(\log \%\ \text{M. C.}) + b, \tag{2}$$

$$\log \Omega = -\alpha(\%\ \text{R. H.}) + \beta. \tag{3}$$

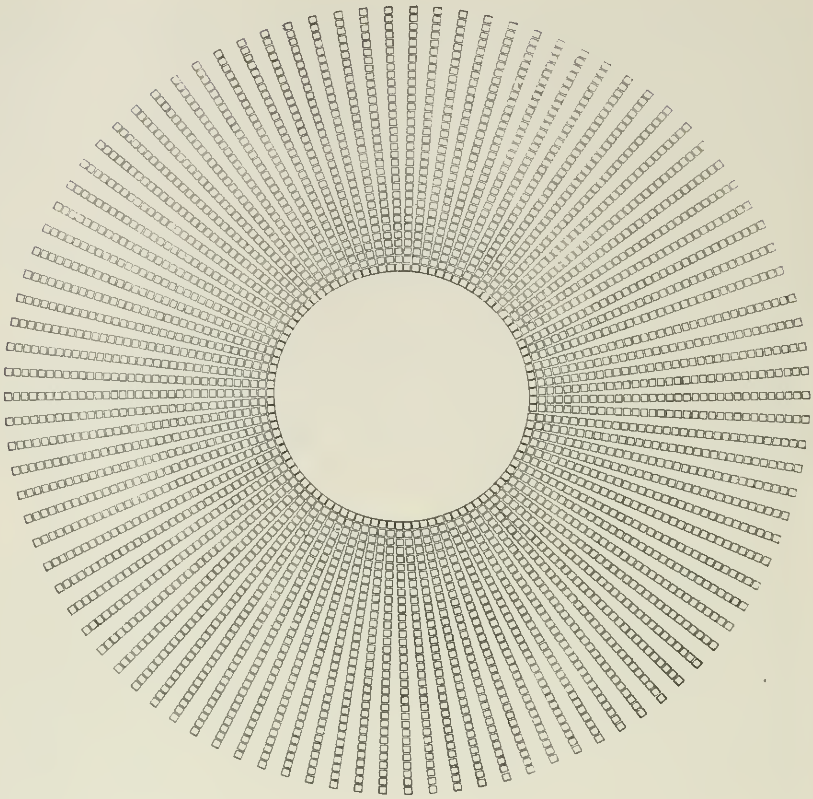
Equation 1 was considered as applying for moisture contents in the vicinity of 3 per cent, Equation 2 between 3 per cent and 10 per cent M. C., and Equation 3 from 10 per cent M. C. to saturation.

The ranges of application of Equations 2 and 3 are evident from Figs. 2 and 1-*B*, respectively; but since the resistivity of cotton becomes enormously high as the moisture content approaches zero, it has been difficult to verify the application of equation 1 below about 2 per cent. However, a recent study along somewhat different lines has provided us with information on this portion of the moisture content curve down to as low as 0.04 per cent M. C. It will be seen from Fig. 3 that equation 1 holds between 1 per cent and 6 per cent M. C. Below 1 per cent M. C., however, the resistance increases more rapidly with decreasing moisture content than is consistent with equation 1.

The data from which Fig. 3 was secured completes a chain of evidence upon which is based a theory of moisture adsorption which appears adequate to explain many properties of cotton. This theory involves, in addition to the data just discussed, a rather specific picture of the cotton hair structure.

2. Structure of the Cotton Fiber

According to Balls,[3] the cotton hair is formed by the outward extension of a single cell from the epidermis of the seed-coat; this extension, unaccompanied by any cell division may continue until the hair is 2000 times as long as it is broad. Up to about half maturity, the cell wall remains very thin but the hair attains most of its length; during the remaining half of the growth period (about one month) the wall thickens from the outside in until it appears to consist of about 30 to 35 concentric "growth rings" (see Fig. 4). Each growth ring consists, further, of parallel strands of fibrils, which run continuously in spiral form from end to end of the hair making one complete turn around the hair in about three diameters, and with periodic reversals in the direction of this spiral. Balls also suggests that side by side in each growth ring, there are about 100 fibrils, each separated from its neighbor by an air space. These fibrils are described by Balls as "dominoes" laid down one on top of the other in a pile-up of growth rings extending from the wall of the central canal or lumen to the outer wall of the fiber. Thus the front and back of each domino are growth ring boundaries, and each domino is separated from its neighbor by an air space. There are also air spaces between each domino in a growth ring. These air spaces are identified as the so-called pits in the wall structure. They are visible under a microscope, and appear to extend from the outer surface of the cotton hair down to the lumen. These air spaces are far larger in magnitude than those separating the front and back of each domino. Only by swelling the fiber in



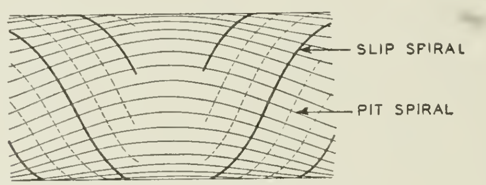


Fig. 4—The fiber structure of cellulose.

(a) An idealized cross-section of the cotton hair. The domino-like blocks shown in cross-section are arranged according to Ball's conception.

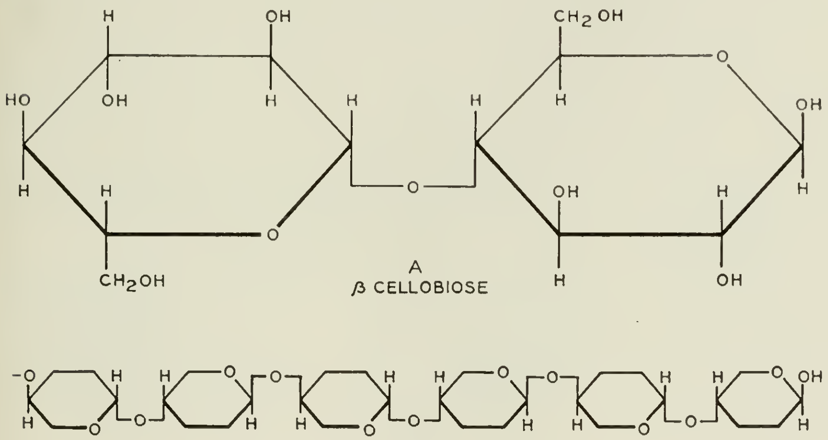
(b) The spiral arrangement of these domino-like blocks or fibrils is shown, together with the separating pits in the wall, and the slip-spiral effect along the fibrils.

Obviously these conventional type figures do not represent the true shape of the cotton of commerce, but they approximate the shape during the growth period, before the boll is opened. As the hair dries out, the central lumen collapses and the hair twists.

caustic soda can a differentiation be detected in the cross-sectional structure of the fiber to indicate this growth-ring character. Thus the fibrils are considered as being separated by air spaces on all sides, the whole cotton hair is spongy, and the surfaces of cellulose bounding these air spaces are internal surfaces of the fiber.

Slip spirals are visible in the hair surface at high magnification. Though decidedly irregular, they appear to cross the pits at approximately right angles, suggesting that there are additional internal surfaces at these points.

Cellulose from all sources appears to consist of definitely arranged crystallites or micellae.* Haworth [9] suggested that cellulose is composed of an elementary group consisting of two $C_6H_{10}O_5$ units, called cellobiose (Fig. 5-*A*). Figure 5-*B* indicates how these cellobiose units are joined together end to end to provide the fibrous structure of native cellulose.



A
β CELLOBIOSE

B
ONE END OF CELLULOSE CHAIN

Fig. 5—Molecular structure of cellulose.

Diffraction and chemical evidence indicate that the cellobiose units are arranged parallel to the *b*-axis of the unit cell, with one cellobiose group at each common edge of adjacent cells and one through the center of each cell. These form long primary valence chains arranged parallel to the fiber axis and are held together laterally by cohesive forces. Figure 6 shows this conception of the unit cell as given by Meyer and Mark.[5]

* Sponsler and Dore,[4] Meyer and Mark,[5] Freudenberg,[6] Herzog,[7] Polanyi.[8]

3. Moisture Adsorption on Fibril Surfaces

It appears reasonable to assume that moisture will first adsorb on dry cotton on the outer surface of the hair, and by diffusion in the vapor state will penetrate into the pits and adsorb on the pit walls. Since cotton swells appreciably in a transverse direction, but hardly at all lengthwise, it is further assumed that moisture in the pits will next penetrate between the fibrils which are contiguous to one another in the radial direction. At equilibrium with any humidity below that required to form a monomolecular layer, it is assumed that the water molecules will be distributed at random on active points over all of the internal surface. For humidities above this value, polymolecular chains of uniform thickness are assumed to adsorb at active points on the fibril surfaces only, since moisture on the growth ring surfaces of these fibrils appears to be responsible for the transverse swelling of the cotton.

The equilibrium moisture content of cotton is reduced if the hydroxyl groups on the cellulose molecules are acetylated or otherwise esterified. Consequently it seems reasonable to assume that each water molecule adsorbed on the cellulose surface is held by a force originating in the oxygen atom of a surface hydroxyl group. As may be seen from Fig. 5-*A*, there are six hydroxyl groups per cellobiose unit, and the percentage moisture equivalent to a monomolecular layer covering the surface of the fibril structure with each water molecule satisfying forces of a surface hydroxyl oxygen will now be estimated.

The fibril cross-section is estimated to be 1240 × 1300 AU, based on average dimensions of the cotton hair.* Assuming the cellobiose units arranged with the a-axis parallel to the fibril width (Fig. 6), there will be $1240/8.3 = 150$ unit cells across the fibril, and $1300/7.9 = 165$ unit cells down through the fibril. Therefore the total number of oxygen atoms per unit cell length in the four fibril surfaces is:

$$6 \times 150 \times 2 + 6 \times 165 \times 2 = 3780.$$

From this the moisture content equivalent to a monomolecular layer is:

$$\frac{3780}{2 \times 150 \times 165} \times \frac{18}{324} \times 100 = 0.42\%.\dagger$$

* The considerations upon which these and subsequent calculations are based are given in detail in separate publications which will appear in the April and May issues of *Textile Research*.

† The Angström unit AU is 10^{-8} cm. From Fig. 6 it appears that only the equivalent of one cellobiose unit may be available for surface adsorption per unit cell in the fibril surface. Furthermore, since molecules in solid or liquid surfaces are subject to unbalanced forces (surface tension) it is assumed that all surface cellobiose units are so oriented that all 6 hydroxyl groups have surface forces capable of adsorbing water molecules.

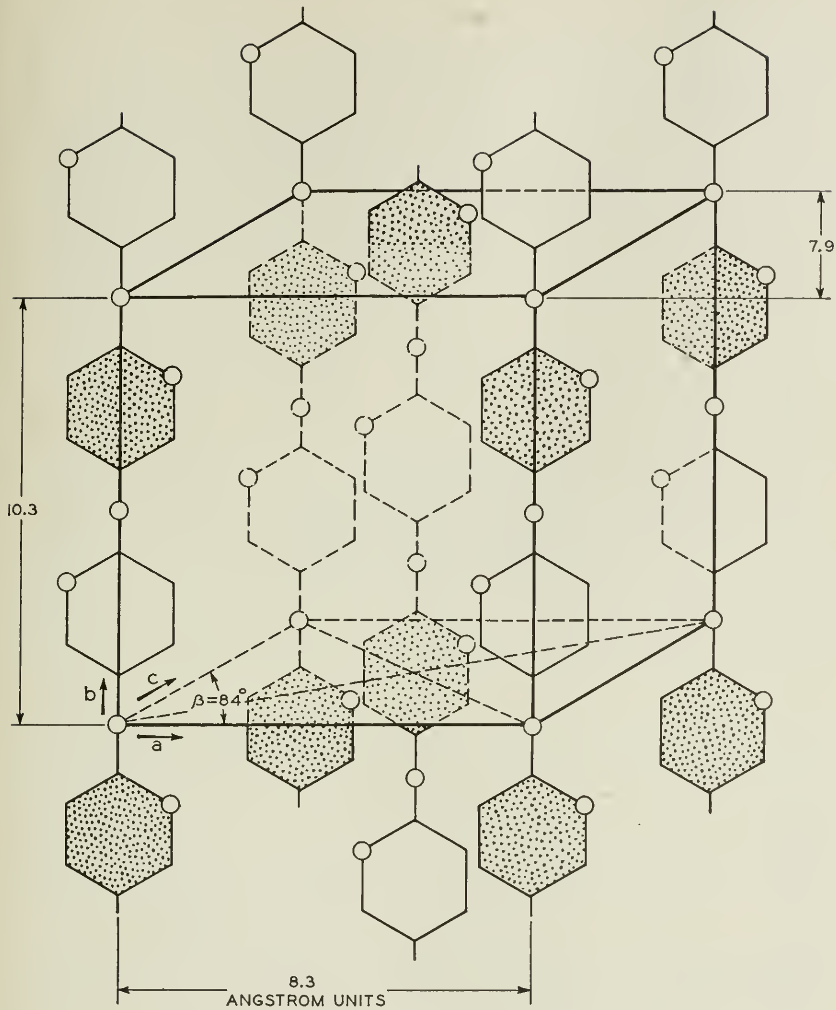


Fig. 6—The cellulose unit cell or crystallite structure.

4. Hair Swelling due to Fibril Surface Moisture

The diameter of the water molecule is reported as 3.8 AU. The surfaces of adjacent fibrils along growth ring boundaries must be separated at least $2 \times 3.8 = 7.6$ AU when a monomolecular layer of water is present on each contiguous surface. This corresponds to a total increase in hair diameter of $33 \times 2 \times 7.6 = 500$ AU. The percentage diameter increase in the hair is:

$$(500/125000) \times 100 = 0.4\%.$$

where 125000 AU is taken as the mean diameter of the dry cotton hair.

According to Collins,[10] the coefficient of hair diameter increase with humidity is about 0.11 per cent per 1 per cent R. H. Therefore, an increase of 0.4 per cent in hair diameter is found at $0.4/.11 = 3.6$ per cent R. H. From adsorption data (Fig. 7) by Urquhart and Williams,[11] this relative humidity corresponds to between 1.1 per cent and 1.2 per cent moisture content. The difference in this value and that

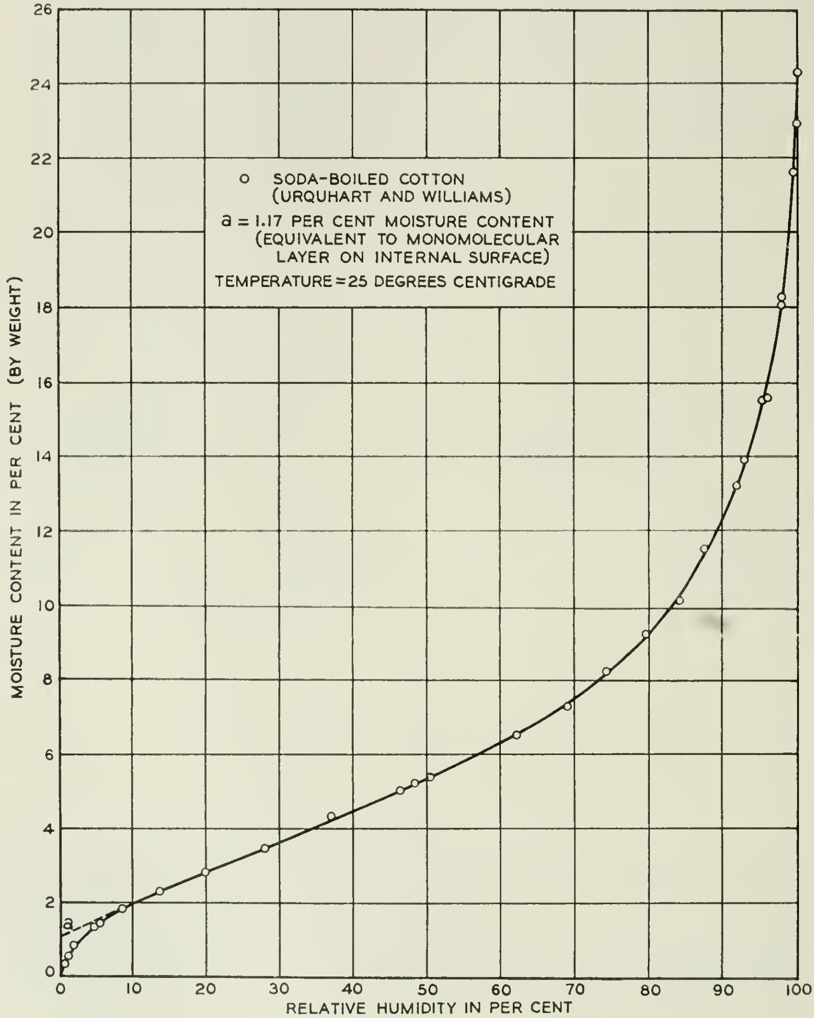


Fig. 7—Moisture adsorption-humidity relation for cotton at 25° C.

calculated under Section 3 is between 0.7 per cent and 0.8 per cent, suggesting that there is additional internal surface within the hair structure upon which moisture may be held without manifesting itself by an increase in diameter. This suggestion appears to be confirmed by quite different considerations.

Brunauer and Emmett[12, 13] consider it likely that the linear portion of van der Waals adsorption isotherms for nitrogen on the surface of ammonia catalysts indicates the building up of additional layers of adsorbed molecules. They state that extrapolation of this linear portion to zero pressure indicates the amount of gas needed to form a monomolecular layer upon this surface. Between 3 per cent and 50 per cent relative humidity the adsorption isotherm for cotton in Fig. 7 is very nearly linear. Applying this method to cotton the intercept (a) has values between 1.4 per cent and 0.35 per cent, depending on temperature. The average value is about 1 per cent being of the same order as that estimated from swelling data.

Since the estimated moisture content equivalent to a monomolecular layer on the internal surface of the cotton hair is so nearly the same when determined by two independent methods, it seems reasonable to postulate an additional internal surface in the cotton hair, amounting in extent to somewhat more than that corresponding to the fibril surfaces.

Slip spirals along the hair, crossing the pits at approximately right angles (see (b) of Fig. 4) suggest that there are discontinuities in the length of the fibril structure, hence further internal surface. Since the additional internal surface suggested by the preceding calculations and estimates does not appear to be involved in the transverse swelling of the cotton hair, it is suggested that this may be held upon the ends of crystallites or micellae which compose the fibril structure.

It is considered of much less importance to pursue the detailed calculations of this possibility than it is to point out that some such distribution of surface within the body of the fibril structure may be involved in adsorption of a small amount of moisture, and this picture is of material value in accounting for some of the properties of cotton.

5. Multimolecular Layers

It is further assumed that above 1.5 per cent moisture content, addition of moisture simply increases the thickness of the moisture layer upon the surfaces of the fibrils. The thickness (n) of the moisture layer on each fibril, expressed in number of water molecules, and the percentage moisture content at 50 per cent R. H. may readily be obtained. At 50 per cent R. H. the hair diameter increase is 5 per cent (plot of Collins' data). Using the following equation formulated in accordance with the calculations in section 4,

$$n = \left(\frac{500}{100} \times 125000\right) 7.6 \times 33 \times 2 = 12 \text{ molecular layers.} \quad (4)$$

Since a layer one molecule thick on the fibril surfaces is equivalent to 0.42 per cent moisture content, 12 such layers are equivalent to 5 per cent M. C. The observed moisture content at 50 per cent R. H. is 5.5 per cent. If to this 5 per cent M. C. thus calculated, is added the 0.7 per cent–0.8 per cent held within the fibril structure, a total of 5.7 per cent to 5.8 per cent is obtained, which checks remarkably well with the observed value (5.5%), considering the method of computation.

Collins reported swelling values for cotton exposed to 100 per cent R. H. (some condensation was visible on the cotton), and for cotton immersed in liquid water. Moisture contents calculated from these data as in equation 4 give 21 per cent and 23 per cent respectively, a surprisingly good agreement with observed values at saturation, which, as reported by various observers, lie between 20–25 per cent, depending upon the degree of wetness of the material, as indicated by condensation of moisture on the surface.

6. Moisture Required to Fill Pits and Lumen.

From Fig. 4-*a*, it appears that cotton may hold considerably more moisture than corresponds to the saturation values calculated under Section *C*. The total moisture calculated to fill the pits and lumen in addition to covering the fibril surfaces is more than 140 per cent. Coward and Spencer[14] have shown that wet cotton retains about 50 per cent of its weight of water after centrifuging, and these authors expressed the opinion that the water was not interstitial, but contained in the hairs themselves. The above calculations indicate that not only 50 per cent but much more than 100 per cent may be retained in the cotton hair, and it is possible that the amount retained after centrifuging or pressing may be some function of the treatment and the surface energy relations of the sorbed moisture.

7. Reduction in Moisture Sorption of Cotton by Acetylation

New [15] has shown that the equilibrium moisture content of cotton is progressively reduced as the acetyl (CH_3CO) content is increased. Below 21 per cent CH_3CO content, acetylation can be carried out in a mild way without appreciable effect upon the strength or physical structure of the hair. Above this value the fiber appears to be increasingly attacked by the acetylating mixture. Moisture adsorption isotherms for cotton containing different percentages of CH_3CO have been obtained by two investigators, New, and Storks.* Figure 8 shows portions of these adsorption isotherms at two humidities (20 per cent and 40 per cent), sufficient to give on extrapolation, monomolecular layer intercepts similar to that of Fig. 7. With increasing acetyl content, the intercepts indicate progressively lower moisture contents, until at 21 per cent or greater CH_3CO content, the extrapolated values are zero.

Considering the structure postulated in Fig. 4, it might be expected that a larger proportion of hydroxyls on fibril surfaces would be acetylated than the average throughout the body of the material, at any given acetyl content. The fact that the intercept is zero for acetyl contents in excess of 21 per cent suggests that all surface hydroxyls have been converted to acetate and no longer adsorb moisture. A consequence of this hypothesis is that the slope of the curve should approach zero with increasing acetyl content and reach zero above 21 per cent. Obviously this is not the case. If such partially acetylated cotton has all cellulose molecules in the fibril surfaces converted to the triacetate, and the remaining cellulose structure to the monoacetate, it is conceivable that the cohesive forces originally binding the fibril surface cellulose molecules to those directly beneath them have been so diminished in strength that water molecules can now adsorb upon the *available* hydroxyls in this second layer. Furthermore, it is assumed that conversion to monoacetate of all cellulose molecules within the fibril structure involves acetylation of all crystallite end hydroxyls, so that adsorption of moisture on cotton having more than 20 per cent acetyl content is upon *available* hydroxyls in the second layer of the fibril surface only.

Based on these considerations, calculations have been made of the moisture content of acetylated cotton containing 21.87 per cent acetyl content at 20 per cent R. H. and 40 per cent R. H. The details of these calculations and discussion are given elsewhere as previously indicated, so that only the essential results are stated below:

At 20 per cent R. H. the estimated moisture content held on available second layer hydroxyls on fibril surfaces of the cotton hair is 1.0 per cent. The actual value, shown on Fig. 8-*a*, is 0.9 per cent. At 40 per cent R. H., the estimated value is 1.85 per cent; the observed value is 1.80 per cent. These agreements are considered as excellent, particularly since the untreated cotton adsorbs 2.83 per cent and 4.50 per cent moisture content, respectively, at these two humidities.

* Unpublished data obtained by K. H. Storks, at the Bell Telephone Laboratories.

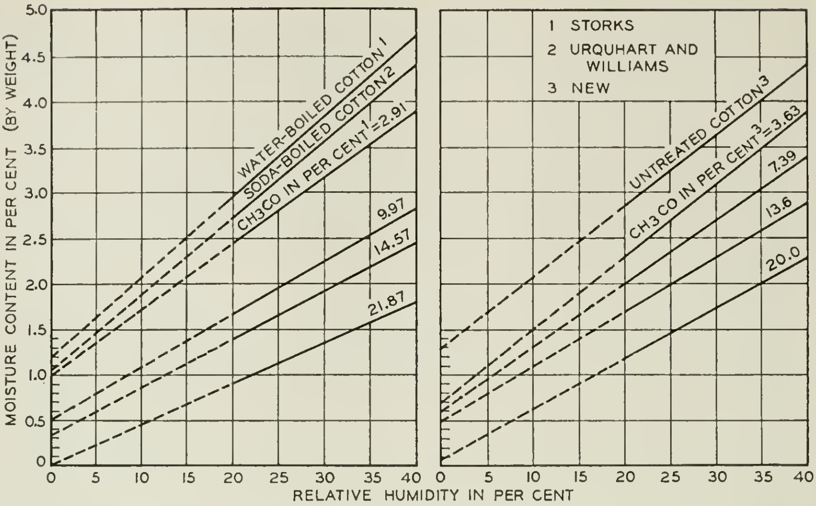


Fig. 8—Moisture adsorption intercepts for acetylated cottons indicating amounts of internal surface.

8. Theory of Conduction of Electricity Through Cotton

It now appears possible to provide a more comprehensive theory for the conduction of electrical current through the moisture paths in a textile than has been suggested previously.[1, 2]

Below 1 per cent moisture content it is likely that equation 1 fails to hold, due to obvious discontinuities in the moisture paths over surfaces containing less than a monomolecular layer of water.* Above 1 per cent, it is not to be concluded, however, that a continuous moisture film exists. It appears that some space is still available between certain water molecules, and the space pattern of water distribution is determined by the type of solid surface and the arrangement of active points or zones upon which each water molecule is held. In the case of cotton, such active points are considered to be hydroxyl groups; for silk they may be amino or carboxyl groups or both, and from Astbury's discussion [16] of the chain structure of these two materials it seems likely that their space patterns for moisture adsorption are different. Furthermore, it is assumed that each of these active points may be anchorage not only for a single water molecule, but for a chain of such molecules, the length of the chain being deter-

* It is of interest to note that if the linear portion of the curve shown on Fig. 3 is extrapolated to the insulation resistance axis, the insulation resistance corresponding to this intercept is found at 0.3 per cent M. C. (see *a*—Fig. 3), this being of the same order of magnitude as that estimated to cover fibril surfaces with a monomolecular layer, suggesting that the linear portion of the curve is specifically related to moisture adsorption on the fibril surfaces.

mined by the relative humidity. This conception is much like the picture of acid or oil molecules standing as a film on a water surface, with the polar end in the water.

Some such function as equation 1 may apply since simply increasing the length of the water chains will not cause a proportionate decrease in the resistance. It seems evident from Fig. 5-*a* that small increments of water might be expected to sufficiently lengthen water chains of minimum separation so as to establish contacts between them along the current path, but that larger increments of water are necessary to accomplish the same result between more widely separated anchorage points. This would explain the gradual change from the relation of equation 1 at low moisture contents to equation 2 at intermediate moisture contents. At some point along the humidity curve it is conceivable that capillary condensation occurs in the pits so that at progressively higher humidities the increasing cross-section of these pits plays a more important part in current conduction than the adsorbed chains of water molecules. This may explain why equation 3 is found to apply at highest humidities.

Adsorption, however, appears to continue throughout the whole of the humidity range, consistent with the hair swelling relations. Thus it is shown that adsorption and capillary condensation need not be considered separately but may go along together with a gradual shifting in importance from one to the other in the current conduction process.

Evershed [17] explained the decrease in resistance of a textile with increasing applied potential as being due to elongation of pools of water in the material under electrical stress forming more continuous current paths. This deviation from Ohm's law may be explained also as being due to the influence of increasing electrical stress upon the oscillation of the free ends of moisture chains, bringing more of them into orderly alignment along the line of applied potential, and establishing shorter current paths through the structure.

The difference in electrical behavior of different textile fibers may be illustrated by a comparison of the adsorption of moisture on cotton and silk surfaces. From Astbury's pictures of the structure of protein molecules as compared with cellulose molecules it appears that although there may be more points per unit surface for moisture to condense upon on protein surface, there are also possibilities of separation of adjacent moisture chains in a manner similar to that discussed for cellulose, and furthermore, there appear to be side chains of hydrocarbons interspersed in the protein chain which may act as barriers to the ready contact of adsorbed water chains on either side of these hydrocarbon chains. This may explain why silk has a higher resistance than cotton for a given moisture content. It might be expected that silk, due to these hydrocarbon barriers in the current path might also have a higher dielectric breakdown under potential stress.

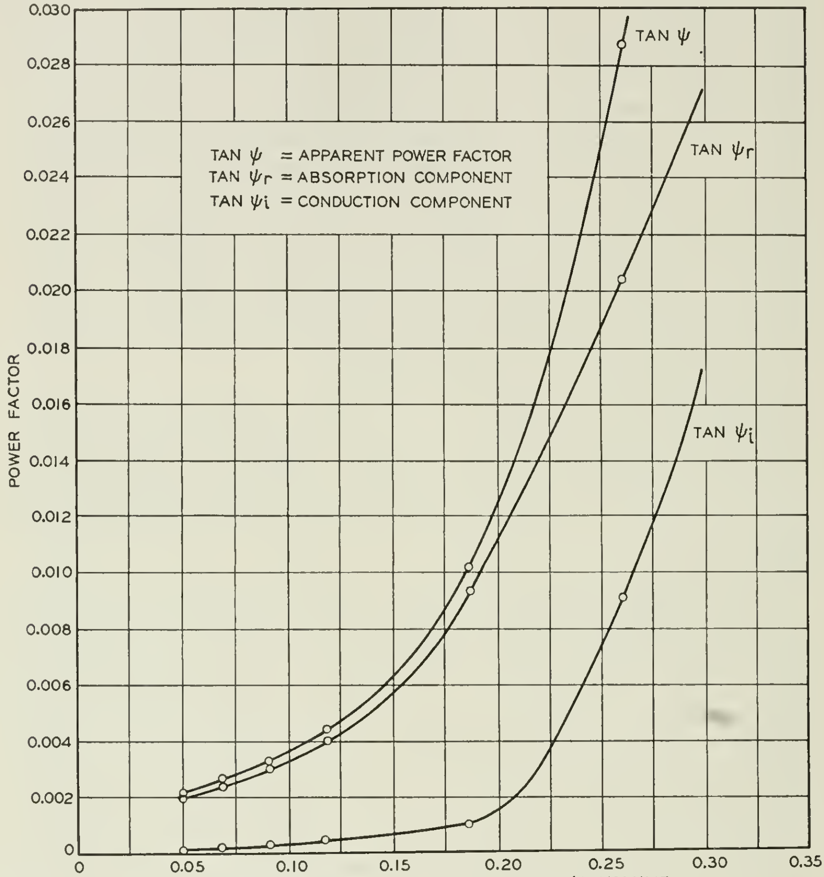


Fig. 9—Relation between power factor components of cable paper and moisture content.

Whitehead,[18] in a study of the effects of moisture on power cable paper insulation, obtained data on the change of power factor with increasing moisture content. Figure 9 shows this relation. It will be noted that the conduction component (tan ψ_i) of the power factor rises sharply as the moisture content of the paper exceeds 0.2 per cent by volume (about 0.3 per cent by weight). This increase occurs in the range consistent with the completion of the monomolecular layer of moisture on the internal surface of the material, particularly since the internal surface of paper and cotton appears to be of the same order of magnitude.

9. General Discussion of Theory

Peirce [19] proposed a two-phase theory for the adsorption of moisture by cotton, based upon the facts that a small quantity of water is adsorbed by dry cotton very much more rapidly than the same amount added to cotton with a moderate water content, that it has much greater effect on the elastic properties, evolves more heat, and is more difficult to remove. He regards the moisture attached to the hydroxyl points as in phase (*a*) while that in phase (*b*) consists of an indefinite number of molecules adsorbed in a looser fashion over all available surfaces, limited only by the conditions of space and of equilibrium with the external concentration of aqueous vapor.

The differentiation between fibril and other internal surfaces is considered as giving a more definite explanation of such two-phase adsorption than that proposed by Peirce. The (*a*) phase may be pictured as moisture added to the active hydroxyls which lie on the fibril surfaces where such surfaces are readily available for moisture adsorption, while the (*b*) phase is associated with a definite number of active hydroxyls but within the body of the fibrils and therefore less accessible than those on the surface. As has been pointed out, it is quantitatively reasonable to suppose these available internal hydroxyls to be located at the ends of the crystallites.

According to the cotton hair structure presented in this paper, there is actually more than twice as much internal surface available for moisture adsorption from the dry state as there is in cotton already in equilibrium with a 1 per cent moisture content. The slow diffusion of moisture to those surfaces most deeply buried in the fibrils may account for the fact that while cotton very nearly reaches equilibrium with any given humidity in a relatively few minutes, the final establishment of equilibrium conditions, involving a change of less than 1 per cent in moisture content, may take more than 24 hours. The first water molecule to attach itself to an active hydroxyl might be expected to evolve more heat than subsequent molecules in the chain since the interaction is between water and cellulose hydroxyls, not between water and water. Also, water held within the fibril structure no doubt is the most difficult to remove since it must diffuse out through this fibril structure.

This picture of moisture adsorption seems to provide a reasonable explanation for a variety of practical problems.

Thus it is known that if fibrous materials, such as wood, paper or fiber board are dried below a certain critical moisture content, permanent changes may be expected in the structure; even serious damage may result from too thorough drying at low temperatures. It seems evident that when the moisture content of such more or less dense structures is reduced to a point where the outer layers have less than 1 per cent of moisture, the internal surfaces of these layers may begin to lose the monomolecular layer of water. The valence forces of surface hydroxyl groups are now no longer satisfied by water molecules, so that such hydroxyls on contiguous surfaces may stick together. On readsorption of moisture, portions of these surfaces may be so permanently attached to one another that swelling will no longer occur in just the same way as originally, and cracking and warping may result.

Bibliography

1. Murphy and Walker, *Jour. Phys. Chem. 32*, 1761 (1928).
2. Walker, *Jour. Text. Inst. 24*, T145 (1933); also *Bell Sys. Tech. Jour. 12*, 431 (1933).
3. Balls, "Studies of Quality in Cotton," Macmillan & Co., p. 24 (1928); also *Egyptian Cotton Supplement*, *14*, 10 (1927).
4. Sponsler and Dore, Colloid Symposium Monograph, *4*, 174 (1926).
5. Meyer and Mark, *Ber. Deut. Chem. Gesell. 61*, 593–614 (1928); also "Der Aufbau Der Hochpolymeren Organischen Naturstoffe auf Grund Molekular-Morphologischer Betrachtungen," Leipzig (1930).
6. Freudenberg, *Ber. 54*, 764 (1921).
7. Herzog, *Ztschr. Phys. Chem.* (A) *139*, 235 (1928).
8. Polanyi, *Ztschr. Phys. 7*, 149 (1921).
9. Haworth, *Nature*, p. 430, Sept. 19, 1925; also *Jour. Soc. Dyers and Colourists, Jubilee Issue*, 1884–1934, 16–23.
10. Collins, *Jour. Text. Inst. 21*, T313 (1930).
11. Urquhart and Williams, *Jour. Text. Inst. 15*, T559 (1924).
12. Brunauer and Emmett, *Jour. Amer. Chem. Soc. 57*, 1754 (1935).
13. Emmett and Brunauer, *ibid. 57*, 2732 (1935).
14. Coward and Spencer, *Jour. Text. Inst. 14*, T30 (1923).
15. New, *Electrical Communication*, *13*, 216–225, 370 (1935).
16. Astbury, "Fundamentals of Fiber Structure," pp. 50 and 96, Oxford Press (1933).
17. Evershed, *Jour. Inst. Elec. Engrs.* (London), *52*, 51 (1914).
18. Whitehead, "Impregnated Paper Insulation," J. Wiley & Sons, p. 31 (1935).
19. Peirce, *Jour. Text Inst. 20*, T133 (1929).

Abstracts of Technical Articles from Bell System Sources

Determination of Ferromagnetic Anisotropy in Single Crystals and in Polycrystalline Sheets.[1] R. M. Bozorth. Following the work of Akulov and of Heisenberg on the magnetic anisotropy of cubic crystals, it is shown that by taking account of an additional term in the expression for the energy of magnetization the [110] direction may under certain conditions be the direction for easiest magnetization in a crystal, instead of [100] or [111] as given by previous theory. This is in accord with experiment. Magnetization curves for single crystals are calculated using the additional term and some peculiarities are recorded. The anisotropy constant appropriate for a single crystal (of iron) has been calculated from measurements on hard-rolled sheet in which there is preferred orientation of the crystals.

Impact Bend Testing of Wire.[2] W. J. Farmer and D. A. S. Hale. This paper comprises a discussion of a machine designed to make rapid determination of the ability of wire to resist permanent deformation by bending.

Two types of machine used in the industry for wire bend testing are described and their features discussed with regard to their suitability for use as standard test methods.

A bend tester operated by the impact of a pendulum has been developed by the Bell Telephone Laboratories in collaboration with Subcommittee IV on Mechanical Tests of the Society's Committee B-4 on Electrical-Heating, Electrical-Resistance and Electric-Furnace Alloys. Results of typical tests with this machine are given, together with information gathered from ultra-rapid motion pictures taken of the machine in operation.

It is concluded that the impact bending machine described offers a simple, rapid and accurate means of measuring the bending properties of wire and that the information acquired from the test is directly applicable to design problems.

Positions of Stimulation in the Cochlea by Pure Tones.[3] John C. Steinberg. The relation between tone frequency and position of

[1] *Phys. Rev.*, December 1, 1936.
[2] *Proc. 39th Ann. Mtg. Amer. Soc. for Testing Materials*, Vol. 36, 1936—Part II, Technical Papers.
[3] *Jour. Acous. Soc. Amer.*, January, 1937.

stimulation on the basilar membrane has been calculated from data on differential pitch sensitivity. The calculations involve assumptions concerning the choice of the upper and lower pitch limits of hearing and the choice of tone levels which should be used in obtaining differential pitch sensitivity data. It is shown that for quite different assumptions the positions of stimulation for tones in the range from 500 to 10,000 cycles are not greatly affected. Outside this range the positions depend on the assumptions. The calculated positions for tones of 1000, 2000 and 4000 cycles fall, respectively, at points on the membrane about $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{2}{3}$ of its length away from the helicotrema. The calculated positions are compared with positions obtained from post-mortem studies of human cochlea and with positions obtained from electric response measurements on the cochlea of anesthetized guinea pigs. The differences between various methods for the most part are no larger than calculated differences between observers.

Some Uses of the Torque Magnetometer.[4] H. J. Williams. The history of torque measurement as an index of ferromagnetic anisotropy is outlined. A simple magnetometer for torque measurement is described in detail and uses for the instrument are discussed. These include the measurement of anisotropy constants, coercive force, complete magnetization curves for single directions, and rotational hysteresis losses. With auxiliary ballistic measurement residual inductions and demagnetizing factors are obtainable.

[4] *Rev. Scientific Instruments*, February, 1937.

Contributors to this Issue

EDWIN H. COLPITTS, who has recently retired as Executive Vice President of the Bell Telephone Laboratories, scarcely needs an introduction. In 1899 he left Harvard to begin his career of research and development in the Bell System. In 1907, when development work was transferred from Boston to the Engineering Department of the Western Electric Company in New York, he also transferred and headed the Physical Laboratory. Later, with the formation of a Research Department, he became its head. In 1933, preliminary to the consolidation of the Department of Development and Research of the American Telephone and Telegraph Company with the Laboratories, Dr. Colpitts was appointed Executive Vice President.

W. B. ELLWOOD, A.B., University of Missouri, 1924; M.A., Columbia University, 1926; Ph.D., Columbia University, 1933. Bell Telephone Laboratories, 1930–. Dr. Ellwood has been engaged in various investigations relating to magnetic materials and measurements.

C. N. HICKMAN, A.B., Winona College, 1914; M.A., Clark University, 1917; Ph.D., Clark University, 1922. Physicist, Bureau of Standards, 1919–22; Physicist, U. S. Navy Yard, 1922–24; Research Physicist, American Piano Company, 1924–30; Bell Telephone Laboratories, 1930–. Since 1930 Dr. Hickman has been engaged in the development of special acoustical instruments.

C. M. HILL, B.S. in Chemistry, Princeton University. American Telephone and Telegraph Company, Long Lines Department, 1929–30; Bell Telephone Laboratories, 1930–. Mr. Hill's work has been on the biometrical problems of wood preservation.

VICTOR E. LEGG, B.A., 1920, M.S., 1922, University of Michigan. Research Department, Detroit Edison Company, 1920–21; Bell Telephone Laboratories, 1922–. Mr. Legg has been engaged in the development of magnetic materials and in their applications, particularly for the continuous loading of cables, and for compressed dust cores.

JOHN LEUTRITZ, B.S. in Chemistry, Bowdoin College, 1929; A.M. in Botany, Columbia University, 1934. U. S. Navy, Medical Corps, 1921–25. Bell Telephone Laboratories, 1929–. Mr. Leutritz' interest has been along biological lines, primarily in respect to wood preservation.

E. L. Norton, S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1922; M.A., Columbia University, 1925. Western Electric Company, Engineering Department, 1922–25; Bell Telephone Laboratories, 1925–. Mr. Norton has been engaged in the study of network and transmission problems. His present work is in connection with signaling circuits and apparatus.

Todos M. Odarenko, University of Technique in Prague, E.E., 1928. New York Telephone Company, 1928–30; Bell Telephone Laboratories, 1930–. Mr. Odarenko has been engaged in the measurement and study of transmission characteristics of existing and newly developed types of transmission lines.

S. A. Schelkunoff, B.A., M.A. in Mathematics, The State College of Washington, 1923; Ph.D. in Mathematics, Columbia University, 1928. Engineering Department, Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925–26. Department of Mathematics, State College of Washington, 1926–29. Bell Telephone Laboratories, 1929–. Dr. Schelkunoff has been engaged in mathematical research, especially in the field of electromagnetic theory.

Albert C. Walker, B.S., Massachusetts Institute of Technology, 1918; Ph.D., Yale University, 1923. Bell Telephone Laboratories, 1923–. Dr. Walker has been engaged in developing and applying methods of improving the electrical properties of textile insulation and methods for the inspection control of commercially purified textiles for telephone apparatus.

R. E. Waterman, B.S. in Chemical Engineering, Williams College and Massachusetts Institute of Technology. Western Electric Company, 1922–25; Bell Telephone Laboratories, 1925–. Mr. Waterman has been engaged in organic and biochemical investigations and for the past few years has been in charge of a group studying the chemical phases of wood preservation.

VOLUME XVI JULY, 1937 NUMBER 3

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XVI *July, 1937* *No. 3*

Scientific Research Applied to the Telephone Transmitter and Receiver *

By EDWIN H. COLPITTS

LET us recall a scene at the Centennial Exhibition in Philadelphia in 1876. Across a room had been strung wires connecting crude instruments, at one end of the room a transmitter and at the other end of the room a receiver. Dom Pedro, Emperor of Brazil, takes up the receiver and listens while Alexander Graham Bell speaks into the transmitter. The Emperor, astonished at hearing Mr. Bell's voice in the receiver, exclaims in amazement, "My God, it talks."

When at the same place, Sir William Thomson (later Lord Kelvin) took up the receiver and listened to Mr. Bell, the words of this distinguished scientist were, "It does speak," and continuing, "it is the most wonderful thing I have seen in America."

Sixty years have passed and, as a result of continued effort, the use of the telephone has become such an everyday matter that even the ability to talk from Tokyo in your country to New York in my country scarcely excites comment or wonder. It is not surprising that, to the layman, the element of distance seems the most striking factor in the technical development of the telephone art. As a matter of fact, while the conquest of distance has involved much scientific effort, and very ingenious and highly developed methods for the transmission of speech currents, the magic of the telephone still resides in the instruments which provide for the conversion of mechanical energy, namely speech sounds of highly complex wave form, into electrical currents of corresponding wave form, and the reverse process of converting these electrical currents into speech sounds. These instruments, the transmitter and the receiver, are basic to the whole telephone art. As they have been improved by development and design, it has become possible not only to render a higher grade of service but to effect economies in other portions of the plant. For example, the

* Another of three Iwadare Foundation lectures delivered during this past spring in Japan by Dr. Colpitts. One lecture was published in the April 1937 issue of this *Journal*.

very extensive use of fine-gauge cables in the plant of the Bell System was, to a large extent, made possible by the development of more efficient transmitters and receivers. Further perfecting of these instruments promises additional improvements in service and some further economies.

Telephony, restricting the term to ordinary two-way talking between individuals, involves an element not present in any other service. It does not greatly concern one customer of an electric light or power company whether another customer chooses to use inadequate or inefficient or poorly located lamps or other equipment. That is, each user of the service is, under any ordinary conditions, independent of all other users. In the case of telephony, however, the problem is entirely different; for one user of the telephone is greatly concerned with not only the apparatus furnished to any one with whom he has occasion to talk but also with other factors affecting the use of this apparatus, such as the amount of noise in the room where the apparatus is located, the user's habits of speech, and whether his ability to hear is normal. Telephone instrumentalities must therefore be so designed and the plant so engineered as to meet reasonably wide variations from what may be termed normal conditions, and ratings of performance should be similarly established.

I believe telephony in your country as in ours will find an increasingly wide field of service, and there is no single factor more important to a sound development of this art than the subscriber apparatus. With your permission, therefore, I will broadly outline certain work of the Bell Telephone Laboratories which has had a very direct bearing on these telephone instrumentalities and the form they are likely to assume. I will first discuss the research program which has been carried on in these laboratories, and then indicate to you the general trend which development and design have taken.

The research program basic to the development and design of transmission instruments has itself been a matter of development as a better understanding of the problems unfolded and as the need for research in this or that direction became apparent. The research problem basic to the development and design of transmission instruments may be described as having the following very broad scope: an understanding of their physical operation viewed as electromechanical structures; an understanding of speech mechanism and an accurate physical definition of speech air waves; an understanding of the hearing processes and a determination of how hearing is affected by factors present in telephony. Also, our research program may be said to have included research upon certain materials, the results of which have an important bearing either upon an understanding of the operation of these instruments or upon their practical design. In addition to the development of many methods of measurement and testing applicable to laboratory research and development, of very great importance has been a development of the testing methods which permit of a better final evaluation of the developments based upon the results of this activity.

Instruments as Electromechanical Structures

The telephone transmitter itself is a complex mechanical and electrical structure. Its general method of operation can be described qualitatively in relatively simple terms, but the operation of few structures is more difficult to define in definite quantitative terms and relationships. For example, we are concerned with acoustical problems such as those involved in the air connection between the lips of the speaker and the diaphragm of the instrument. This air connection may involve a short column of air as in those instruments which have a telephone mouthpiece. Connection between the column of air and the working parts of the transmitter may be partially closed by a perforated section. When we come to consider the operation of the instrument itself, there is involved the mechanical vibration of the diaphragm as it operates on the carbon, and further, the whole question of electric conduction in the small mass of granular carbon itself.

In the case of the receiver which converts telephonic currents into speech sounds, we have very similar acoustical, mechanical and electrical problems with the exception, of course, of the mechanical and electrical problems introduced by the carbon of the transmitter.

A large amount of research work has been carried on in the Laboratories relating broadly to the transmitter and the receiver as electromechanical physical structures. The theory of these devices as vibrating systems has been developed so that their overall performance can be related to the various structural features. Consequently, our development and design engineers are now enabled to predetermine by calculation how certain modifications in structure will affect the physical performance of the instrument. In other words, the design process has become very much less "cut and try."

Research has been undertaken and substantial progress has been made on a study of microphonic action in carbon. In order to develop a complete theory of the operation of the transmitter, it is necessary to understand fully what takes place between each carbon granule in the carbon chamber.

Speech Sounds

Let me outline briefly some of the results of these studies on speech. The source of any voiced sound is in the larynx. On both sides of this larynx there are two muscular ledges called the vocal cords. When we breathe, these two ledges are widely separated, but when a voiced sound is produced, they come close together, forming a long narrow slit. As they come close together, the air passing through the resulting slit is set into vibration producing a sound. It has been generally supposed that the pitch of the tone thus produced was determined by the natural frequency of vibration of the two vocal cords, and that by changing the tension of these cords, the pitch of the tone can be raised or lowered at will. As most of you know, their natural frequency of vibration is the rate that they would vibrate to and fro if they were plucked and set into vibration like a banjo string or an elastic band. Our studies revealed that the natural pitch of these cords while a tone is being produced is considerably below that of the pitch of the tone. It is true that the pitch of the tone produced is affected, somewhat, by the elasticity of the vocal cords, but it is principally controlled by the size of the air opening between them. The little plug of air between the two vocal cords vibrates through a very much larger amplitude than the amplitude of the cords themselves and is the real source of the sound. The mass of this small plug is controlled by the size of the opening and by the elastic forces pushing it to and fro—namely, the air pressures on either side of it. It is evident that these oscillating pressures will be influenced by the size and shape of the trachea leading into the lungs on one side and by the size and shape of the tongue, mouth, and nasal cavities on the other. The mechanical action involved is analogous to the electrical action in a vacuum-tube oscillator. The sound which is generated at the vocal cords is modified as it passes through the throat, mouth, and nasal passages. The real character of the sound which enables us to identify words is wholly dependent upon the manner in which this cord tone is modified by the changing sizes, shapes, and characters of these passages and the outlet to the outside air.

After the various speech sounds leave the mouth, they are transmitted to the ear of the listener by means of air vibration. As an example of the type of disturbance created in the air, consider the sentence, "Joe took Father's shoe bench out." This silly sounding sentence is chosen because it is used in our laboratories for making tests on the efficiency of telephone systems. The sentence, together with its mate, "She was waiting at my lawn," contains all the fundamental sounds in the English language that contribute appreciably toward the loudness of speech. As the sound wave produced by speaking this sentence travels along, each particle of air over which it passes executes a vibration through its original or undisturbed position. The successive positions occupied by the particle as it moves in the complicated series of vibrations corresponding to a spoken sound can be visualized in laboratory investigations from oscillographic records of the corresponding telephone currents.

Each successive particle of air along the line in which the sound is traveling executes a similar complicated series of vibrations but any particular oscillation is performed at a later instant by the particle which is farther away from the source of the sound. The disturbance in the air which represents a spoken sound may then be pictured

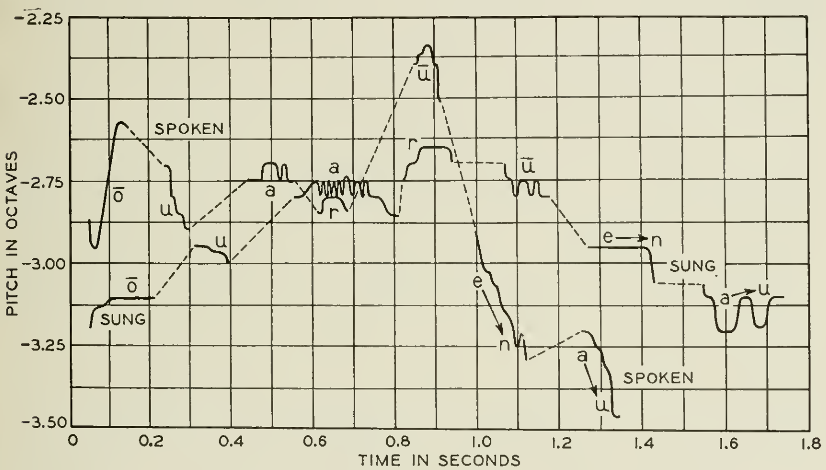


Fig. 1—Melodic curves showing the variation of pitch with time as the sentence "Joe took Father's shoe bench out" is spoken and sung.

either, as was first described, in terms of the successive positions of a single particle or in terms of the displacements at any instant of each of the particles along the line of travel of the sound wave. For example, for the sentence "Joe took Father's shoe bench out," the disturbance carrying the sound *j* in the word "Joe" is about fifteen hundred feet from the mouth by the time the sentence is finished. I have a record here which was taken in our laboratories which shows the intricate motion of each particle of air as this sentence is transmitted through the air.

If we analyze the wave when the sentence "Joe took Father's shoe bench out" is spoken, the variations in pitch of the speech sounds can be determined from the vibration rate. Such an analysis is shown in Fig. 1. The variations in pitch are represented on the vertical axis. The duration of the sounds in fractions of a second is represented on the horizontal axis. It will be seen that the pitch rises and falls as the various sounds are spoken. This representation of the pitch variation is called the fundamental melodic stream. It is the melody in the same sense as this term is used in music, although it is evident that the pitch changes do not take place in musical intervals as would be the case if the sentence were sung.

To show the contrast, a graph was made when the sentence was intoned on the musical intervals *do, re, mi, fa, mi, re, do.* An analysis of the graph gave the result shown in Fig. 1. In the case of the sung sentence the pitch changes are in definite intervals on the musical scale, while for the spoken sentence the pitch varies irregularly, depending upon the emphasis given. The pitch of the fricative and stop consonants is ignored in the musical score, and since these consonants form no part of the music, they are generally slid over, making it difficult for a listener to understand the meaning of the words. Some of our friends in the musical profession may object to this statement of the situation, but I think it will be agreed that a singer's principal aim is to produce beautiful vowel quality and to manipulate the melodic stream so as to produce emotional effects. To do this, it is necessary in singing to lengthen the vowels and to shorten and give less emphasis to the stop and fricative consonants. It is for this reason that it is more difficult to understand song than speech.

There are two secondary melodic streams of speech represented by the second and third curves from the bottom of Fig. 2, which are due to the resonances imposed upon the speech sound by the throat and mouth cavities. The numbers on these curves give the number of the harmonic which is reenforced. These two secondary melodic streams are not sensed as changes in pitch, but rather as changes in the vowel quality. Then there is a fourth stream, or, it would probably be better to say, a fourth series of interrupted sounds which are very high in pitch and are the sounds which enable us to identify the fricative consonants. The secondary melodic streams produced while speaking the same sentence are approximately the same for different persons, even for a man and a woman, while the fundamental melodic stream is usually quite different. This latter stream is not used in identifying words, but it is used sometimes to give different meanings to the same words.

As one listens to this sentence he hears the variations in loudness as well as in pitch. Loudness is related to the amplitudes and frequencies of the components of the tone, but this relationship is very complicated. It is dependent upon the action of the ear, including the nerve mechanism carrying the message to the brain. This relationship has been under study for a number of years so that we are now able to calculate from physical measurements the loudness for a typical ear and also to devise instruments for measuring approxi-

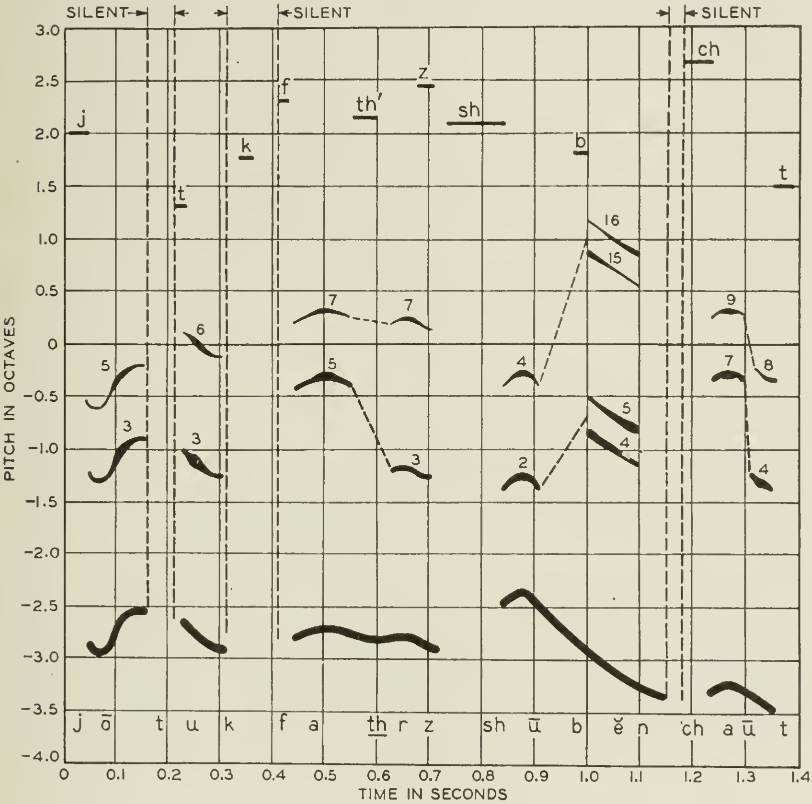


Fig. 2—Melodic curves showing the variation of pitch with time as the sentence "Joe took Father's shoe bench out" is intoned on the musical intervals do, re, mi, fa, mi, re, do. The pitch changes in regular intervals rather than in irregular intervals as shown in Fig. 1.

mately the loudness of any sound. The result of using such a device for recording the variations of loudness in the spoken sentence which we have been discussing is shown in Fig. 3. For comparison, the variations in pitch are also shown in this figure.

If the fifteen-hundred-foot wave carrying the sentence above mentioned could all be collected into an energy collector, the question arises, "How much energy would be involved?" It is not possible here to describe the devices by which we were able to measure accurately the energies and frequencies involved in speech, but the results of this research work are interesting. When this sentence is spoken fairly rapidly, it will contain about two hundred ergs of energy. About 500,000,000 ergs of energy pass through the filament of an ordinary incandescent lamp each second. This shows that the acoustic

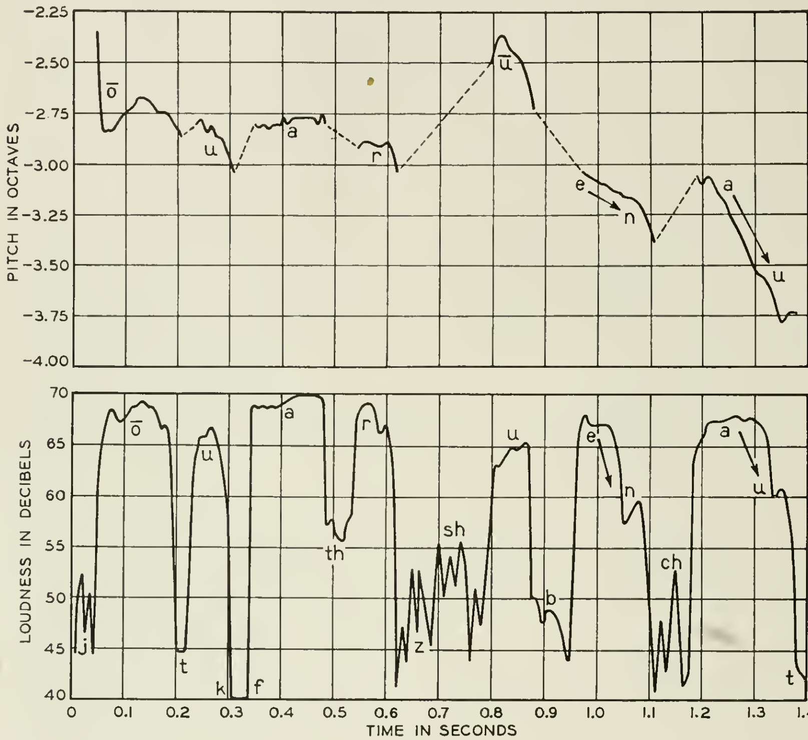


Fig. 3—Graph of the loudness of the various sound elements when the sentence "Joe took Father's shoe bench out" is spoken.

energy in this sentence is very small. Putting it in another way, it would require five hundred persons speaking this sentence continuously for a year to produce sufficient speech energy to heat a cup of tea.

An examination of the wave produced by this sentence shows that the vowels contain considerably more energy than the consonants. Exact measurements have indicated that in ordinary conversation the ratio of the intensity of the faintest speech sound, which is *th* as in "thin," to the loudest sound, which is *aw* as in "awl," is about one to five hundred. The actual power used in producing the various sounds depends, of course, upon the speaker and the emphasis with which he pronounces the sound. The power in an accented syllable is three or four times that in a similar unaccented syllable. Measurements upon a number of voices during a conversation have indicated that the average power in the speech produced is ten microwatts (one one-hundred-thousandth watt). Some speak with more and others with less than this power. In Table I is shown how various

TABLE I

Relative Speech Powers Used by Individuals in Conversation

Ratio of power of individual speakers to average power . .	Below 1/16	1/16 to 1/8	1/8 to 1/4	1/4 to 1/2	1/2 to 1	1 to 2	2 to 4	4 to 8	Above 8
Per cent of speakers	7	9	14	18	22	17	9	4	0

voices in a sample group vary from the average. It is seen that seven per cent speak with less than one sixteenth the average power, eighteen per cent use powers lying between one quarter and one half the average, and four per cent between four and eight times the average. No speakers were found to use more than eight times the average for conversational purposes.

Now let us consider the variations for a typical speaker. As a conversation proceeds, the speech power varies from zero during the silent intervals to peak values which frequently are one hundred times the average power. Extensive measurements of these peak powers upon a number of speakers indicated a distribution about the average as shown in Table II. For example, if we should examine the speech

TABLE II

Peak Powers in Conversational Speech

Power Boundaries In Terms of Average Power	Per Cent of Intervals
Below 1/8	12
1/8 to 1/4	4.0
1/4 to 1/2	4.5
1/2 to 1	5.5
1 to 2	8.3
2 to 4	12.7
4 to 8	18.6
8 to 16	17.0
16 to 32	10.5
32 to 64	5.1
64 to 128	1.7
Above 128	.1

during each one-eighth-second-interval throughout a typical conversation, we should find that for seventeen per cent of them the peak power would lie between eight to sixteen times the average over a long interval. It is seen that the most frequently occurring value of the peak power is about ten times the average.

Although a typical voice of a man and a typical voice of a woman are alike in that they use the same average power and variations of power from this average, they are different in other respects which we shall now consider. It is well known that the pitch of the voice of a woman is about one octave higher than that of a man. It was not known, however, until our experiments revealed it, that the intensity of the components having vibration rates above three thousand cycles per second was definitely greater for voices from women than from men. The following investigation shows the extent of this difference.

An apparatus has been devised in our laboratory which will receive the speech during a conversation and then sort out the components into groups depending upon their intensity and pitch. Those lying in each half-octave band on the pitch scale are automatically grouped together and the group power measured. Also, by means of another automatic device, a sorting process is accomplished within the group placing together all the components having powers between certain power boundaries so that they operate a particular recording meter. It was by means of an apparatus of this latter type that the results in Table II were obtained. It was found that the powers were distributed in each of these pitch bands in approximately the same manner as indicated in Table II for speech as a whole.

The relative values of the average speech power in each of the half-octave bands are shown in Fig. 4. The horizontal positions give the pitch in octaves above or below a tone having a vibration rate of one thousand cycles per second. The vertical positions give the fraction of the total power which comes into each half-octave band. For example, consider the half-octave from − 2.25 to − 1.75, which is the octave with its midpoint at middle "C" on the musical scale. The fraction of the power coming into this half-octave is about one quarter. It will be noted that for both types of voices the maximum power occurs in the second octave below one thousand cycles. This particular octave contains about one half of the total speech power. The octaves on either side of this one containing the maximum power contain slightly less than one quarter of the total power. No other octave contains more than about three per cent of the total power. It is seen that for the band of lowest pitch the voices from men contain about eight times the power of those from women. Also, as stated above, for pitches above one—that is, for tones having vibration rates above two thousand cycles per second—the voice power for women is greater than for men. For the half-octave in the region of pitch three octaves above one thousand cycles, it is about ten times greater.

For some reason which is not very evident, women use higher pitch sounds for producing the fricative consonants, and this results in the greater power shown in the regions of higher pitch. Every one who is familiar with such transmission systems knows well that these high-frequency components are nearly always eliminated. While

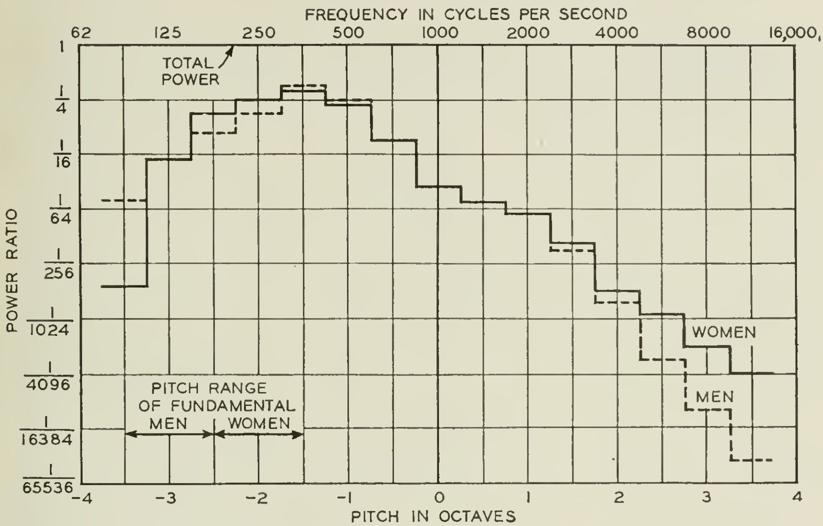


Fig. 4—Distribution of speech power in fractions of the total power for half-octave intervals above and below 1000 cycles.

these sounds are not of controlling importance in properly understanding speech, it is evident that the women's voices are somewhat handicapped as compared with men in systems which eliminate them.

Hearing

Paralleling our research on speech sounds, an investigation of hearing has been under way in Bell Telephone Laboratories. Broadly speaking, the aim has been to arrive at an accurate physical description and a measure of the mechanical operation of human ears in such terms that we may relate them directly to our electrical and acoustical instruments. We have measured the keenness of the sound-discriminating sense, and determined what is the smallest distortion which the mind can perceive, and how it reacts to somewhat larger distortions. This information is utilized in determining a reasonable basis of design both for separate instruments and for transmission systems as a whole, to give a proper balance between cost and performance.

I can only indicate a few of the important results of our investigations. One of the first steps was to determine in a quantitative way the performance of our ears as machines. It was obviously important to know how faint a sound the ear can hear, and also how loud a

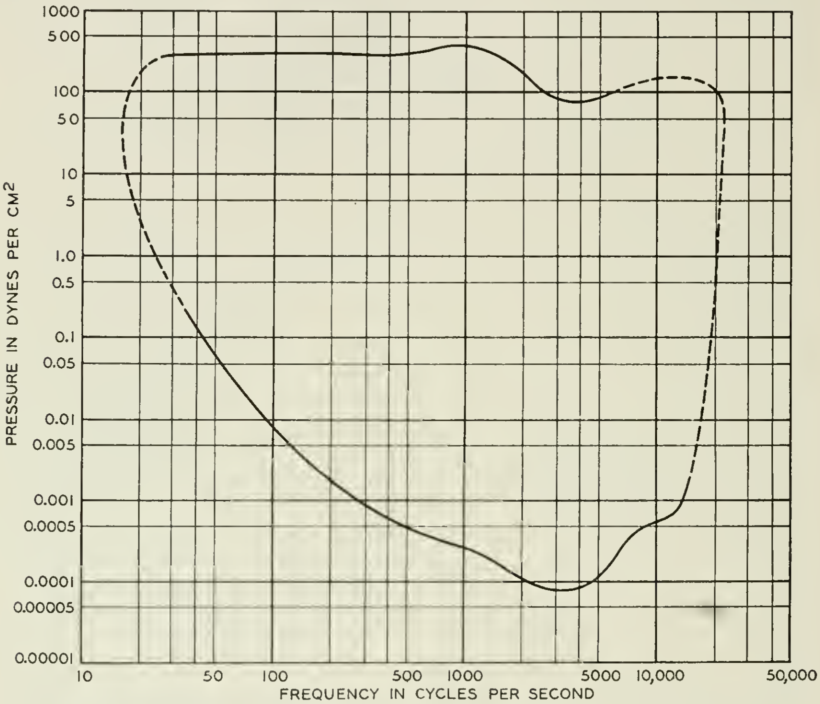


Fig. 5—Auditory sensation area for the typical ear of a young adult.

sound the ear can tolerate. With the advent of the vacuum tube, it was possible to develop methods of accurately measuring the intensity of faint sounds and of readily producing such sounds. Figure 5 gives the results of a large number of measurements made to determine the limits of hearing. This graph is called the auditory sensation area. The lower solid curve represents the minimum sound that an average young person can hear. The abscissa gives the frequency of the pure tone, and the ordinate the sound pressure in dynes per square centimeter. The top solid curve represents the maximum intensity of sound that the ear is capable of handling. This curve was determined by noting that intensity which produced a feeling sensation. Intensities slightly higher than this result in pain and in some instances serious injury to the ear. The dotted lines on either side complete the enclosure and represent the upper and lower limits of pitch that can be heard. It is obvious from this figure that the upper or lower limit of pitch is greatly dependent upon the intensity at which the sound is produced. It will be seen that near the middle range of frequencies, the pressure range is one million to one. The pitch range of pure tones is from about 16 to 25,000 cycles per second.

These results are for young adults, and it may be of interest to note that as one becomes older the hearing acuity, at the higher frequencies particularly, becomes less. In the table below is shown some measurements to determine what the effect of age would be upon the hearing acuity:

TABLE III

Db Loss in Hearing with Age

Frequency	60 to 1024 Cycles	2048 Cycles	4096 Cycles	8192 Cycles
Ages 20–29 (96 ears)........	0	0	6	6
Ages 30–39 (162 ears).......	0	0	16	11
Ages 40–49 (84 ears)........	0	2	18	16
Ages 50–59 (28 ears)........	0	5	30	32

These are average values obtained from measurements on a large number of persons.

Another important measurement of average hearing is that concerned with minimum perceptible differences in pitch and in intensity. Careful measurements on large groups of people have given us reliable data of this form. In Fig. 6 are shown the results of such measurements. They are plotted on the auditory sensation area. The ordinates are decibels above the reference pressure and the abscissas are centi-octaves above or below a pitch of 16.35 cycles per second. A frequency scale is also given for reference purposes. The numbers within the area indicate the minimum changes in the intensity level in db that the average ear is able to detect over that region of the auditory area. It will be seen that near the threshold fairly large changes are necessary to be perceptible, while at fairly high intensities about 1/4 decibel is all that is necessary for the change to be perceived.

In Fig. 7 are given similar data for minimum perceptible differences in pitch. The numbers in the figure in this case are given in centi-octaves; that is, each unit corresponds to 1/100 of an octave. The results of this line of investigation have an important bearing on the physiological theory of hearing which I cannot enter into, and another important result has been the development of methods of determining the degree of impairment of hearing.

In telephony we are, of course, not directly concerned with simple sounds, but with the highly complex sounds of speech, and these are

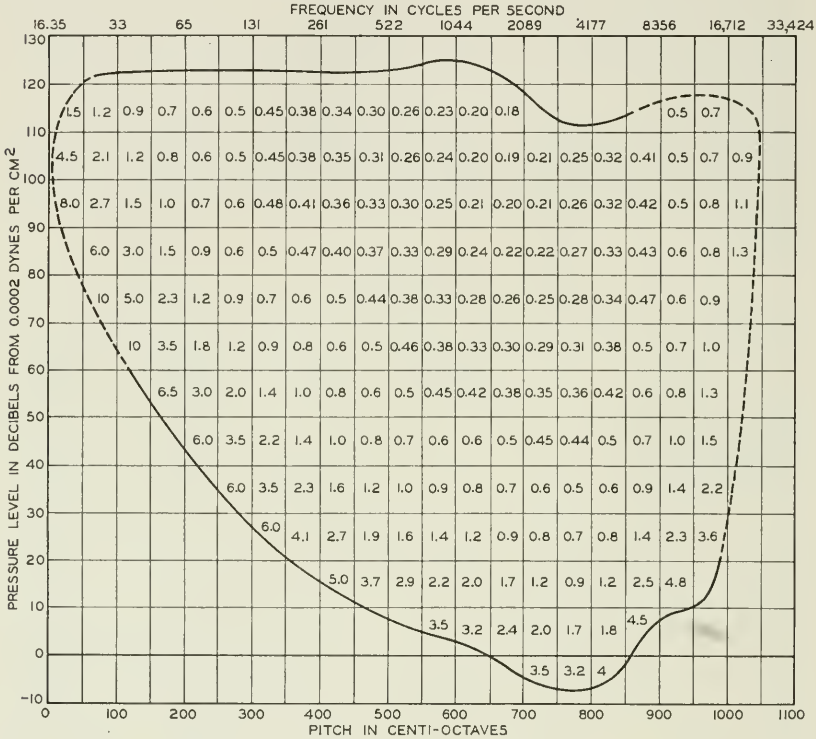


Fig. 6—Minimum changes in intensity level in db that the ear is able to appreciate at various positions in the auditory area.

on actual telephone circuits generally associated with extraneous sounds which we may group under the broad term of noise. Further, telephone instruments are not perfect, and could be made to approach perfection only at a great expense. In order to arrive at a quantitative understanding of the importance of departures from perfection in telephone transmission elements and conditions of use, we have in very general terms proceeded as follows: We set up transmission systems so nearly perfect that even the keenest ear could not find a flaw in their transmission performance, and then introduce measured imperfections or variations.

By this general process, we were able to determine the effect of noise of chosen intensities either as noise present in the telephone receiver or as noise in the room. Similarly, the effect of a line or other characteristic such that voice frequencies above a certain value or below a certain value were not transmitted, was determined. The

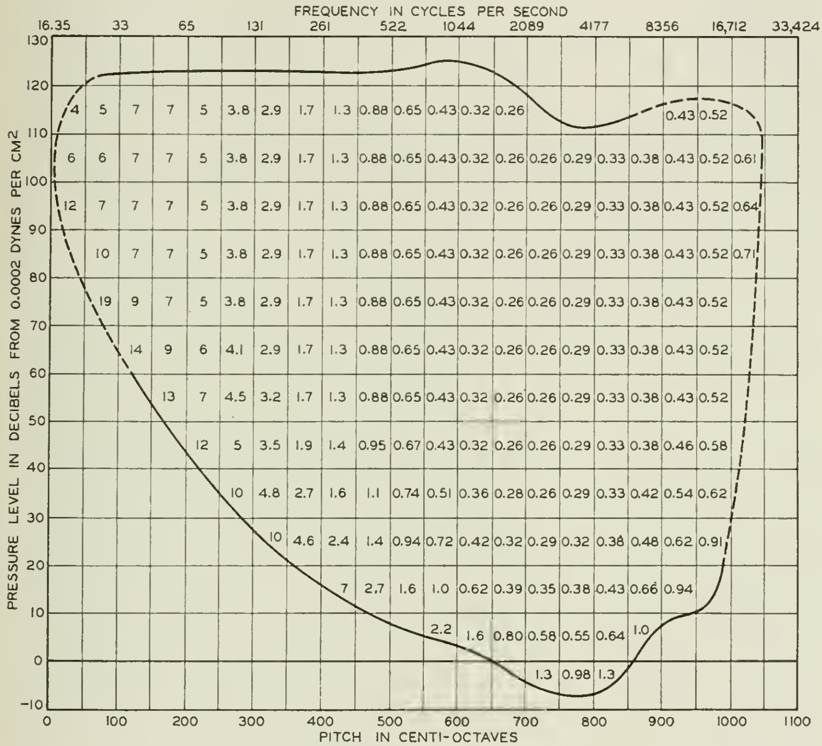


Fig. 7—Minimum perceptible differences in pitch in centi-octaves for various positions of the auditory area.

effect of introducing a highly resonant element or of a non-linear element was studied. The range in loudness of speech necessary for best reception was likewise measured.

As noise became recognized as a very real factor, a standard basis for noise measurements was established. Consequently we are now able to measure noise on a telephone circuit or in a room, and state the result in terms of a standard unit.

Materials

In the practical design of modern telephone instruments we owe a large debt to the chemist and the metallurgist. Modern molding materials and processes are utilized in order to secure forms of apparatus satisfactory from the standpoint of appearance and of mechanical strength. The newer types of permanent magnet steel, to the development of which your countrymen have contributed so largely, provide possibilities of light-weight and very efficient magnetic structures.

It is a most striking circumstance that commercial telephony is dependent upon the performance of a small mass of carbon granules in the transmitter. No single material entering into the construction of telephone apparatus has therefore greater importance. In America at least, transmitter carbons are largely derived from a certain specially selected anthracite coal. In its natural state, this coal exhibits none of the characteristics required for its use in a transmitter. These characteristics or properties are secured by heat treatment. These heat-treatment processes were for many years the result of empirical development and were not well understood or, as we now recognize, adequately controlled. This resulted in a product of uncertain quality. An important task of the Laboratories was therefore to study each step in the process of producing carbon and to develop a process definitely specified at each step, which would be capable of giving the desired uniform quality. The results so far obtained have had very important reactions upon transmitter performance. The Laboratories have also set themselves the more elementary task of understanding the fundamental properties of carbon contacts. One important element of this research is to determine the causes of resistance changes produced when the compressive force on a mass of carbon granules is changed. It is too early to report results from this research, but it seems clear that granular carbon will be an important element in the design of transmitters for many years to come, and we should seek to obtain complete fundamental knowledge of its operation.

Testing Methods

Broadly speaking, methods of testing have been developed, first, to enable the development and design engineer to determine quantitatively the various performance factors of the apparatus under development, and second, to determine how well the apparatus which has been developed performs under service conditions. In the Laboratories, we have over the last twenty years developed methods for measuring the physical constants of the apparatus involved so that we can analyze this apparatus as electromechanical structures. Further, in the design of telephone transmission apparatus we are concerned with a power transmission system in which the design engineer has no control over the power source, the human voice, nor over the receiving agency, the human ear. His control is limited to the conveyance of power from speaker to listener. In the Laboratories, therefore, we have recognized that it is necessary, at least without present knowledge, to supplement physical measurements by measurements involving speech sounds and the human ear. Some years ago, these tests consisted of comparisons between different instruments or transmission elements made by the process of talking first over a circuit containing, for example, one instrument, and then over the same circuit containing a different instrument. Dependence was placed wholly upon the listener's skill to detect differences in volume, quality and intelligibility. It was recognized that this method of testing left much to be desired. Owing to the limitations of the human ear, small volume differences could not be detected, but even more important, this simple test furnished no very accurate measure of speech distortion affecting intelligibility, and obviously no definite information as to the relation between volume, various types of distortion, and overall effectiveness.

Dr. George A. Campbell, in 1910, proposed a method of testing which has been highly developed in our Laboratories. This method, termed "articulation testing," measures the relation between the reproduced and impressed sounds from the standpoint of effects on intelligibility of different kinds of distortion. This method has been described in a number of publications. Briefly, in this method, lists of syllables chosen at random and usually meaningless monosyllables are called over the circuits to be rated, and the percentage of syllables correctly understood gives a measure of the circuit performance. Further, the method has been extended to give quantitative measures in terms of the recognizability of reproduced speech sounds, of the effects of loudness of these sounds, and of the noise which may be present.

While various physical tests and the articulation test method are exceedingly useful tools in the hands of the research and development engineer, they do not give a direct measure of the transmission service performance of a circuit in terms of the ability of the user to carry on a conversation under actual commercial conditions. This ability of the user to carry on what may be termed a successful telephone conversation depends not only upon the performance of the telephone instruments and circuits but also, to a substantial extent, upon the users' own performances—the subject material of conversation, how they talk into the transmitters, and how they hold the receivers—and upon the room noise conditions. In other words, there are a number of factors random in nature which, while beyond the control of those who design and engineer the telephone plant, must be taken account of in rating the service performance.

A large amount of thought and effort has been given to the problem of how best to determine transmission service performance. Very briefly stated, we have been led to the following steps: In order to take suitably weighted account of all the factors involved, service performance ratings should be based on service results, that is, transmission service performance should be measured by the success which users of the telephone circuit have in carrying on conversations over the circuit. With the various factors in mind, we have fixed upon what we have termed "effective transmission" ratings for transmission plant design. These ratings are based on a determination of the *repetition rate* in normal telephone conversations.

As the effect of a change in a circuit depends upon its initial characteristics, it is necessary in order to be able to compare numerical results to have a basic circuit for reference. By suitable choice of basic circuit, it is possible to express the effects of changes in any one transmission characteristic in terms of the attenuation of the trunk. For example, the effect of changes in sidetone level in the subscriber's set can be expressed as so many decibels change in trunk attenuation. Mr. W. H. Martin's paper, "Rating the Transmission Performance of Telephone Circuits," in the *Bell System Technical Journal*, January, 1931, discusses the method and general principles. It should be noted that the application of the method requires careful consideration of many factors and the accumulation and analysis of a very substantial amount of data. Based on these data, we have arrived at the following relationship:

$$\text{Relative effective loss in db} = 50 \log_{10} (r)$$

where r is the ratio of the repetition rates for the two conditions compared.

Association of Transmitter and Receiver

In order to furnish a convenient two-way talking circuit over a single pair of wires, the transmitter and the receiver at each end of the circuit must be continuously associated in the circuit. This has been accomplished by various circuit arrangements since the early days of the telephone, and as every user of the telephone knows, leads to the condition that when speaking into the transmitter one hears his own voice in the receiver. Local speech so heard is designated as sidetone. The Laboratories have carried on research in order to determine the effect of sidetone on the overall efficiency of the circuit. We find that sidetone above a certain volume decreases the conversational efficiency of the circuit. Parallel with the study of the effects of sidetone, research has been carried on on methods which could be applied to limit sidetone in amount to more nearly its optimum value. This has led to the development of what are known as anti-sidetone circuits, which do not eliminate sidetone but reduce it to an amount which is more nearly that found to be desirable.

An important step in the association of the transmitter and the receiver is represented by the handset which provides a rigid mechanical connection between the two units. This rigid mechanical connection introduces mechanical coupling between the receiver and the transmitter, which had to be given very serious consideration in order to avoid speech distortion.

Trends in Instrument Development

I have broadly indicated to you fields of research which underlie the development and design of the telephone transmitter and receiver. It will now be of interest for us to note what application is likely to be made of the results of what has amounted to an enormous total of scientific effort. In this connection, it may be well again to emphasize that station apparatus is intimately associated with the user, and has therefore to be designed to fit him, his habits of using the telephone, and the conditions attending such use. The handset has to be designed to fit his head, the holes in the dial to fit the size of his finger, the bell to be loud enough, and so on. Our effective transmission rating system has been set up in an attempt to rate the performance of the telephone when employed by the customer in the way he wants to use it, under the conditions surrounding him. For this reason, this method of rating has been found particularly valuable in the development work on instruments.

Because of the wide range of customer usage and conditions, a number of factors have to be taken into account in the design of the apparatus. Also, because this apparatus is located on the customer's premises, where it is relatively inaccessible to the telephone personnel, it must be capable of standing up without undue trouble under this wide range of usage and conditions. To strike a proper balance in meeting all these factors requires an intimate knowledge of the field conditions as well as of the development and manufacturing possibilities. A continuing close contact with field experience is employed to modify the designs towards securing the proper balance to meet these factors.

In order to indicate more clearly the present trends in design, I shall refer briefly to the earlier art. In the early development of transmitters and receivers, the matter of getting efficiency was of primary importance since this could be evaluated directly in terms of the amount of copper required in the connecting line. The early transmitters, which were of the same construction as the receiver, depended on the generator action of a diaphragm and coil and developed sufficient power to be heard over only a few miles of heavy-gauge wire. Some amplification was necessary before telephone communication could begin to assume the proportions of a widespread service. This amplification was obtained at a reasonable cost in the carbon contact transmitter. Transmitters of this type are in the order of 60 db more efficient as transducers of acoustic to electric energy than the earlier type.

Both the transmitter and the receiver operate by means of diaphragms which have natural periods of vibration. These resonances and the resonances of the air spaces on each side of the diaphragm were used to obtain as efficient a transfer of energy as possible. In the early design, a great deal of attention was also given to locating these resonances at the portion of the frequency range where they would tend to increase the intelligibility of the reproduced sound. As a result, both instruments were made very efficient in the region of 1000 cycles, which lies within the range where the ear and the sensation of loudness are most sensitive.

It was recognized that these resonances caused undesirable distortion, but under the conditions the resulting increase in efficiency more than compensated for this disadvantage. As time went on, the diaphragm resonances came to be looked upon as practically inherent in commercial transmitters and receivers, because no way was known of eliminating them without making a very material sacrifice in the efficiency of the instrument.

About twenty years ago, the development of the vacuum tube amplifier and the high quality condenser transmitter made it possible to demonstrate and measure quantitatively the advantages of reducing distortion. These high-quality instruments, the improvement in measuring technique and the development of improved methods of designing vibratory systems offered the promise of providing instruments in which the resonance effect could be reduced without unduly affecting efficiency.

The first commercial instrument for station use, which demonstrated the possibility of carrying out this promise, was the transmitter employed in the handset first supplied by the Bell System in 1927. This transmitter had to meet the requirement of giving the same transmission service as transmitters of the deskstand type, and at the same time meet the very exacting requirements imposed by the handset to make it free from howling and capable of performing over a wide range of positions. The diaphragm resonance was damped to a large extent by the use of paper rings and, by lightening the structure, the point of maximum response was moved up in frequency so that it no longer coincided with the peak of the receiver. The effect of this was not only to broaden the response characteristic and improve intelligibility, but also to reduce the gain in the local howling circuit which is, of course, a maximum when both transmitter and receiver have their greatest efficiency at the same frequency. The same separation of peaks resulted in the received speech being less loud, but in spite of this the overall performance was equivalent to that of the best deskstand type of instrument then available.

With this accomplishment, further work was directed toward maintaining the lower distortion and increasing the efficiency. The transmitter introduced in 1934 represented a marked improvement along this line. This instrument still further broadens the transmitted frequency range and is used with about the same efficiency in desk-stands, handsets, wall sets, and coin-collect sets.

A new type of handset will be introduced in the Bell System in 1937 which, in addition to having a more pleasing and simplified design, will incorporate the new transmitter mounted in such a way as to make fullest use of its ability to transmit efficiently over a wide-frequency band.

During this evolution of the transmitter, the knowledge which had been gained as to the importance of transmitting different widths of frequency band over commercial telephone circuits led to the establishment of the range from 250 to 2750 cycles for designs of new circuits. It was not the intention in the establishment of this range that circuits should not do better than this where it is possible without materially increasing cost, but that all circuits should be at least as good as this. The establishment of this frequency range took into account a number of factors of which a very important one is that the overall utilization of this range from the sound entering the transmitter to the sound output of the receiver provides a grade of transmission which is highly satisfactory for the reproduction of conversational material.

The establishment of this frequency range played a part not only in the design of circuits, but also in guiding the evolution of the transmitter and receiver. The transmitter last referred to meets this requirement very well. In fact, its efficiency is fairly uniform for a frequency range extending beyond 4000 cycles.

The next step in the process was to improve the performance of the receiver. A pronounced resonance at 1000 cycles was no longer necessary since means had been found to improve the efficiency of instruments in other ways than by concentrating all the resonances at one frequency. The importance of the higher frequencies in transmitting and reproducing the transient sounds characteristic of the consonants in speech led to placing more emphasis on these frequencies and attempting to produce more uniformly the band of frequencies which was set as a limit for circuits. This has now been accomplished in a practical fashion in the receiver which is being introduced in 1937.

The effect of this evolution in the design of station instruments may be brought out by a comparison of the overall response characteristic —that is, the relation of the sound delivered to the ear to the sound available at the transmitter—for a typical telephone connection having, in one case, both terminal instruments of the 1920 type and, in the other case, the terminal instruments of the coming new 1937 type. In this typical circuit, the trunk has been taken as free from distortion so that its effect will not influence the indicated performance of the instruments, although the circuit does include two 22-gauge loops each three miles long.

At the resonance point of the old instruments, just over 1000 cycles, the overall response in going to the new instruments is reduced by almost 30 db while the response in the range from 2000 to 3000 cycles is increased by over 20 db. In the frequency range from 500 to 2000 cycles, the circuit employing the older instruments shows a variation of overall response of over 30 db. For the new type, the variation for this same frequency range is reduced to 15 db, and, furthermore, this variation of 15 db applies approximately for the range of 250 to 2750 cycles which was mentioned as the transmission range requirement for the design of new circuits. In regard to the variation of 15 db in this frequency range, there is good indication that this response is more desirable than one of no variation, from the standpoint of having the telephone performance approach that of direct air transmission.

In addition to these improvements in frequency response and efficiency, the intensive development program on these instruments has improved materially the stability of the carbon transmitter under service conditions. This is an important factor in extending the useful life of these instruments and in reducing the cost of maintaining the desired transmission performance.

You will perhaps pardon me if, in concluding, I say a few words which I hope will not seem unduly laudatory of the work of my associates in the Bell Laboratories. The facts seem to be that twenty years ago or thereabouts, there was very little general scientific interest in sound and sound devices. As a result of work begun in

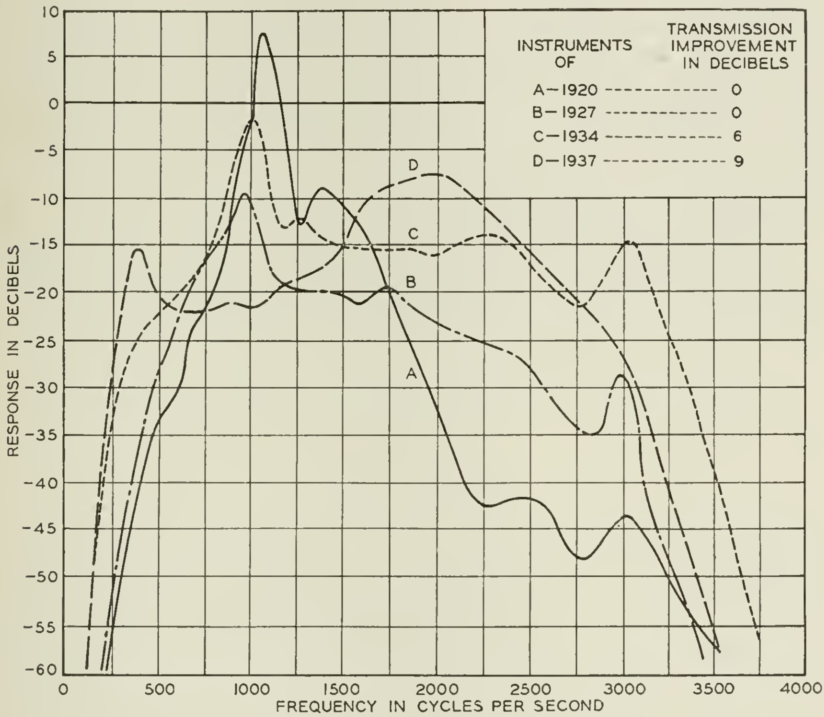


Fig. 8—Comparison of the response-frequency characteristics of telephone instruments since 1927.

these Laboratories, and as the possibilities of interesting and important applications became apparent, broad scientific interest was stimulated, and we have seen and welcomed increasing research activity in sound and acoustics in many of the university laboratories and in new industries based upon the results of scientific research in sound initiated by us. A number of my associates have attained world-wide recognition for their scientific and technical accomplishments. Our scientific investigations were undertaken to enable us to develop further the telephone art, and the results of these investigations are serving to guide us not only in the development of telephone instruments but in all developments of telephone transmission. The Laboratories' scientific and design work has contributed in large measure to the improvement of methods of recording and reproducing sound in the phonograph and sound-picture arts. The art of radio broadcasting owes a large debt to the work of the Laboratories, not only for the fundamental scientific knowledge contributed but also for actual instrumentalities employed. To those with impaired hearing, the Laboratories' investigations have made possible improved means for determining the extent of their impairment, and improved hearing aids. Finally, at least in America, we are becoming what I may term as "noise conscious." In our cities, noise is being recognized as a factor affecting comfort, efficiency, and possibly even health. The development of accurate methods for the measurement of noise is contributing to studies looking towards the reduction of noise.

Lecturer's Note: The lecturer wishes to acknowledge assistance given in the preparation of this material, particularly by Dr. Harvey Fletcher and Mr. W. H. Martin of the Bell Telephone Laboratories' staff.

The Use of Coaxial and Balanced Transmission Lines in Filters and Wide-Band Transformers for High Radio Frequencies

By W. P. MASON and R. A. SYKES

At the high radio frequencies, filters and transformers become difficult to construct from conventional electrical coils and condensers, on account of the small sizes of the elements, the large effects of the interconnecting windings and the low ratios of reactance to resistance realizable in coils. It is shown in this paper that selective filters and wide-band transformers can be constructed using transmission lines and condensers as elements. The ratio of reactance to resistance in these elements can be made very high; consequently very selective filters and transformers with small losses, can be constructed. The effect of the distributed nature of the elements is taken account of in the design equations and methods are described for obtaining single-band filters and transformers. Experimental measurements of such filters and transformers are shown. The experimental loss curve is shown of a coaxial filter used in the Provincetown-Green Harbor short-wave radio circuit for the purpose of connecting a transmitter and receiver to the same antenna.

I. Introduction

AT the higher radio frequencies, coil and condenser networks become difficult to construct on account of the small sizes of the elements and the large effects of the interconnecting windings. The Q realizable in high-frequency coils is about the same as can be obtained at the lower frequencies but due to the smaller percentage band widths, it is desirable to obtain a higher Q. There has been a tendency to replace coils by small lengths of transmission lines, and these have been used to some extent as tuned circuits, and as single-frequency transformers.[1,2,3]

It is the purpose of this paper to describe work which has been done in constructing selective filters and wide-band transformers from lengths of transmission lines and condensers. Due to the high ratio of reactance to resistance obtainable in both of these types of elements,

[1] "Transmission Lines for Short-Wave Radio Systems," E. J. Sterba and C. B. Feldman, *B. S. T. J.*, Vol. XI, No. 3, July 1932, page 411.

[2] "Resonant Lines for Radio Circuits," F. E. Terman, *Elec. Engg.*, Vol. 53, pp. 1046–1053, July 1934.

[3] "A Unicontrol Radio Receiver for Ultra-High Frequencies Using Concentric Lines as Interstage Couplers," F. W. Dunmore, *Proc. I. R. E.*, Vol. 24, No. 6, June 1936.

very selective networks can be obtained at the high radio frequencies. The effect of the distributed nature of the elements is considered and methods are described for obtaining single-band filters and transformers. Experimental measurements of such filters and transformers are shown, and these results indicate that such structures should be of use in short-wave radio circuits.

II. Characteristics of Transmission Lines

To facilitate an understanding of the following discussion, the equations of transmission lines as they apply to filter structures will be briefly reviewed first. The equations of propagation for any uniform transmission line can be expressed in the form of equations between the output voltage e_2, the output current i_2, the input voltage e_1, and the input current i_1 by the relations

$$\begin{aligned} e_2 &= e_1 \cosh Pl - i_1 Z_0 \sinh Pl, \\ i_2 &= i_1 \cosh Pl - \frac{e_1}{Z_0} \sinh Pl, \end{aligned} \tag{1}$$

where l is the length of the line, P the propagation constant, and Z_0 the characteristic impedance of the line. In terms of the distributed resistance R per unit length of line, L the distributed inductance, G the distributed conductance, and C the distributed capacitance, P and Z_0 can be expressed by the relations

$$P = \sqrt{(R + j\omega L)(G + j\omega C)}; \qquad Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}}, \tag{2}$$

where ω is 2π times the frequency.

The distributed conductance G is usually very low and can be neglected for coaxial or balanced transmission lines in dry atmospheres. For copper coaxial lines, the values of R, L, and C have been calculated [4] to be

$$\begin{aligned} R &= 41.6 \times 10^{-9} \sqrt{f} \left(\frac{1}{a} + \frac{1}{b} \right) \text{ ohms per centimeter,} \\ L &= 2 \log_e \frac{b}{a} \times 10^{-9} \text{ henries per centimeter,} \\ C &= \frac{1.11 \times 10^{-12}}{2 \log_e \dfrac{b}{a}} \text{ farads per centimeter,} \end{aligned} \tag{3}$$

where b is the inside radius of the outer conductor and a the outside radius of the inner conductor. If we define the Q of the conductor as

[4] See reference 1, page 415 and page 417.

the ratio of the series inductive reactance to the series resistance, the ratio will be

$$Q = .302b\sqrt{f}\,(\log_e k)/(k+1), \tag{4}$$

where $k = b/a$. It will be noted that this is the value of Q measured for a short-circuited conductor for low frequencies. As an example, a conductor 3 inches in diameter with the optimum ratio of $k = 3.6$ will have a Q of 3,200 at 100 megacycles. The value of the characteristic impedance Z_0 of a coaxial line is

$$Z_0 \doteq \sqrt{\frac{L}{C}} = 60 \log_e \frac{b}{a} \cdot \tag{5}$$

For a balanced transmission line the values of R, L, and C are,[4] if D is much larger than a,

$$\begin{aligned} R &= \frac{83.2 \times 10^{-9}\sqrt{f}}{a} \text{ ohms per centimeter,} \\ L &= 4 \log_e \frac{D}{a} \times 10^{-9} \text{ henries per centimeter length,} \\ C &= \frac{1.111 \times 10^{-12}}{4 \log_e \dfrac{D}{a}} \text{ farads per centimeter,} \end{aligned} \tag{6}$$

where D is the spacing between wires, and a the radius of one of the pair of wires. With these values, the expressions for Q and Z_0 become

$$Q = .302\sqrt{f}a \log_e \frac{D}{a}\,; \qquad Z_0 = 120 \log_e \frac{D}{a} \cdot \tag{7}$$

Another combination of some interest is obtained by using the inside conductors of two coaxial conductors adjacent to each other. Such a construction results in a balanced and shielded transmission line. All of the constants are double those given by equation (3), except for the capacitance which is halved.

III. Filters Employing Transmission Lines as Elements

One of the first uses of transmission lines as elements in wave filters is described in a patent of one of the writers.[5] In this patent are considered the characteristics obtainable by combining sections of trans-

[5] U. S. Patent 1,781,469 issued to W. P. Mason. Application filed June 25, 1927; Patent granted Nov. 11, 1930. Wave filters using transmission lines only are very similar to acoustic wave filters as pointed out in an article, "Acoustic Filters," *Bell Laboratories Record*, April 1928. Most of the equations and results of a former paper on acoustic filters, *B. S. T. J.*, April 1927, p. 258, are also applicable to transmission line filters.

mission lines in ladder filter structures. The results obtained are briefly reviewed here.

One of the simplest filters considered is shown on Fig. 1. The filter consists of a length $2l_1$ of transmission line shunted at its center by a short-circuited transmission line of length l_2. To determine the transmission bands of a filter, it is necessary to neglect the dissipation occurring in the elements and hence we assume that R and G are zero. In any case these values are small for transmission lines since they have a high Q. Neglecting R and G, equations (1) become

$$
\begin{aligned}
e_2 &= e_1 \cos \frac{\omega l}{v} - j i_1 Z_0 \sin \frac{\omega l}{v}, \\
i_2 &= i_1 \cos \frac{\omega l}{v} - j \frac{e_1}{Z_0} \sin \frac{\omega l}{v},
\end{aligned}
\tag{8}
$$

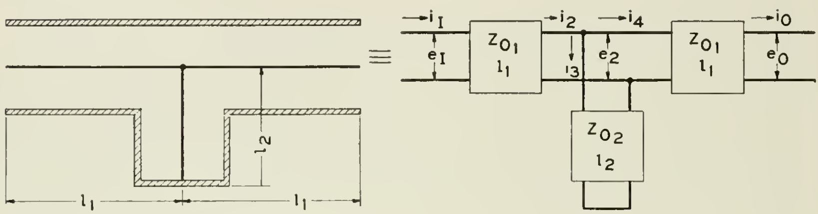


Fig. 1—Band-pass filter constructed from coaxial conductors.

where v, the velocity of propagation, and Z_0, the characteristic impedance, have the values

$$
v = \frac{1}{\sqrt{LC}}; \qquad Z_0 = \sqrt{\frac{L}{C}}. \tag{9}
$$

Using these equations the characteristics of the filter illustrated by Fig. 1 are easily calculated. With reference to Fig. 1 and equations (8) we can write the equations of the network as

$$
\begin{aligned}
e_2 &= e_I \cos\frac{\omega l_1}{v} - j i_I Z_{0_1} \sin\frac{\omega l_1}{v}, \\
i_2 &= i_I \cos\frac{\omega l_1}{v} - j \frac{e_I}{Z_{0_1}} \sin\frac{\omega l_1}{v}, \qquad & i_2 &= i_3 + i_4, \\
e_0 &= e_2 \cos\frac{\omega l_1}{v} - j i_4 Z_{0_1} \sin\frac{\omega l_1}{v}, \qquad & i_3 &= \frac{e_2}{j Z_{0_2} \tan \frac{\omega l_2}{v}}. \\
i_0 &= i_4 \cos\frac{\omega l_1}{v} - j \frac{e_2}{Z_{0_1}} \sin\frac{\omega l_1}{v},
\end{aligned}
\tag{10}
$$

Combining these equations and eliminating all terms except the input and output voltages and currents we have

$$e_0 = e_I \left[\cos\frac{2\omega l_1}{v} + \frac{Z_{01}}{2Z_{02}} \frac{\sin\dfrac{2\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}} \right] - j i_I Z_{01} \left[\sin\frac{2\omega l_1}{v} + \frac{Z_{01}}{Z_{02}} \frac{\sin^2\dfrac{\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}} \right];$$

$$i_0 = i_I \left[\cos\frac{2\omega l_1}{v} + \frac{Z_{01}}{2Z_{02}} \frac{\sin\dfrac{2\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}} \right] - j \frac{e_I}{Z_{01}} \left[\sin\frac{2\omega l_1}{v} - \frac{Z_{01}}{Z_{02}} \frac{\cos^2\dfrac{\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}} \right]. \tag{11}$$

The properties of symmetrical filters such as considered here are usually specified in terms of the propagation constant Γ and the iterative impedance K. These are similar to the line parameters used in equation (1) and it can be shown that the same relations exist between the input and output voltages and currents that exist in equation (1). Hence comparing equation (1) with equation (11) we find

$$\cosh \Gamma = \cos\frac{2\omega l_1}{v} + \frac{Z_{01}}{2Z_{02}} \frac{\sin\dfrac{2\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}},$$

$$K = Z_{01} \sqrt{\frac{1 + \dfrac{Z_{01}}{2Z_{02}} \dfrac{\tan\dfrac{\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}}}{1 - \dfrac{Z_{01}}{2Z_{02}} \dfrac{\cot\dfrac{\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}}}}. \tag{12}$$

The propagation constant Γ has the same significance as for a uniform line, namely that

$e^{-\Gamma} = \dfrac{e_n}{e_{n-1}} = \dfrac{i_n}{i_{n-1}}$ = ratio of currents or voltages between any two sections in an infinite sequence of sections, while K, the iterative impedance, is the impedance measured looking into an infinite sequence of such sections. The propagation constant Γ is in general a complex number $A + jB$. The real part is the attenuation constant and the complex part the phase constant. For a pass band, A must be zero, which will occur when $\cosh \Gamma$ is between $+1$ and -1. Hence to locate the pass band of the dissipationless filter, the value of the first equation of (12) must lie between $+1$ and -1.

In order to point out what types of filters result from a combination of transmission lines, two simplifying cases are considered. The first case is when $l_1 = l_2$. Then the equation for the propagation constant becomes

$$\cosh \Gamma = \left(1 + \frac{Z_{0_1}}{Z_{0_2}}\right)\left(\cos^2 \frac{\omega l_1}{v}\right) - \sin^2 \frac{\omega l_1}{v}$$
$$= \cos \frac{2\omega l_1}{v} \times \left[1 + \frac{Z_{0_1}}{2Z_{0_2}}\right] + \frac{Z_{0_1}}{2Z_{0_2}}. \quad (13)$$

The edges of the pass band occur when

$$\tan \frac{\omega l_1}{v} = \sqrt{\frac{Z_{0_1}}{2Z_0}} \quad (14)$$

and the centers of the bands occur when

$$\sin \frac{\omega l_1}{v} = 1; \quad \frac{\omega l_1}{v} = \frac{(2n+1)\pi}{2} \quad \text{and} \quad f_m = \frac{(2n+1)v}{4l_1}, \quad (15)$$

where $n = 0, 1, 2,$ etc. A plot of the propagation constant for several ratios of Z_{0_1}/Z_{0_2} is shown in Fig. 2. As is evident the filter is a multi-band filter with bands centered around odd harmonic frequencies. The ratio of Z_{0_1} to Z_{0_2} determines the band width of the filter. This ratio cannot be made very large because the characteristic impedance of a coaxial line or a balanced line cannot be widely varied from the mean value. For example when the ratio of b/a varies from 1.05 to 100 the characteristic impedance of a coaxial conductor changes from about 3 ohms to 275 ohms. This represents about as extreme a range as can be obtained. For a balanced conductor the impedance may range from 90 to 1100 ohms by taking extreme values of the ratio D/a. Taking 100 as the extreme range between Z_{0_1} and Z_{0_2} the band width for this case cannot be made less than 20 per cent.

We next consider the case given by taking $Z_{0_1} = 2Z_{0_2}$. For this case the filter parameters take on the simple form

$$\cosh \Gamma = \frac{\sin \dfrac{\omega(2l_1 + l_2)}{v}}{\sin \dfrac{\omega l_2}{v}} \text{ and } K = Z_{01}\sqrt{-\tan\frac{\omega l_1}{v}\tan\frac{\omega(l_1+l_2)}{v}}. \quad (16)$$

For this case the lower side of the pass band occurs when $\cosh \Gamma = +1$ and the upper side when $\cosh \Gamma = -1$. Hence the frequency limits

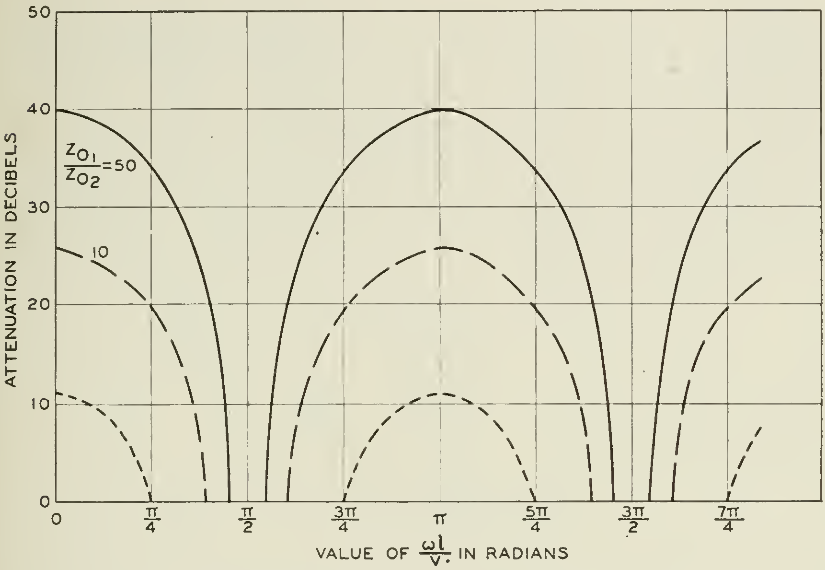


Fig. 2—Attenuation characteristic of band-pass filter.

of the first pass band can be obtained by solving the transcendental equation

$$\sin\frac{\omega(2l_1 + l_2)}{v} \mp \sin\frac{\omega l_2}{v} = 0,$$

which has the solutions

$$f_1 = \frac{v}{4(l_1 + l_2)}; \qquad f_2 = \frac{v}{4l_1}. \quad (17)$$

Hence by making the shunting line very short, it is possible to obtain a narrow band with this type of filter. The mid-frequency of the band occurs when

$$\sin\frac{\omega(2l_1 + l_2)}{v} = 0 \qquad \text{or when} \qquad f_m = \frac{v}{4l_1 + 2l_2}.$$

The characteristic impedance of this type of filter becomes very high for narrow-band filters as is shown by equation (16). At the mean frequency f_m the impedance of the filter is

$$K = Z_{0_1}\sqrt{-\tan\left[\frac{\pi/2}{1+\dfrac{l_2}{2l_1}}\right]\tan\left[\frac{\pi}{2}\left(\frac{1+\dfrac{l_2}{l_1}}{1+\dfrac{l_2}{2l_1}}\right)\right]}$$

$$\doteq \frac{4l_1}{\pi l_2} Z_{0_1} \text{ for narrow bands.} \quad (18)$$

Such a filter would be useful for an interstage coupler to couple together the plate of one screen grid or pentode tube to the grid of another one. One such arrangement is shown in Fig. 3. This method of coupling together two stages of vacuum tubes has an advantage over using a coaxial conductor or a coil and condenser as a tuned circuit on several counts. In the first place the width of the band passed can be accurately controlled and a flatter gain characteristic is obtained. As will be shown later, distributed capacity in the plate and grid of the

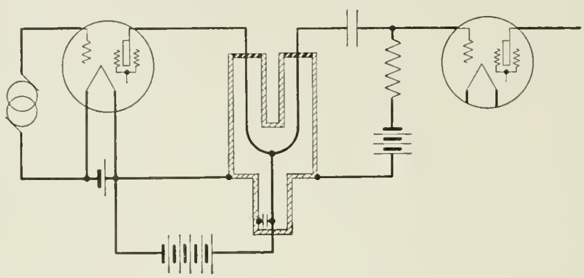
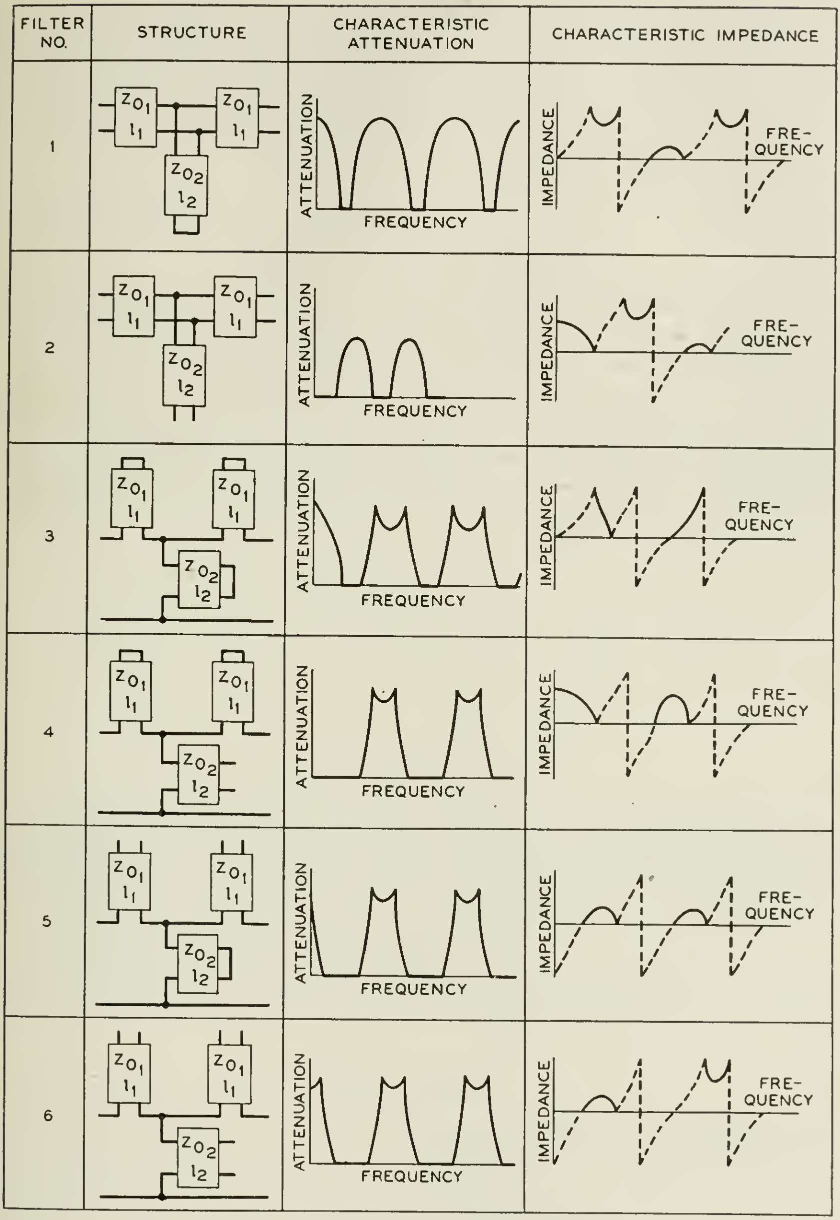


Fig. 3—Coaxial band-pass filter used to couple vacuum tubes.

vacuum tube can be absorbed in the filter by making the line length l_1 shorter. Since only half of the total distributed capacity has to be absorbed on each end of the filter, a higher impedance can be built up in the filter for the same band width, and hence more gain per section can be obtained than with a tuned circuit.

The filter will have other pass bands at $2f_m$, $3f_m$, etc., but these do not usually cause any trouble at the short-wave frequencies because the gain in the vacuum tube is falling off very rapidly and no appreciable signal is passed. If desired, however, another band-pass filter can be added which has the same fundamental bands but different overtone bands, and this will eliminate the effect of the additional pass bands. For very narrow bands, the line length l_2 can be made longer and hence more realizable by making the characteristic impedance Z_{0_2} lower. If anything is to be gained by making the impedance on the plate side different from that on the grid side this can be accomplished by making the filter an impedance transforming device, as discussed in the next section.



NOTE:
DOTTED LINES INDICATE REACTIVE IMPEDANCE. SOLID LINES INDICATE RESISTIVE IMPEDANCE.

Fig. 4—A list of filters constructed from transmission lines.

The filter discussed above shows some of the possibilities and limitations of filters constructed from transmission lines. Many other types are also possible. Figure 4 lists a number of these and the types of filter characteristics they give. The design equations for a number of them are considered in detail in the patent referred to above and hence will not be worked out here.

IV. Impedance Transforming Band-Pass Filters Employing Transmission Lines as Elements

In a good many cases it is desirable to transform from one impedance to another over a wide range of frequencies. Examples of such uses are when antennas are connected to transmission lines, or when transmission lines are to be connected to vacuum tubes, etc. Transforming band-pass filters constructed out of transmission lines which will transform over wide frequency ranges are therefore of practical interest. Previously two types of single-frequency transformers, constructed

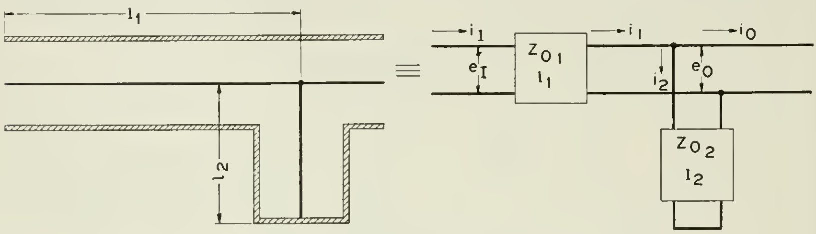


Fig. 5—A wide-band transformer—constructed from coaxial conductors.

from transmission lines, have been suggested [6] but these differ from the types investigated here in that they have a specified ratio of impedances for a single frequency only.

One of the simplest types of band transforming filters is shown by Fig. 5. It consists of a series transmission line of characteristic impedance Z_{0_1} and a shunt line having the characteristic impedance Z_{0_2}. Since the impedance of the short-circuited line is

$$jZ_{0_2} \tan \frac{\omega l_2}{v},$$

the equations for the structure are

$$\begin{aligned} e_1 &= e_I \cos \frac{\omega l_1}{v} - j i_I Z_{0_1} \sin \frac{\omega l_1}{v}; \qquad i_1 = i_2 + i_0; \qquad e_1 = e_0; \\ i_1 &= i_I \cos \frac{\omega l_1}{v} - j \frac{e_I}{Z_{0_1}} \sin \frac{\omega l_1}{v}; \qquad i_2 = - \frac{j e_0}{Z_{0_2} \tan \dfrac{\omega l_2}{v}}. \end{aligned} \tag{19}$$

[6] See reference 1, pages 430 and 431.

Combining these equations we have

$$e_0 = e_I \cos\frac{\omega l_1}{v} - j i_I Z_{0_1} \sin\frac{\omega l_1}{v},$$

$$i_0 = i_I \cos\frac{\omega l_1}{v}\left[1 + \frac{Z_{0_1}\tan\frac{\omega l_1}{v}}{Z_{0_2}\tan\frac{\omega l_2}{v}}\right] - j e_I \frac{\sin\frac{\omega l_1}{v}}{Z_{0_1}}\left[1 - \frac{Z_{0_1}\cot\frac{\omega l_1}{v}}{Z_{0_2}\tan\frac{\omega l_2}{v}}\right]. \qquad (20)$$

In order to interpret this equation in terms of transformer theory, it can be shown that equations (20) are identical to the equations for a perfect transformer and a symmetrical filter. To show this, consider the circuit of Fig. 6, which consists of two half-sections of a symmetrical

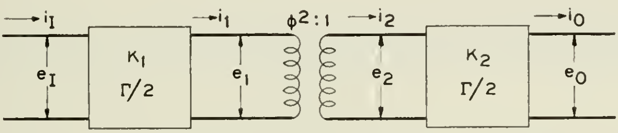


Fig. 6—Transformer and filter.

filter separated by a perfect transformer having an impedance step-down φ^2 to 1. The characteristic impedance of the first filter is φ^2 times that of the second filter. The equations for the first section, the transformer, and the last section are respectively

$$e_1 = e_I \cosh\frac{\Gamma}{2} - i_I K_1 \sinh\frac{\Gamma}{2}; \quad i_1 = i_I \cosh\frac{\Gamma}{2} - \frac{e_I}{K_1}\sinh\frac{\Gamma}{2};$$

$$i_2 = \varphi i_1; \quad e_2 = e_1/\varphi; \qquad (21)$$

$$e_0 = e_2 \cosh\frac{\Gamma}{2} - i_2 K_2 \sinh\frac{\Gamma}{2}; \quad i_0 = i_2 \cosh\frac{\Gamma}{2} - \frac{e_2}{K_2}\sinh\frac{\Gamma}{2}.$$

Combining these equations on the assumption that $K_1/K_2 = \varphi^2$, we have

$$e_0 = \frac{1}{\varphi}\left[e_I \cosh\Gamma - i_I K_1 \sinh\Gamma\right];$$

$$i_0 = \varphi\left[i_I \cosh\Gamma - \frac{e_I \sinh\Gamma}{K_1}\right]. \qquad (22)$$

Similar results can be obtained by using the image impedance parameters of a dissymmetrical filter. For a dissymmetrical filter the most general relationship between the input and output voltages and currents can be written in the form

$$e_2 = e_1A - i_1B; \qquad i_2 = i_1C - e_1D \qquad \text{where} \qquad AC - BD = 1. \tag{23}$$

In terms of these parameters the image transfer constant θ and the image transfer impedances are given by the relationships [7]

$$\cosh\theta = \sqrt{AC}; \qquad K_1 = \sqrt{\frac{BC}{AD}}; \qquad K_2 = \sqrt{\frac{AB}{CD}}. \tag{24}$$

The transformation ratio between the two ends of the network is then

$$\frac{K_1}{K_2} = \frac{C}{A} = \varphi^2 \tag{25}$$

in agreement with the results of equation (22).

Applying these results to equation (20) we find for the most general case

$$\cosh\theta = \cos\frac{\omega l_1}{v}\sqrt{1 + \frac{Z_{0_1}\tan\frac{\omega l_1}{v}}{Z_{0_2}\tan\frac{\omega l_2}{v}}};$$

$$K_1 = Z_{0_1}\sqrt{-\tan\frac{\omega l_1}{v}\left[\frac{\tan\frac{\omega l_1}{v} + \frac{Z_{0_2}}{Z_{0_1}}\tan\frac{\omega l_2}{v}}{1 - \frac{Z_{0_2}}{Z_{0_1}}\tan\frac{\omega l_1}{v}\tan\frac{\omega l_2}{v}}\right]};$$

$$\varphi^2 = 1 + \frac{Z_{0_1}\tan\frac{\omega l_1}{v}}{Z_{0_2}\tan\frac{\omega l_2}{v}}. \tag{26}$$

Two special cases are of interest here: the first when $Z_{0_1} = Z_{0_2}$, and the second when $l_1 = l_2$. The first case corresponds to the transformer disclosed by P. H. Smith,[8] which can be used to transform over a wide

[7] These relations were first proved for wave transmission networks in a paper of the writer's, *B. S. T. J.*, April 1927. See equations 67 and 68, page 291. They are proved by other methods in "Communication Networks," E. A. Guillemin, p. 139.

[8] See reference 1, p. 431.

range of impedance for a single frequency. For this case

$$\begin{aligned}
\cosh\theta &= \cos\frac{\omega l_1}{v}\sqrt{1+\frac{\tan\frac{\omega l_1}{v}}{\tan\frac{\omega l_2}{v}}};\\
K_1 &= Z_{0_1}\sqrt{-\tan\frac{\omega l_1}{v}\tan\frac{\omega(l_1+l_2)}{v}};\\
\varphi^2 &= 1+\frac{\tan\frac{\omega l_1}{v}}{\tan\frac{\omega l_2}{v}}.
\end{aligned}\tag{27}$$

The pass band of the filter lies between the values

$$f_1 = \frac{v}{4(l_1+l_2)} \qquad \text{and} \qquad f_2 = \frac{v}{4l_1} \tag{28}$$

and hence between these frequencies the structure will act as a transformer. The ratio of the transformer, however, varies with frequency, and hence two given impedances can only be matched at one frequency. By adjusting the values of l_1 and l_2 it is possible to transform between any two resistances at a given frequency.

For a number of purposes it is desirable to transform a wide band of frequencies between two constant impedances. This requires a transformer with a constant transformation ratio over the whole band of frequencies. As can be seen from equation (26) this can be accomplished with the structure considered above if we let $l_1 = l_2 = l$. For this case

$$\begin{aligned}
\cosh\theta &= \sqrt{1+\frac{Z_{0_1}}{Z_{0_2}}}\cos\frac{\omega l}{v};\\
K_1 &= Z_{0_1}\varphi\sqrt{\frac{1}{1-\frac{Z_{0_1}}{Z_{0_2}}\cot^2\frac{\omega l}{v}}};\\
\varphi^2 &= 1+\frac{Z_{0_1}}{Z_{0_2}}.
\end{aligned}\tag{29}$$

The mid-band frequency occurs when $\cosh\theta = \cos(\omega l/v) = 0$. Hence l is $\frac{1}{4}$ wave-length at the mid-band frequency. At the mid-band frequency the impedance of the transformer is

$$K_{1_0} = Z_{0_1}\varphi. \tag{30}$$

This is the impedance that the transformer should be connected to since this is the characteristic impedance of the filter. The width of the pass band is determined by

$$1 \geqq \cosh \theta \geqq -1. \tag{31}$$

The cut-off frequencies of the filter transformer are given by the formula

$$\cos \frac{\omega l}{v} = \pm \frac{1}{\varphi} \cdot \tag{32}$$

For relatively narrow bands, the ratio of the band width to the mean frequency is given by the simple formula

$$\frac{f_2 - f_1}{f_m} \doteq \frac{4}{\pi \varphi}, \tag{33}$$

as can readily be shown from equation (32) by using the approximation formula for the cosine in the neighborhood of the angle $\pi/2$.

Hence the structure shown in Fig. 5 is equivalent to a perfect transformer whose ratio is $\varphi^2 = 1 + (Z_{0_1}/Z_{0_2})$ to 1 and a filter whose band width is given by equation (33). Such a filter will have a flat attenuation loss over about 80 per cent of its theoretical band when it is terminated on each side by resistances equal to $Z_{0_1}\varphi$ and Z_{0_1}/φ on its high- and low-impedance sides respectively. Due to the high Q obtainable in the transmission lines, the loss in the band of the filter can be made very low and hence such a transformer will introduce a very small transmission loss. Furthermore, since it is constructed only of transmission lines, it can carry a large amount of power. The complete design equations for the transformer are

$$\varphi^2 = 1 + \frac{Z_{0_1}}{Z_{0_2}}; \qquad Z_{0_1} = \frac{R_I}{\varphi}; \qquad Z_{0_2} = \frac{R_I}{\varphi(\varphi^2 - 1)};$$
$$l = \frac{v}{4f_m}; \qquad R_0 = \frac{R_I}{\varphi^2}, \tag{34}$$

where R_I and R_0 are respectively the input and the output resistances that the transformer works between.

Many other types of transforming networks containing only transmission lines are also possible. Another simple network which is the inverse of the one considered above is shown in Fig. 7. It consists of a length of line l_1 and characteristic impedance Z_{0_1} in series with a balanced open-circuited line of length l_2 and characteristic impedance Z_{0_2}. It is easily shown by employing the equations for a line that when $l_1 = l_2 = l$

$$e_0 = e_I\left(1+\frac{Z_{0_2}}{Z_{0_1}}\right)\cos\frac{\omega l}{v} - j i_I Z_{01}\sin\frac{\omega l}{v}\left[1-\frac{Z_{0_2}}{Z_{0_1}}\cot^2\frac{\omega l}{v}\right];$$

$$i_0 = i_I\cos\frac{\omega l}{v} - j\frac{e_I}{Z_{0_1}}\sin\frac{\omega l}{v}. \tag{35}$$

For this case

$$\varphi^2 = \frac{1}{1+\dfrac{Z_{0_2}}{Z_{0_1}}};\quad \cosh\theta = \frac{\cos\dfrac{\omega l}{v}}{\varphi};\quad K_1 = Z_{0_1}\varphi\sqrt{1-\frac{Z_{0_2}}{Z_{0_1}}\cot^2\frac{\omega l}{v}}. \tag{36}$$

The band width of the filter for narrow bands is given approximately by

$$\frac{f_2 - f_1}{f_m} \doteq \frac{4\varphi}{\pi}.$$

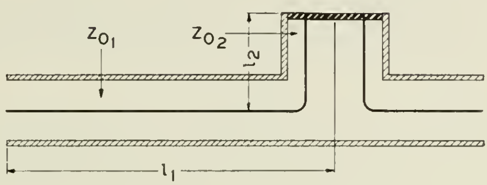


Fig. 7—A wide-band transformer constructed from coaxial conductors.

The design equations for the transformer are

$$\varphi^2 = \frac{1}{1+\dfrac{Z_{0_2}}{Z_{0_1}}};\qquad Z_{0_1} = R_I/\varphi;\qquad Z_{0_2} = R_I(1-\varphi^2)/\varphi^3;$$

$$l = v/4f_m;\qquad R_0 = R_I/\varphi^2. \tag{37}$$

The first transformer discussed is a step-down transformer while the one considered here has a step up from the input of the line to the output. The first filter had a mid-shunt impedance characteristic on each end, i.e., the impedance of the band is infinity at the two edges; whereas the transformer with the series open-circuited line has a mid-series impedance, since the impedance given by the last expression in equation (36) goes to zero at the two cut-off frequencies. The range of transformation is about the same for each and hence one type has no particular advantage over the other.

It is often desirable in filter work to be able to have the impedance of one end of the filter somewhat different from that of the other, i.e., to have the filter act as a transformer of a moderate ratio. An example of this occurs when using a structure composed of short lengths of line to connect two high-frequency pentodes. As shown by Salzberg and Burnside,[9] the output impedance of a high-frequency pentode at 100 megacycles may be in the order of 30,000 ohms due to the high-frequency shunting loss of the tube. On the other hand, due to active grid loss, the impedance looking into the grid of the next tube may be in the order of 20,000 ohms. Hence in order to obtain the most gain from such tubes, the coupling circuit should be able to work from 30,000 ohms when connected with the output to 20,000 ohms for the input of the next tube. If we employ the simple coupling circuit shown in Section III, Fig. 1, this can be made impedance transforming by making the second-series conductor of a different characteristic impedance from the first-series conductor. Such a combination is shown in Fig. 8. The equations of the combination are easily solved

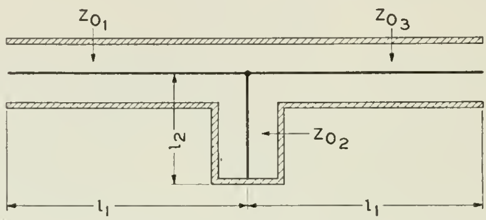


Fig. 8—A narrow-band transformer for coupling vacuum tubes.

with the result that

$$
\begin{aligned}
i_0 &= i_I\left[\cos^2\frac{\omega l_1}{v} - \frac{Z_{01}}{Z_{03}}\sin^2\frac{\omega l_1}{v} + \frac{Z_{01}}{Z_{02}}\,\frac{\sin\dfrac{\omega l_1}{v}\cos\dfrac{\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}}\right] \\
&\qquad - je_I\left[\frac{Z_{01}+Z_{03}}{Z_{01}Z_{03}}\sin\frac{\omega l_1}{v}\cos\frac{\omega l_1}{v} - \frac{\cos^2\dfrac{\omega l_1}{v}}{Z_{02}\tan\dfrac{\omega l_2}{v}}\right]; \\
e_0 &= e_I\left[\cos^2\frac{\omega l_1}{v} - \frac{Z_{03}}{Z_{01}}\sin^2\frac{\omega l_1}{v} + \frac{Z_{03}}{Z_{02}}\,\frac{\sin\dfrac{\omega l_1}{v}\cos\dfrac{\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}}\right] \\
&\qquad - ji_I\left[(Z_{01}+Z_{03})\sin\frac{\omega l_1}{v}\cos\frac{\omega l_1}{v} + \frac{Z_{01}Z_{03}}{Z_{02}}\,\frac{\sin^2\dfrac{\omega l_1}{v}}{\tan\dfrac{\omega l_2}{v}}\right].
\end{aligned}
\tag{38}
$$

[9] "Recent Developments in Miniature Tubes," *Proc. I. R. E.*, Vol. 23, No. 10, p. 142, Oct. 1935.

For this case the image transfer constant and the impedance ratio of the transforming filter are given by the equations

$$\cosh \theta = \sqrt{\cos^2 \frac{2\omega l_1}{v} + \left(\frac{Z_{0_1} + Z_{0_3}}{2Z_{0_2}}\right) \frac{\sin \frac{2\omega l_1}{v} \cos \frac{2\omega l_1}{v}}{\tan \frac{\omega l_2}{v}} + \sin^2 \frac{2\omega l_1}{v} \left[\frac{Z_{0_1} Z_{0_3}}{4Z_{0_2}{}^2 \tan^2 \frac{\omega l_2}{v}} - \frac{(Z_{0_1} - Z_{0_3})^2}{4Z_{0_1} Z_{0_3}}\right]};$$

$$\varphi^2 = \left[\frac{\cos^2 \frac{\omega l_1}{v} - \frac{Z_{0_1}}{Z_{0_3}} \sin^2 \frac{\omega l_1}{v} + \frac{Z_{0_1}}{Z_{0_2}} \frac{\sin \frac{\omega l_1}{v} \cos \frac{\omega l_1}{v}}{\tan \frac{\omega l_2}{v}}}{\cos^2 \frac{\omega l_1}{v} - \frac{Z_{0_3}}{Z_{0_1}} \sin^2 \frac{\omega l_1}{v} + \frac{Z_{0_3}}{Z_{0_2}} \frac{\sin \frac{\omega l_1}{v} \cos \frac{\omega l_1}{v}}{\tan \frac{\omega l_2}{v}}}\right]. \tag{39}$$

When we solve the expression for $\cosh \theta$ for the cut-off frequencies, we find that one of them is given by the expression

$$f_2 = \frac{v}{4l_1} \tag{40}$$

and the other by

$$\tan \frac{2\omega l_1}{v} = \frac{-2\left(\frac{Z_{0_1} + Z_{0_3}}{Z_{0_1} Z_{0_3}}\right) Z_{0_2} \tan \frac{\omega l_2}{v}}{1 - \frac{(Z_{0_1} + Z_{0_3})^2}{Z_{0_1}{}^2 Z_{0_3}{}^2} Z_{0_2}{}^2 \tan^2 \frac{\omega l_2}{v}}. \tag{41}$$

If we consider the special case $Z_{0_2} = Z_{0_1} Z_{0_3}/(Z_{0_1} + Z_{0_3})$ equation (41) reduces to the simple form

$$\tan \frac{2\omega l_1}{v} = -\tan \frac{2\omega l_2}{v} \tag{42}$$

or for narrow bands

$$\frac{2\omega_1 l_1}{v} = \pi - \frac{2\omega_1 l_2}{v} \quad \text{and} \quad f_1 = \frac{v}{4(l_1 + l_2)} \tag{43}$$

in agreement with equation (17).

At the two cut-off frequencies, it can be shown that

$$\varphi^2 = \frac{Z_{0_1}{}^2}{Z_{0_3}{}^2} \tag{44}$$

and throughout the band the value of φ does not differ much from this value for narrow-band filters.

Hence the structure of Fig. 8 acts as a narrow-band coupling unit which introduces a transformation from input to output. For narrow bands it is easily shown that the image impedance at the middle of the pass band is given by the expression

$$K_1 = \frac{4}{\pi}\frac{(f_m)}{(f_2 - f_1)}\sqrt{Z_{0_1}Z_{0_3}}\frac{Z_{0_1}}{Z_{0_3}}. \tag{45}$$

The design equations for this transforming filter are

$$l_1 = \frac{v}{4f_2}; \quad l_2 = \frac{v}{4}\left[\frac{1}{f_1} - \frac{1}{f_2}\right]; \quad Z_{0_1} = \frac{\pi}{4}\frac{R_I(f_2 - f_1)}{\sqrt{\varphi}f_m};$$

$$Z_{0_3} = \frac{\pi}{4}\frac{R_I(f_2 - f_1)}{f_m\varphi^{\frac{3}{2}}}; \quad Z_{0_2} = \frac{\pi R_I(f_2 - f_1)}{4f_m\varphi^{\frac{1}{2}}(1 + \varphi)}; \quad \frac{R_I}{R_0} = \varphi^2. \tag{46}$$

V. Filters and Transformers Employing Transmission Lines and Condensers

Condensers can be constructed for high radio frequencies which have little dissipation and hence they can be combined with short sections of transmission lines to produce filters and transformers. Combinations of lines with condensers have the advantages that much more isolated bands can be obtained, in general narrower pass bands can be obtained, and at the lower frequencies shorter sections of lines can be employed if they are resonated by capacities. Also, when such structures are used as interstage coupling units working between vacuum tubes, they usually will have to incorporate the grid to filament and plate to filament capacities as part of the coupling circuit. Hence it is desirable to consider combinations of transmission lines and condensers as filters and transformers.

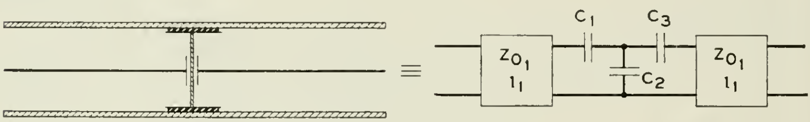


Fig. 9—A transformer or filter using condensers and coaxial conductors.

One of the simplest and most useful types of band-pass filter using transmission lines and condensers is shown in Fig. 9. This structure has been used as a wide-band transformer and as a very narrow-band filter, and experimental curves are given in the next section. The equations connecting the output voltage and current with the input voltage and current can easily be calculated by the methods given above and are

$$\begin{aligned} e_0 = e_I \Bigg[& \frac{C_2 + C_3}{C_3} \cos^2 \frac{\omega l}{v} - \frac{Z_{0_2}}{Z_{0_1}} \left(\frac{C_1 + C_2}{C_2} \right) \sin^2 \frac{\omega l}{v} \\ & + \frac{\sin \frac{2\omega l}{v}}{2} \left(\frac{C_1 + C_2 + C_3}{\omega Z_{0_1} C_1 C_3} - \omega C_2 Z_{0_2} \right) \Bigg] \\ - j i_I Z_{0_1} \Bigg[& \frac{\sin \frac{2\omega l}{v}}{2} \left[\frac{C_2 + C_3}{C_3} + \frac{Z_{0_2}}{Z_{0_1}} \left(\frac{C_1 + C_2}{C_1} \right) \right] \\ & - \left[\frac{C_1 + C_2 + C_3}{\omega Z_{0_1} C_1 C_3} \cos^2 \frac{\omega l}{v} + Z_{0_2} \omega C_2 \sin^2 \frac{\omega l}{v} \right] \Bigg]; \end{aligned} \tag{47}$$

$$\begin{aligned} i_0 = i_I \Bigg[& \frac{C_1 + C_2}{C_1} \cos^2 \frac{\omega l}{v} - \frac{Z_{0_1}}{Z_{0_2}} \left(\frac{C_2 + C_3}{C_3} \right) \sin^2 \frac{\omega l}{v} \\ & + \frac{\sin \frac{2\omega l}{v}}{2} \left[\frac{C_1 + C_2 + C_3}{\omega Z_{0_2} C_1 C_3} - Z_{0_1} \omega C_2 \right] \Bigg] \\ - \frac{j e_I}{Z_{0_1}} \Bigg[& \frac{\sin \frac{2\omega l}{v}}{2} \left[\frac{C_1 + C_2}{C_1} + \frac{Z_{0_1}}{Z_{0_2}} \left(\frac{C_2 + C_3}{C_3} \right) \right] \\ & + \sin^2 \frac{\omega l}{v} \left[\frac{C_1 + C_2 + C_3}{\omega Z_{0_2} C_1 C_3} \right] + \omega C_2 Z_{0_1} \cos^2 \frac{\omega l}{v} \Bigg]. \end{aligned}$$

In order that the structure shall transform uniformly over a band of frequencies we must have

$$\begin{aligned} \varphi^2 = \left(\frac{C_1 + C_2}{C_1} \right) \left(\frac{C_3}{C_2 + C_3} \right) = \frac{Z_{0_1}}{Z_{0_2}} \\ = \text{impedance transformation ratio.} \end{aligned} \tag{48}$$

With this substitution, equations (47) simplify to

$$
\begin{aligned}
e_0 = e_I &\left[\frac{C_2 + C_3}{C_3} \cos \frac{2\omega l}{v} + \frac{\sin \dfrac{2\omega l}{v}}{2} \left[\frac{C_1 + C_2 + C_3}{\omega Z_{01} C_1 C_3} - \omega C_2 Z_{02} \right] \right] \\
&- j i_I Z_{01} \left[\left(\frac{C_2 + C_3}{C_3} \right) \sin \frac{2\omega l}{v} \right. \\
&\qquad \left. - \left[\frac{C_1 + C_2 + C_3}{\omega Z_{01} C_1 C_3} \cos^2 \frac{\omega l}{v} + Z_{02} \omega C_2 \sin^2 \frac{\omega l}{v} \right] \right]; \\
i_0 = \varphi^2 &\left[i_I \left[\frac{C_2 + C_3}{C_3} \cos \frac{2\omega l}{v} \right. \right. \\
&\qquad \left. + \frac{\sin \dfrac{2\omega l}{v}}{2} \left(\frac{C_1 + C_2 + C_3}{\omega Z_{01} C_1 C_3} - Z_{02} \omega C_2 \right) \right] \\
&- \frac{j e_I}{Z_{01}} \left[\left(\frac{C_2 + C_3}{C_3} \right) \sin \frac{2\omega l}{v} + \frac{C_1 + C_2 + C_3}{\omega Z_{01} C_1 C_3} \sin^2 \frac{\omega l}{v} \right. \\
&\qquad \left. \left. + \omega C_2 Z_{02} \cos^2 \frac{\omega l}{v} \right] \right].
\end{aligned}
\tag{49}
$$

Comparing these equations with equations (24) we have for the image parameters

$$
\begin{aligned}
\cosh \theta &= \varphi \left[\frac{C_2 + C_3}{C_3} \cos \frac{2\omega l}{v} \right. \\
&\qquad \left. + \frac{\sin \dfrac{2\omega l}{v}}{2} \left[\frac{C_1 + C_2 + C_3}{\omega Z_{01} C_1 C_3} - \omega C_2 Z_{02} \right] \right]; \\
K_1 &= Z_{01} \sqrt{\frac{\dfrac{C_2 + C_3}{C_3} \sin \dfrac{2\omega l}{v} - \left[\dfrac{C_1 + C_2 + C_3}{\omega Z_{01} C_1 C_3} \cos^2 \dfrac{\omega l}{v} + Z_{02} \omega C_2 \sin^2 \dfrac{\omega l}{v} \right]}{\dfrac{C_2 + C_3}{C_3} \sin \dfrac{2\omega l}{v} + \dfrac{C_1 + C_2 + C_3}{\omega Z_{01} C_1 C_3} \sin^2 \dfrac{\omega l}{v} + Z_{02} \omega C_2 \cos^2 \dfrac{\omega l}{v}}}; \\
K_2 &= \frac{K_1}{\varphi^2}.
\end{aligned}
\tag{50}
$$

These equations give the image parameters for a general transforming band-pass filter. The two uses to which such a structure will ordinarily be put are either to obtain a transformer with as wide a pass band as possible for a given impedance transformation or else to obtain a filter without transformation ratio. For the transformer case it can be shown that the widest pass-band occurs when $C_3 \rightarrow \infty$, or in other words the condenser C_3 is short-circuited. In order to obtain a simple design, it is assumed that each conductor is an eighth of a wave-length at the mid-band frequency or that

$$\frac{\omega_m l}{v} = \frac{\pi}{4}. \tag{51}$$

The mid-band frequency occurs when $\cosh\theta = 0$. Upon substituting the relation $Z_{0_1} = l/vC_0$ where C_0 is the total distributed capacity of the input line of the transformer, $\cosh\theta$ vanishes when

$$\frac{C_0}{C_1} = \frac{\pi^2}{16}\frac{C_2}{\varphi^2 C_0}. \tag{52}$$

Solving for the frequencies for which $\cosh\theta = \pm 1$, it is easily shown that the ratio of the band width to the mean frequency is given by the expression

$$\frac{f_2 - f_1}{f_m} \doteq \frac{\dfrac{4}{\pi\varphi}}{1 + \dfrac{C_2}{2\varphi^2 C_0}}. \tag{53}$$

The image impedance K_1 at the mid-frequency of the band is from equation (50)

$$K_{1_0} = Z_{0_1}\sqrt{\frac{1 - \dfrac{\pi}{4}\dfrac{C_2}{\varphi^2 C_0}}{1 + \dfrac{\pi}{4}\dfrac{C_2}{\varphi^2 C_0}}}. \tag{54}$$

From the above equations and noting that $\varphi^2 = 1 + C_2/C_1$, the design equations of the transformer become

$$Z_{0_1} = R_1\sqrt{\frac{\varphi + \sqrt{\varphi^2 - 1}}{\varphi - \sqrt{\varphi^2 - 1}}}; \qquad Z_{0_2} = \frac{Z_{0_1}}{\varphi^2}; \qquad l = \frac{v}{8f_m};$$

$$C_0 = \frac{33.3l}{Z_{0_1}} \text{ in } \mu\mu\text{f}; \qquad C_1 = \frac{4C_0}{\pi}\sqrt{\frac{\varphi^2}{\varphi^2 - 1}}; \tag{55}$$

$$C_2 = \frac{4C_0}{\pi}\sqrt{(\varphi^2 - 1)\varphi^2},$$

where R_1 is the input impedance from which the transformer must work.

When the structure of Fig. 9 is used as a filter without transformation, we have $C_1 = C_3$; $Z_{0_1} = Z_{0_2} = Z_0$. For this case the image parameters become

$$\cosh \theta = \frac{C_1 + C_2}{C_1} \cos \frac{2\omega l}{v} + \frac{\sin \frac{2\omega l}{v}}{2} \left[\frac{2C_1 + C_2}{\omega Z_0 C_1^2} - \omega C_2 Z_0 \right];$$

$$K_1 = K_2 = K = Z_0 \sqrt{\frac{1 - \left[\dfrac{(2C_1 + C_2) \cot \frac{\omega l_1}{v}}{2\omega Z_0 C_1 (C_1 + C_2)} + \dfrac{Z_0 \omega C_1 C_2}{2(C_1 + C_2)} \tan \frac{\omega l_1}{v} \right]}{1 + \dfrac{(2C_1 + C_2) \tan \frac{\omega l_1}{v}}{2\omega Z_0 C_1 (C_1 + C_2)} + \dfrac{Z_0 \omega C_1 C_2}{2(C_1 + C_2)} \cot \frac{\omega l_1}{v}}}. \tag{56}$$

For narrow-band filters it is easily shown that

$$\frac{f_2 - f_1}{f_m} = \Delta = \frac{\frac{4}{\pi}}{1 + \frac{C_2}{C_1} + \frac{C_2}{2C_0}}; \qquad \frac{(2C_1 + C_2)C_0}{C_1^2} = \frac{\pi^2}{16} \frac{C_2}{C_0}. \tag{57}$$

At the mid-band of the filter, since the constants were worked out on the assumption that each conductor was an eighth of a wave-length at the mid-band frequency, the mid-band filter impedance can be obtained from the last part of (56) by setting $\omega l / v = \pi / 4$, giving

$$K_0 = Z_0 \sqrt{\frac{1 + \frac{C_2}{C_1} - \frac{\pi}{4} \frac{C_2}{C_0}}{1 + \frac{C_2}{C_1} + \frac{\pi}{4} \frac{C_2}{C_0}}}. \tag{58}$$

Hence solving for the constants of the filter on the assumption that Δ is a small quantity, we find

$$C_1 = \frac{4C_0}{\pi}; \qquad C_2 = \frac{16C_0}{\pi(2 + \pi)\Delta}; \qquad Z_0 = \frac{8R}{(\pi + 2)\Delta};$$

$$l = \frac{v}{8f_m}; \qquad C_0 = \frac{33.3l}{Z_0} \text{ in } \mu\mu\text{f}, \tag{59}$$

where R is a resistance equal to K at the mean-frequency of the filter. For narrow bands this gives a very large value for C_2, the shunt capacity. A more practical arrangement is to replace the two series condensers C_1 and the shunt condenser C_2 by a π network consisting of two shunt condensers C_A separated by a series condenser C_B. These have the values

$$C_A = \frac{C_1 C_2}{2C_1 + C_2}; \qquad C_B = \frac{C_1^2}{2C_1 + C_2}. \tag{60}$$

With this arrangement we find that for narrow bands $C_A \doteq C_1$ and C_B is a very small capacity. This can readily be obtained physically by inserting a partition with a small hole in it at the middle of the section. Then C_A will be the capacity of the inside conductors to the partition, and C_B will be the capacity of one inside conductor to the other looking through the small hole. By adjusting the size of this hole, this capacity can be made as small as desired.

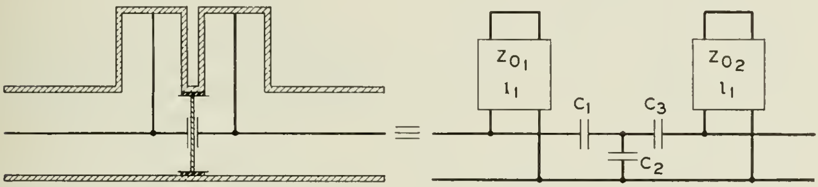


Fig. 10—A shunt terminated transformer.

The transformer discussed above is suitable for transforming from line impedances down to very low impedances, but cannot be used to transform from line impedances up to very high impedances such as the impedance of a vacuum tube. This generally requires a shunt type of termination rather than the series type discussed above. One such transformer is shown in Fig. 10. It consists of two shunt lines connected together by a T or π network of capacities. The constants of such a transformer are readily calculated, and for the condition of maximum transformation for a given band width—which occurs when $C_3 \to \infty$, and for eighth wave-length conductors on each end—these have been found to be

$$\varphi^2 = 1 + \frac{C_2}{C_1} = \frac{Z_{01}}{Z_{02}}; \qquad Z_{01} = \frac{K_{10}}{\varphi}; \qquad Z_{02} = \frac{K_{10}}{\varphi^3};$$
$$C_1 = \frac{4C_0}{\pi}; \qquad C_2 = \frac{4}{\pi} C_0(\varphi - 1); \qquad l = \frac{v}{8f_m}. \tag{61}$$

The theoretical band width for this type transformer is given approximately by the expression

$$\frac{f_2 - f_1}{f_m} \doteq \frac{4}{\varphi(\pi + 2)}. \tag{62}$$

Such a transformer is also suitable for connecting together vacuum tubes of high impedance.

Another type of transformer of some interest is one which will transform from very high impedances to very low impedances. Such a transformer is shown in Fig. 11. It has a shunt conductor on the high-impedance end and a series conductor on the low-impedance end. Such a transformer does not have a constant transformation ratio over the whole band, but for about 80 per cent of the theoretical band width

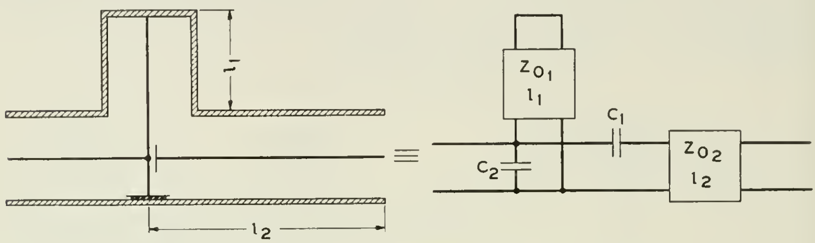


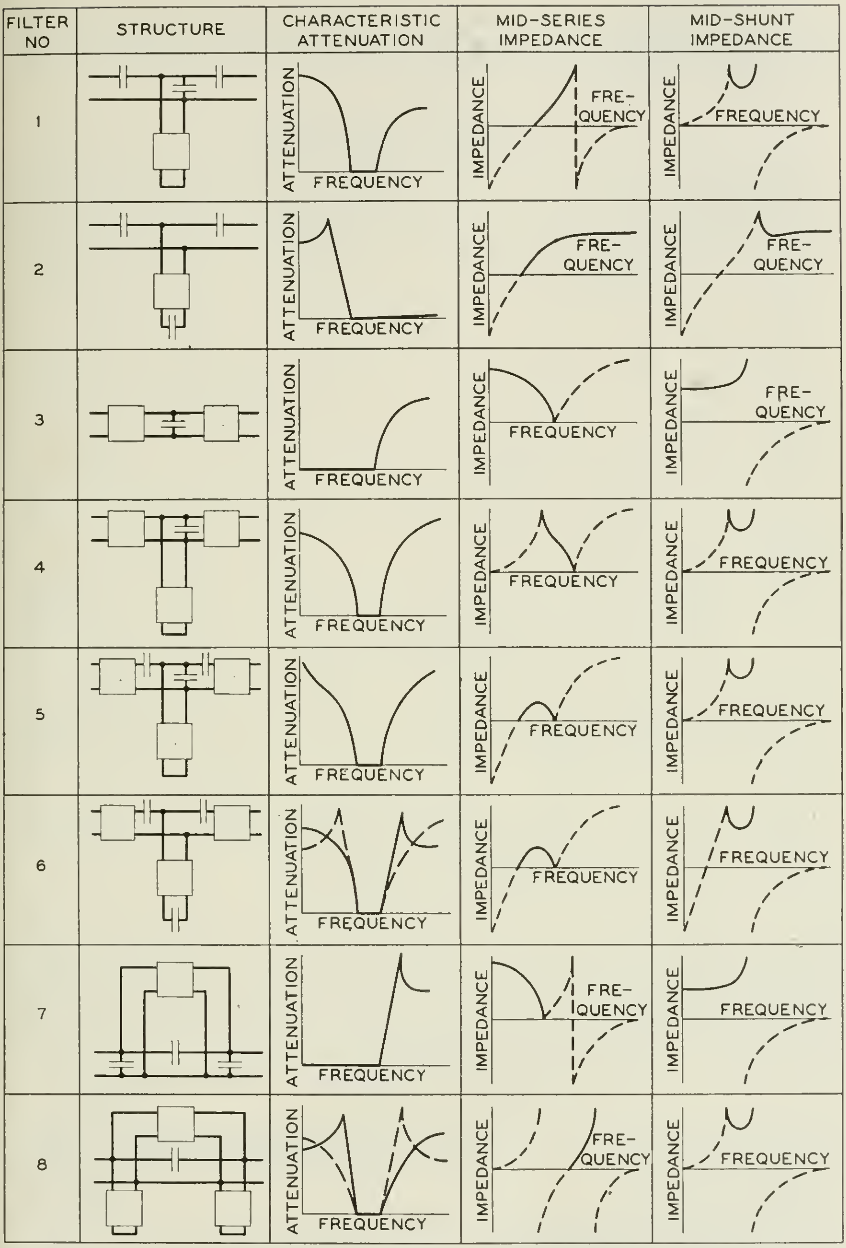
Fig. 11—A series shunt type transformer suitable for wide bands.

the transformation ratio is approximately constant. The design equations for such a transformer are

$$\begin{aligned} &\frac{K_{10}}{K_{20}} = \varphi^2; \quad f_m = \sqrt{f_1 f_2}; \quad Z_{01} = \frac{K_{10}(f_2 - f_1)}{f_m}; \quad l_1 = \frac{v}{8f_m}; \\ &Z_{02} = \frac{K_{20} f_m}{(f_2 - f_1)\sqrt{1 - 1/\varphi}}; \quad l_2 = \frac{v\sqrt{1 - 1/\varphi}}{2\pi f_m}; \\ &C_1 = \frac{(f_2 - f_1)\varphi}{2\pi f_m^2 K_{10}}; \\ &C_2 = \frac{1}{2\pi K_{10}(f_2 - f_1)}\left[1 - \frac{(f_2 - f_1)^2(\varphi - 1)}{f_m^2}\right]. \end{aligned} \tag{63}$$

The transformer is especially useful since it will give the widest transmission band for a given transformation ratio of any of the transformers discussed.

Many other types of filters, transforming and nontransforming, are also possible using transmission lines and condensers. A partial list of such filters is shown in Fig. 12 together with their attenuation and



NOTE DOTTED LINES INDICATE
REACTIVE IMPEDANCE. SOLID LINES
INDICATE RESISTIVE IMPEDANCE

Fig. 12—A list of filter structures employing transmission lines and condensers.

iterative impedance characteristics. All of the band-pass filters can be made impedance-transforming by varying the ratios of the impedance elements for the two ends.

VI. Experimental Results

Several filters and transformers of the types discussed above have been tested experimentally and have been found to give operating characteristics in accordance with the calculated results. Figure 13

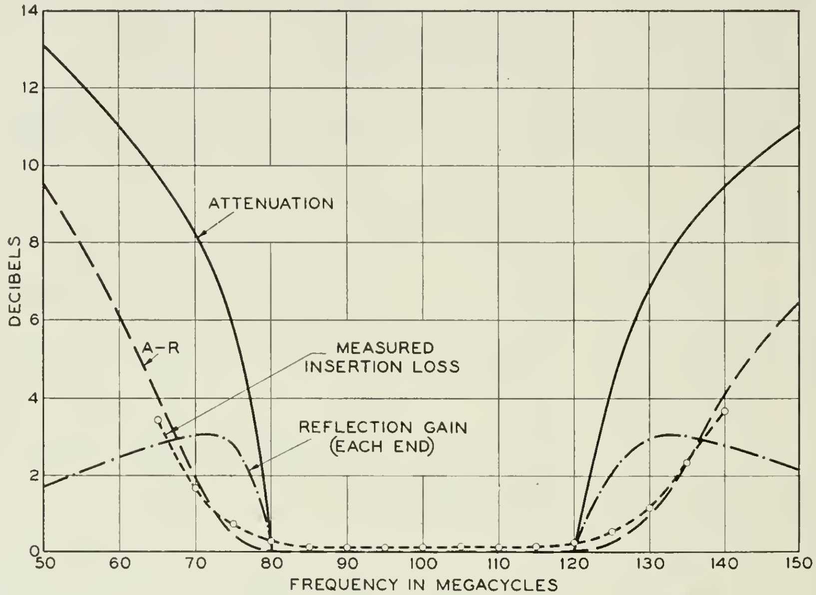


Fig. 13—Measured insertion loss of a wide-band transformer.

shows the measured insertion loss of a wide-band low-impedance transformer which transforms from 70 ohms down to an impedance of 17.5 ohms. The useful transformation band is from 80 megacycles to 120 megacycles. The measuring circuit is shown in Fig. 14. It consists

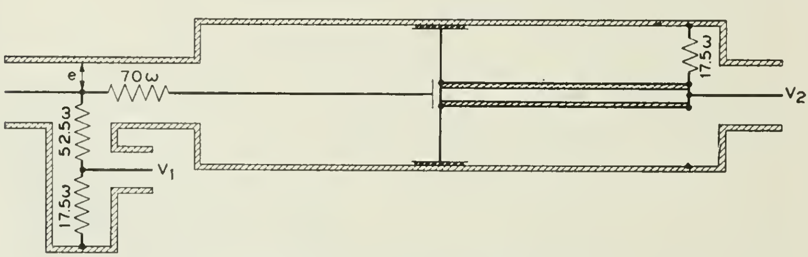


Fig. 14—Measuring circuit for a transformer.

of a source of high-frequency voltage impressed on a divided circuit. In one branch is a series resistance of 70 ohms connected to the 70-ohm side of the transformer. The output of the transformer is connected

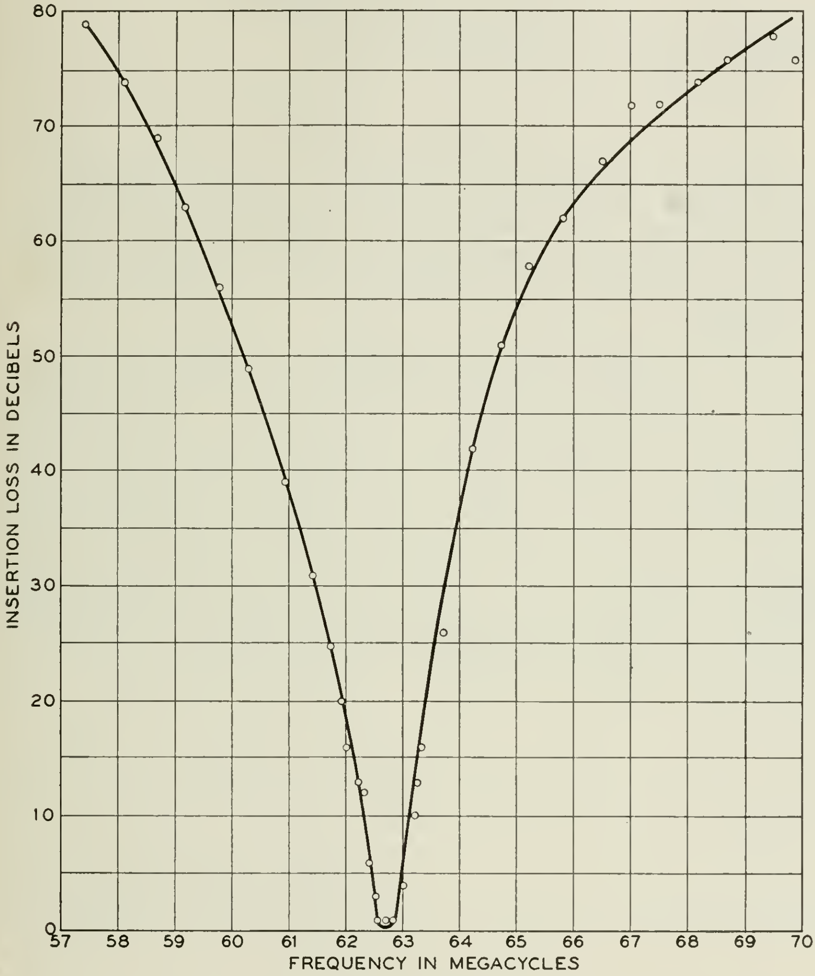


Fig. 15—Measured insertion loss of filter used on Green Harbor–Provincetown radio link.

to a 17.5 ohm resistance and a high-impedance voltmeter is connected across it. The other branch contains a 52.5-ohm and a 17.5-ohm resistance in series, with a high-impedance voltmeter shunted across the 17.5 ohms. If the transformer were a perfect transformer, the current from the output of the transformer would be

$$i_0 = \frac{e}{140} \times \sqrt{\frac{70}{17.5}} = \frac{e}{70}, \tag{64}$$

where e is the voltage applied to ground at the common point. The voltage across the output should be

$$e_0 = \frac{e}{70} \times 17.5 = \frac{e}{4}. \tag{65}$$

But this is just the voltage that should occur across the voltmeter V_1. Hence the difference in reading between V_2 and V_1 will be a measure of the loss introduced by the transformer. From Fig. 13 we see that this is in the order of 0.1 db, which represents a small loss for a transformer.

Several of the narrow-band filters of the type shown in section V, Fig. 9, have also been constructed and tested. One of these has been used on an experimental radio system at Green Harbor, Massachusetts, since 1935, for the purpose of connecting a transmitter and receiver on the same antenna. This filter has been constructed and tested by Messrs. F. A. Polkinghorn and N. J. Pierce using the design data developed here. The filter used consisted of three sections of the type shown in Fig. 9 connected in tandem. The resulting insertion loss of the filter and associated transformers is shown in Fig. 15. The loss at mid-band is in the order of 1 db and an insertion loss of over 50 db is obtained 2 megacycles on either side of the center of the pass band.

A Ladder Network Theorem

By JOHN RIORDAN

The theorem of this paper gives four-terminal representation of ladder networks satisfying a prescribed condition on the side impedances, in terms of the three parameters specifying the network connected as a transducer, the driving-point impedance between short-circuited transducer terminal pairs, and an impedance ratio involving the side impedances only. This mode of representation has a special advantage in applications to electric railway networks in that the transducer parameters which alone involve the ladder shunt impedances (under the stated conditions) may be calculated in a relatively simple fashion, and extensive networks reduced to manageable form. The theorem is stated and proved, and its applications are sketched in some detail.

THE theorem of this paper gives a four-terminal representation of ladder networks satisfying a certain condition with respect to the side impedances. Ladder networks appearing in transmission and filter theory generally are connected as transducers, that is, such that the entry and exit terminals on the ladder sides are associated in pairs; the networks are two-terminal pairs. As is well known, passive transducers may be completely specified by three parameters (as is the case for three-terminal networks, with which transducers are similar in some, though not all, respects), the choice of which has been the occasion for much study and ingenuity.[1] The present theorem does not assume transducer connection and is thus quite distinct from earlier work; indeed it arose outside the communication field in the problem of the calculation of short-circuit currents and network current distribution of electric railway networks, where at present it seems to have chief application.

This paper gives a statement of the theorem, an indication of its applications, and finally its proof.

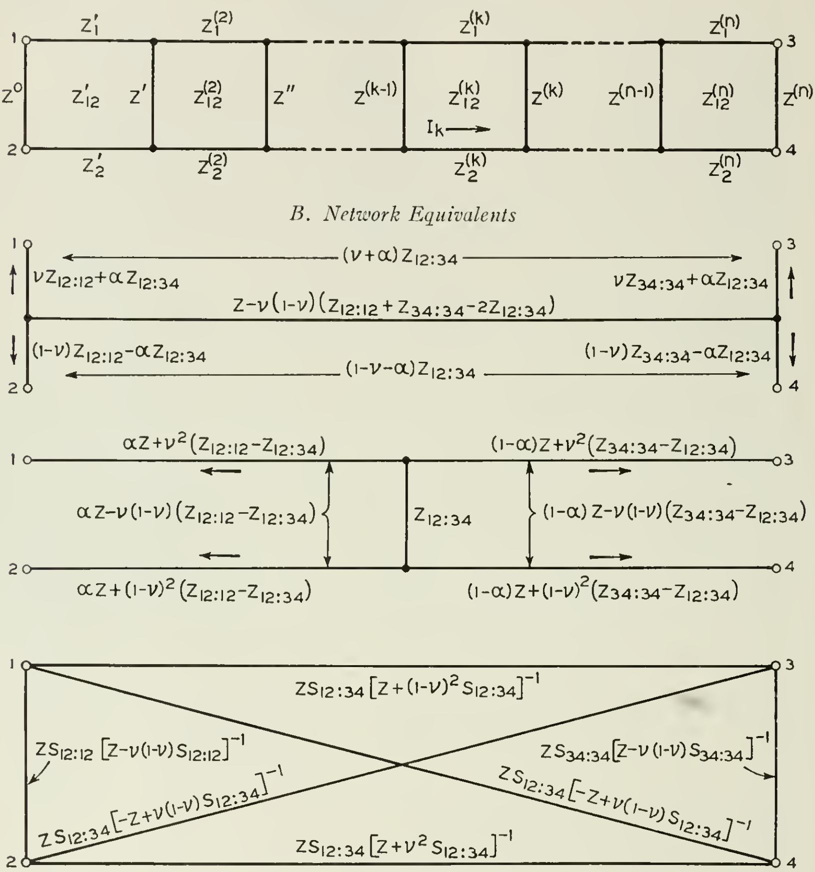
The Theorem

A ladder network, composed of any number of arbitrary shunt impedances forming sections whose side impedances $Z_1^{(k)}$, $Z_2^{(k)}$ *and* $Z_{12}^{(k)}$, $k = 1$,

[1] Five types of equivalent networks by which a transducer may be replaced, including T, π, transformer and artificial line networks, and their interrelations are given on Table I of "Cisoidal Oscillations" by G. A. Campbell, *Trans. A. I. E. E.*, **30**, pp. 873–909 (1911). The most significant addition to the table would appear to be the image impedance representation due to O. J. Zobel.

2 ⋯, are such that $[Z_1^{(k)} - Z_{12}^{(k)}][Z_2^{(k)} - Z_{12}^{(k)}]^{-1}$ *is a constant, may be completely specified by its three transducer parameters (with transducer terminal pairs each made up of adjacent terminals on opposite ladder sides*[2]*), the driving-point impedance between short-circuited transducer*

A. Network Diagram



Notation

$\nu = \nu_k = [Z_1^{(k)} - Z_{12}^{(k)}][Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}]^{-1}$

$$Z = \sum_1^n \frac{Z_1^{(k)}Z_2^{(k)} - (Z_{12}^{(k)})^2}{Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}}$$

α = Arbitrary Constant

$S_{12:12} = (Z_{12:12}Z_{34:34} - Z^2_{12:34})/Z_{34:34}$

$S_{34:34} = (Z_{12:12}Z_{34:34} - Z^2_{12:34})/Z_{12:12}$

$S_{12:34} = (Z_{12:12}Z_{34:34} - Z^2_{12:34})/Z_{12:34}$

Fig. 1—A sample ladder network illustrating notation; and some network equivalents.

[2] The theorem also holds when the terminals of each pair are non-adjacent, that is, for terminal pairs 1, 4 and 3, 2 of Fig. 1*A*; this result is of no importance in the railway applications.

terminal pairs, and the constant

$$[Z_1^{(k)} - Z_{12}^{(k)}][Z_2^{(k)} - Z_{12}^{(k)}]^{-1}.$$

The current in any branch of the network for any condition of energization of the network terminals is a linear function of the currents in the same branch for energization at sending and receiving transducer terminals.

It should be observed that the result stated is independent of the number or values of the shunt impedances (except as they are included in the transducer parameters); hence in the diagram on Fig. 1*A* illustrating the ladder network in question, any of the shunt impedances may be allowed to vanish or become infinite, and their number $n + 1$ may be increased or decreased at pleasure provided that one shunt remains (this excludes the trivial case in which, the sides being completely insulated from each other, the network degenerates to a pair of single impedances).

When the impedances of the sides are linearly extended impedances, as is the case in electric railway applications, the section impedances may be written:

$$Z_1^{(k)} = s_k z_1,$$
$$Z_2^{(k)} = s_k z_2,$$
$$Z_{12}^{(k)} = s_k z_{12},$$

where z_1, z_2 and z_{12} are self and mutual impedances of the sides per unit length. The condition, $[Z_1^{(k)} - Z_{12}^{(k)}][Z_2^{(k)} - Z_{12}^{(k)}]^{-1} = \text{const.}$, is replaced by the condition that the shunt impedances connect corresponding points on the sides.

Since a four-terminal network requires six independent quantities for its specification, the conditions (a) that the network be of ladder type and (b) that the given section impedance ratio be constant may be regarded, at least intuitively, as replacing two (or more) of the measurable impedances at the terminals.[3]

With the sending and receiving transducer terminal pairs short-circuited, and with the side impedances satisfying the given condition, no current flows in any of the shunt impedances and the driving-point impedance required for the theorem is

$$Z = \sum_{k=1}^{n} \frac{Z_1^{(k)} Z_2^{(k)} - Z_{12}^{(k)2}}{Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}} = \frac{\sum Z_1^{(k)} \sum Z_2^{(k)} - (\sum Z_{12}^{(k)})^2}{\sum Z_1^{(k)} + \sum Z_2^{(k)} - 2\sum Z_{12}^{(k)}}, \qquad (1)$$

[3] It is interesting to observe that, if short circuits between terminals are permitted, there are 64 measurable impedances for a four-terminal network. The network may be specified by any six of these which are independent; hence the number of ways of specifying the network is something less than the number of combinations of 64 things taken 6 at a time, which equals 74,974,368. The number of non-independent sets which make up this large total appears at the moment to be the smaller part, and possibly a very small part indeed. These remarks are inspired by Mr. R. M. Foster.

where the summations in the last expression extend over all the sections; this impedance then is simply the parallel impedance of the sides taken in their entirety.

The current in branch k of line 2, designated by I_k on Fig. 1A, for any condition of energization is expressed in terms of the currents in the same branch and in the same direction for unit current supplied between terminals 1 and 2, and 3 and 4 [terminals 3, 4 (1, 2) open, respectively], designated by $i_{k:12}$ and $i_{k:34}$, respectively, by the following equation:

$$I_k = \nu(I_3 + I_4) - i_{k:12}[\nu I_1 - (1 - \nu)I_2] - i_{k:34}[\nu I_3 - (1 - \nu)I_4], \quad (2)$$

where I_1, I_2, I_3 and I_4 are currents flowing out of the network from the respective terminals, and ν is the current in side 2 for unit current between short-circuited transducer terminals, as given on Fig. 1B. Thus I_k is a linear function of currents $i_{k:12}$ and $i_{k:34}$, as stated in the second half of the theorem.

Three types of networks completely equivalent to any ladder network satisfying the condition of the theorem are shown on Fig. 1B. The transducer impedances employed in the representation by these networks are the driving-point impedances between transducer terminals 1 and 2, and 3 and 4 [terminals 3, 4 (1, 2) open, respectively] and the corresponding transfer impedance between the ends of the transducer. These impedances are designated $Z_{12:12}$, $Z_{34:34}$ and $Z_{12:34}$, following a notation for Neumann integrals used by G. A. Campbell.[4] For present purposes the notation has the advantage of putting into evidence the terminals between which current is supplied and the terminals between which voltage is measured; thus $Z_{12:12}$ may be read as the voltage drop from 1 to 2 for unit current from 1 to 2 (terminals 3, 4 open), $Z_{12:34}$ the voltage drop from 3 to 4 for unit current from 1 to 2 under the same conditions.[5] By the reciprocity theorem $Z_{12:34} = Z_{34:12}$.

[4] "Mutual Impedance of Grounded Circuits," *Bell System Technical Journal*, 2, 1–30 (Oct. 1923).

[5] Further, the subscripts may be handled algebraically to give results following from the superposition theorem. For this purpose the numbers in each part of the two-part subscript are taken as separated by a minus sign and the colon is taken as a sign of multiplication; thus:

$$Z_{12:12} = Z_{(1-2)(1-2)} = Z_{11} + Z_{22} - 2Z_{12},$$

the last expression being formed by writing out the indicated product and separating the terms. The equation expresses the fact that the impedance of a circuit may be subdivided into the self-impedances of sides (real or fictional) associated with its terminals minus twice their mutual impedance. Moreover, any additional subscripts desired may be intercalated by adding and subtracting the same numeral; the expansion of bracketed terms then gives a relation between circuit impedances; thus:

$$\begin{aligned} Z_{(1-2)(1-2)} &= Z_{[(1-3)+(3-2)][(1-3)+(3-2)]} \\ &= Z_{(13+32)(13+32)} \\ &= Z_{13:13} + Z_{32:32} + 2Z_{13:32}. \end{aligned}$$

The first two equivalent networks are of the H type; [6] as seven impedances are shown on each, whereas only six are required for complete representation, an arbitrary constant α has been introduced so that the mutual impedance of the uprights may be varied at pleasure. Thus in the first H network, the condition $\alpha + \nu = 0$ puts all the mutual impedance between the uprights below the crossbar; the condition $1 - \nu - \alpha = 0$ puts it all above. The same type of shift may be made in the second H network.

The third equivalent is the network of direct impedances ("Cisoidal Oscillations," loc. cit. designation (b)); these are expressed in terms of the transducer parameters with opposite pairs of terminals short-circuited, which following Campbell are denoted by S's. Thus $S_{12:12}$ is the driving-point impedance between terminals 1 and 2 with terminals 3 and 4 short-circuited; $S_{12:34}$ is the ratio of current from 3 to 4 to voltage from 1 to 2 with 1, 2 energized and 3, 4 short-circuited, or its reciprocity theorem equivalent.

These three equivalents correspond respectively to transformer, T and π transducer equivalent networks. For the first H type the transducer condition that currents into terminals 1 and 2, and 3 and 4, shall be equal and opposite entails zero current in the H crossbar, which may be removed, leaving a transformer connection. For the second H type the transducer condition allows grouping the impedances of branches 1 and 2 and their mutual impedance, and of 3 and 4 and their mutual impedance, into single branches, say, branches 1 and 3, which gives the T equivalent network. The reduction of the direct impedance network is not so immediate.

Periodic Ladder Networks

When the network is periodic, the transducer impedances and current distribution may be expressed completely in terms of the

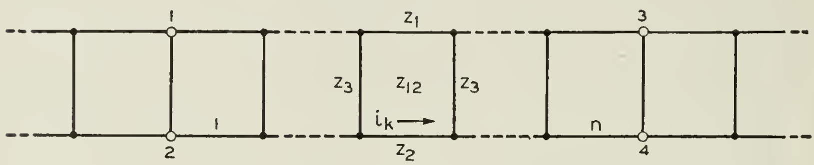
The justification of the operation lies in the fact that, as regards the current half of the subscript, a unit current from 1 to 2 is equivalent by the superposition theorem to unit currents 1 to 3 and 3 to 2 and similarly the voltage 1 to 2 for unit current 1 to 2 is the same as the sum of voltages 1 to 3 and 3 to 2. Thus the notation is a shorthand for application of the superposition theorem. Its use is illustrated further in the course of the proof of the theorem.

[6] This is a form of equivalent network falling under designation (c) of the list of equivalents for an arbitrary number of terminals given by G. A. Campbell ("Cisoidal Oscillations," p. 889, loc. cit.), which is described as branches radiating from a common concealed point, one to each of the terminals, with mutual impedances between pairs. This is not a unique representation since the number of elements is redundant, and the set of mutual impedances may be given values appropriate for particular purposes provided that the self-impedances are adjusted correspondingly. In the present application the mutual impedances of branches to terminals 1 and 4, and 2 and 3, have been set at zero and the mutual impedances of branches 1 and 2 and 3 and 4 in the first H diagram, and of branches 1 and 3, 2 and 4 in the second, have been eliminated in forming the cross bar of the H.

section impedances, giving a certain concreteness to the application of the theorem, which may be valuable. Partly on this account, and partly because periodic iterative impedances are in themselves useful in applications, a particular type of periodic network is considered in this section.

The network considered is infinite in extent, with section series impedances z_1, z_2 and z_{12} and shunt impedance z_3, as shown on Fig. 2*A*.

A. Network Diagram



B. Network Impedances $Z_{ab:cd}$

cd \ ab	12	13	14	24
12	K			
13	$\nu(K-T)$	$Z+2\nu^2(K-T)$		
14	$T+\nu(K-T)$	$Z-\nu(1-2\nu)(K-T)$	$Z+K-2\nu(1-\nu)(K-T)$	
23	$-K+\nu(K-T)$	$Z-\nu(1-2\nu)(K-T)$	$Z-T-2\nu(1-\nu)(K-T)$	$Z+(1-\nu)(1-2\nu)(K-T)$
24	$-(1-\nu)(K-T)$	$Z-2\nu(1-\nu)(K-T)$	$Z+(1-\nu)(1-2\nu)(K-T)$	$Z+2(1-\nu)^2(K-T)$
34	T	$-\nu(K-T)$	$(1-\nu)K+\nu T$	$(1-\nu)(K-T)$

C. Current Distribution

Unit Current Between:	i_k
1 and 2	$-\dfrac{K_1}{K_1+K_2}e^{-k\alpha}$
1 and 3	$\nu-\dfrac{K_1}{K_1+K_2}[\nu e^{-k\alpha}+\nu e^{-(n+1-k)\alpha}]$
1 and 4	$\nu-\dfrac{K_1}{K_1+K_2}[\nu e^{-k\alpha}-(1-\nu)e^{-(n+1-k)\alpha}]$
2 and 3	$\nu+\dfrac{K_1}{K_1+K_2}[(1-\nu)e^{-k\alpha}-\nu e^{-(n+1-k)\alpha}]$
2 and 4	$\nu+\dfrac{K_1}{K_1+K_2}[(1-\nu)e^{-k\alpha}+(1-\nu)e^{-(n+1-k)\alpha}]$
3 and 4	$\dfrac{K_1}{K_1+K_2}e^{-(n+1-k)\alpha}$

Fig. 2—Periodic ladder network of infinite extent; network diagram, impedances and current distribution.

The infinite network is the simplest to formulate since there are no points of reflection; it is of course symmetrical with respect to the terminal pairs 1, 2 and 3, 4.

The impedance across either of these terminal pairs is the parallel impedance of the full-series and full-shunt iterative impedances (or one-half the mid-shunt iterative impedance). The full-series and full-shunt iterative impedances are given by the following formulas:

$$\begin{aligned} &\textit{Full-series } K_1 = \tfrac{1}{2}[\sqrt{z(z + 4z_3)} + z] = K_2 + z, \\ &\textit{Full-shunt } K_2 = \tfrac{1}{2}[\sqrt{z(z + 4z_3)} - z] = \frac{K_1 z_3}{K_1 + z_3}, \end{aligned} \tag{3}$$

where, for brevity, $z = z_1 + z_2 - 2z_{12}$.

Then

$$Z_{12:12} = Z_{34:34} = K = \frac{K_1 K_2}{K_1 + K_2} = z_3\left[1 + \frac{4z_3}{z}\right]^{-1/2}. \tag{4}$$

The voltage across lines is propagated as exp $(-\,k\alpha)$ where α is the section propagation constant; hence

$$Z_{12:34} = T = Ke^{-n\alpha}. \tag{5}$$

The propagation factor exp $(-\,\alpha)$ is defined in terms of the iterative impedances by

$$e^{-\alpha} = \frac{K_2}{K_1}\cdot \tag{6}$$

The currents $i_{k:12}$ and $i_{k:34}$ are given by the following formulas:

$$i_{k:12} = -\frac{K_1}{K_1 + K_2}e^{-k\alpha}, \tag{7}$$

$$i_{k:34} = \frac{K_1}{K_1 + K_2}e^{-(n+1-k)\alpha}. \tag{8}$$

This completes the formulation, since the remaining quantities, ν and Z, are given immediately by

$$\nu = \frac{z_1 - z_{12}}{z_1 + z_2 - 2z_{12}},$$

$$Z = \sum_{k=1}^{n} \frac{z_1 z_2 - z_{12}^2}{z_1 + z_2 - 2z_{12}} = n\frac{z_1 z_2 - z_{12}^2}{z_1 + z_2 - 2z_{12}}\cdot$$

Figure 2*B* shows driving-point and transfer impedances for energization between terminals, omitting certain impedances equal by symmetry. Figure 2*C* shows the corresponding *k*-section currents in side 2.

Applications to Electric Railroad Networks

A.-c. electric railroad networks in one-line diagram are predominantly of the ladder type. The series elements of sides 1 and 2 represent, for two-wire networks, impedances of sections of transmission lines and traction circuits, respectively; the shunt elements represent transformer impedances. For three-wire networks, the series elements may represent trolley-feeder (or feeder-rail) and trolley-rail impedance elements, the shunt elements autotransformer impedances.

The theorem may be used for representing portions of a network or a whole network of ladder form,[7] when the series impedances satisfy the condition of the theorem. As the circuits are linearly extended this is almost always the case except where the traction circuits change character, from two to four tracks, for example. For approximate purposes the H networks may be used even in these cases provided that the parameter ν is properly chosen. In many cases the transfer impedance $Z_{12:34}$ is negligible and a value of ν may be associated with each pair of terminals; the values for the sections immediately adjoining the terminal pairs 1–2 and 3–4 (sections 1 and n on Fig. 1A) are of dominant importance and serve for rough purposes. If the transfer impedance is not negligible a mean of these values may be sufficiently accurate.

In two-wire networks, generator circuits are connected directly to the transmission line (side 1), and the short circuits of chief interest (grounding points on the one-line diagram) are those on the traction circuits (side 2). Thus, for a single generating point the network is energized between points on sides 1 and 2, such as 1 and 4, for example; if the impedance in the generator connection is Z_g and the impedance of the short circuit is zero, the short-circuit driving-point impedance and the traction circuit currents are as follows:

$$\begin{aligned} Z_0 &= Z_g + Z_{14:14} \\ &= Z_g + Z + \nu^2 Z_{12:12} + (1-\nu)^2 Z_{34:34} + 2\nu(1-\nu)Z_{12:34}, \end{aligned} \tag{9}$$

$$I_k = [\nu + \nu i_{k:12} + (1-\nu) i_{k:34}]\frac{E}{Z_0}, \tag{10}$$

where E is the generator voltage. The impedance may be obtained immediately from either of the H networks—as the sum of the self-

[7] For multiple transmission-line two-wire networks the ladder form is obtained when all transmission lines are bussed at all generating stations and substations or when the generators, step-up transformers and substation transformers connected to each line are of similar impedances and are similarly connected. When these conditions are not met the network is of multiple-side ladder form, for the representation of which an extension of the theorem would be required. Similar remarks apply to three-wire networks.

impedances of legs 1 and 4 and of the crossbar. The current expression follows from equation (2) with $I_2 = I_3 = 0$; $-I_1 = I_4 = E/Z_0$.

For multiple generating points (or for multiple points of short

A. Short Circuit Between Generator Points

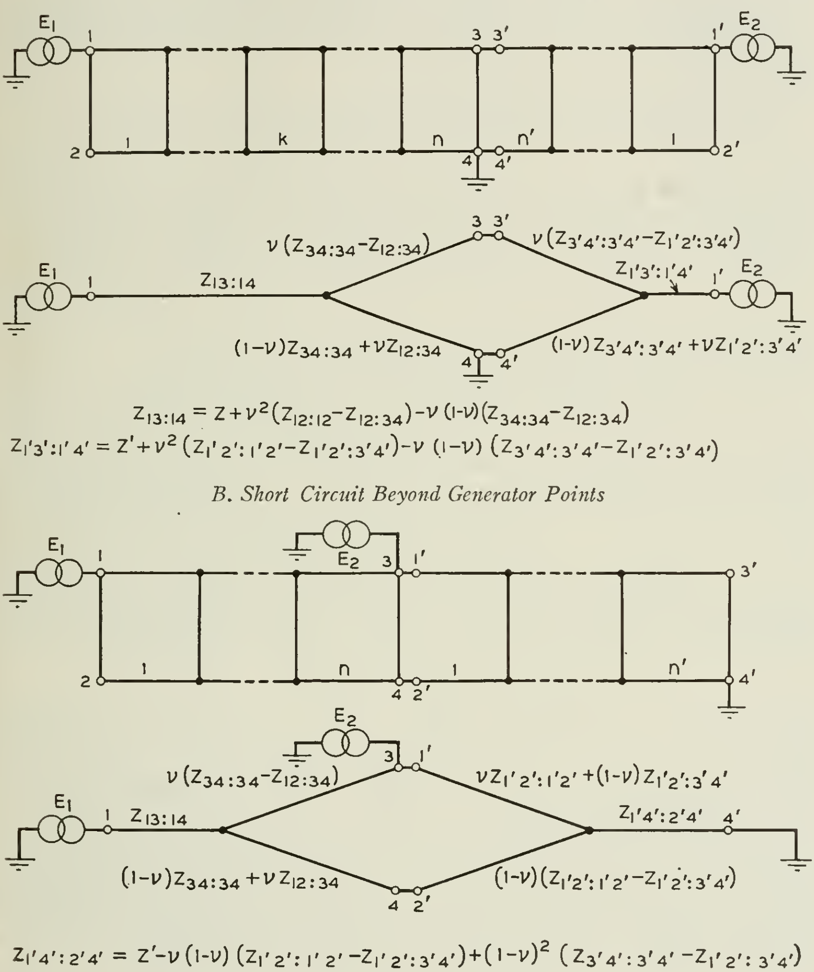


Fig. 3—Equivalent networks for electrified railways; two-wire system, two sources.

circuit) the theorem may be used to represent portions of the network. Examples for the case of two generators are shown on Fig. 3; on Fig. 3*A* the short circuit is located between generator points, on Fig. 3*B* beyond them. On Fig. 3*A* the network is supplied with duplicate pairs of terminals at the short-circuit point, separated by an infinitesimal difference; the parts of the original network thus formed are represented by Y-connected impedances which may be derived from the first H network. On Fig. $3B$ the network is broken at the intermediate generator point and similarly represented. A similar process may be followed for any number of generator points but in some cases it may be expedient to superpose additional generators; the network impedances required for superposition may be formulated in the manner followed in the proof of the theorem.

The solution of these reduced networks supplies the currents I_1, I_3, I_4; I_1', I_3', I_4', etc., from which the current in branch k of side 2 of any of the ladder sections may be found from equation (2). Thus, for example the current I_k in the kth section of the ladder network with terminals 1, 2, 3 and 4 on Figs. $3A$ and $3B$ is formulated as follows:

$$I_k = [\nu + \nu i_{k:12} - \nu i_{k:34}]I_3 + [\nu + \nu i_{k:12} + (1 - \nu)i_{k:34}]I_4, \quad (11)$$

which follows from equation (2) with $I_2 = 0$; $- I_1 = I_3 + I_4$. Similar formulas apply to the other two ladder networks on Fig. 3.

In three-wire networks, generators are usually connected to the traction network by three-winding transformers which may be represented on the network diagram by three impedances connected in star. The traction network may be represented on a trolley-feeder, trolley-rail or a feeder-rail, trolley-rail base and it is well known that the three-winding transformer equivalent impedances for the two bases are related. Using the notation shown on Figs. $4A$ and $4A'$, with primes distinguishing the feeder-rail, trolley-rail base, the relations are as follows:

$$\begin{aligned} V_{tf}Z_a &= V_{tf}Z_a' + V_{tr}Z_b', \\ V_{tf}Z_b &= V_{fr}Z_b', \\ V_{tf}^2Z_c &= V_{fr}^2Z_c' - V_{tr}V_{fr}Z_b', \end{aligned} \quad (12)$$

where V_{tr}, V_{tf} and V_{fr} are the trolley-rail, trolley-feeder and feeder-rail circuit voltages, respectively.

From Figs. $4B$ and $4B'$ showing the reduced networks for trolley-rail short-circuits on the two bases for a single source feed, it is apparent that the impedances involved in the equivalent networks must be similarly related. The relations are found to be as follows:

$$\begin{aligned} V_{tf}^2Z_{12:12} &= V_{fr}^2Z'_{12:12}, \\ V_{tf}^2Z_{34:34} &= V_{fr}^2Z'_{34:34}, \\ V_{tf}^2Z_{12:34} &= V_{fr}^2Z'_{12:34}, \\ V_{fr}(1 - \nu) &= V_{tf}(1 - \nu'), \\ Z &= Z', \\ i_{k:12} &= i'_{k:12}, \\ i_{k:34} &= i'_{k:34}. \end{aligned} \quad (13)$$

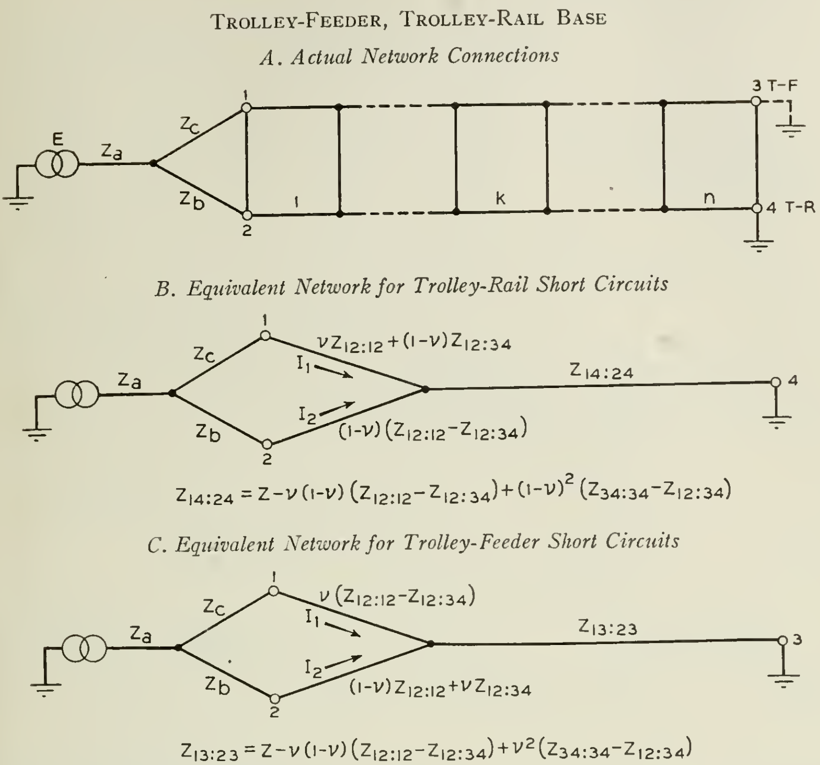


Fig. 4—Equivalent networks for electrified railways; three-wire system, single source, trolley-feeder, trolley-rail and feeder-rail, trolley-rail bases.

Thus the complete set of short-circuit currents (trolley-rail, trolley-feeder and feeder-rail short circuits) may be made from a single determination of the transducer impedances and current distributions on either of the two bases, whenever the theorem is applicable.

For multiple generator three-wire systems, and for three-wire systems with auxiliary transmission lines, the theorem may be used to represent portions of the network, possibly broken as in the two-wire cases illustrated above, four-terminal representation being necessary in general. The application follows the lines indicated above.

Proof of Theorem

For energization between terminals 1 and 2, the sum of the currents in sides 1 and 2 at any point on the ladder is zero; the current $i_{k:12}$ may be taken as flowing in a mesh made up of the k-section sides and its terminating shunt impedances. For unit current supplied, the

FEEDER-RAIL, TROLLEY-RAIL BASE

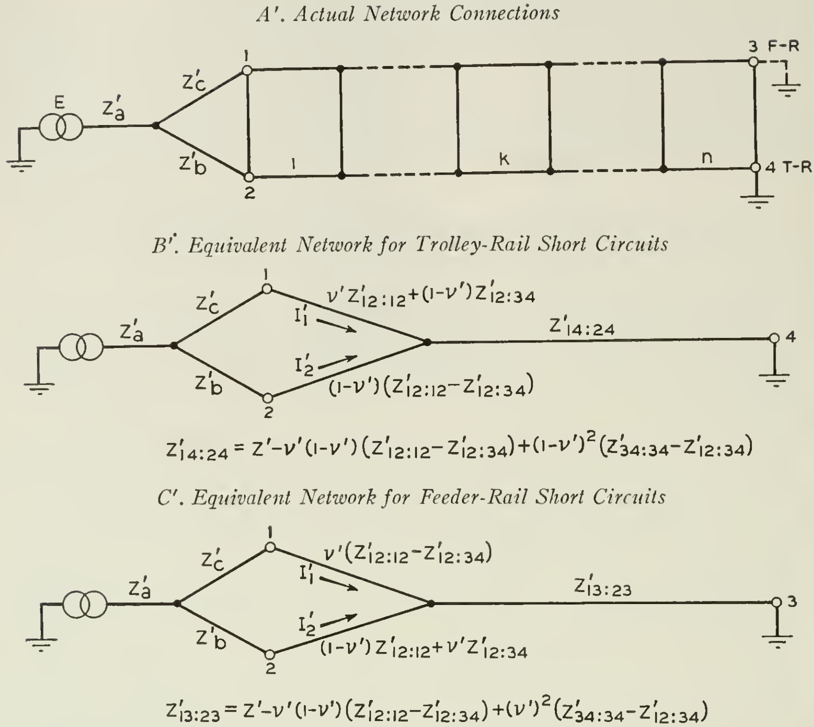


Fig. 4—*Continued from page* 346.

driving-point impedance between terminals 1, 2 and the transfer impedance to terminals 3, 4 for the network shown on Fig. 1*A* may be formulated immediately as:

$$\begin{aligned} Z_{12:12} &= (1 + i_{1:12})Z^0, \\ Z_{12:34} &= -\, i_{n:12}Z^{(n)}. \end{aligned} \tag{14}$$

The positive sense for currents $i_{1:12}$ and $i_{n:12}$ is taken as indicated on Fig. 1*A*, namely, in the direction from terminal 2 to terminal 4 on side 2.

From the voltage equation around the loop formed from sides 1 and 2 in their entirety and the terminal shunt impedances, the difference of these impedances may be expressed by:

$$Z_{12:12} - Z_{12:34} = -\sum_{k=1}^{n} (Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)})i_{k:12}. \tag{15}$$

The transfer impedances with respect to the side terminals 1, 3 and 2, 4 are formulated as:

$$
\begin{aligned}
Z_{12:13} &= -\sum_{k=1}^{n} (Z_1^{(k)} - Z_{12}^{(k)}) i_{k:12},\\
Z_{12:24} &= \sum_{k=1}^{n} (Z_2^{(k)} - Z_{12}^{(k)}) i_{k:12}.
\end{aligned}
\tag{16}
$$

From the condition $[Z_1^{(k)} - Z_{12}^{(k)}][Z_2^{(k)} - Z_{12}^{(k)}]^{-1} = \text{const.}$, a constant ν may be defined such that:

$$
\begin{aligned}
\nu = \nu_k &= [Z_1^{(k)} - Z_{12}^{(k)}][Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}]^{-1},\\
1 - \nu &= [Z_2^{(k)} - Z_{12}^{(k)}][Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}]^{-1},
\end{aligned}
$$

and equations (16) may be combined with (15) to give:

$$
\begin{aligned}
Z_{12:13} &= \nu Z_{12:12} - \nu Z_{12:34},\\
Z_{12:24} &= -(1 - \nu) Z_{12:12} + (1 - \nu) Z_{12:34}.
\end{aligned}
\tag{17}
$$

The remaining transfer impedances follow by superposition; thus

$$
\begin{aligned}
Z_{12:14} &= Z_{12:13} + Z_{12:34} = \nu Z_{12:12} + (1 - \nu) Z_{12:34},\\
Z_{12:23} &= Z_{12:13} - Z_{12:12} = -(1 - \nu) Z_{12:12} - \nu Z_{12:34}.
\end{aligned}
\tag{18}
$$

It may be observed that

$$
Z_{12:24} = Z_{12:13} + Z_{12:34} - Z_{12:12}. \tag{19}
$$

Similarly

$$
\begin{aligned}
Z_{34:13} &= -\nu Z_{34:34} + \nu Z_{12:34},\\
Z_{34:14} &= (1 - \nu) Z_{34:34} + \nu Z_{12:34},\\
Z_{34:23} &= -\nu Z_{34:34} - (1 - \nu) Z_{12:34},\\
Z_{34:24} &= (1 - \nu) Z_{34:34} - (1 - \nu) Z_{12:34}.
\end{aligned}
\tag{20}
$$

These impedances, together with $Z_{34:34}$ and $Z_{34:12}$, form a set of 12 impedances of which only five are independent. There are three independent impedances determined by energization at each pair of terminals, including $Z_{12:34}$ and $Z_{34:12}$, which are equal by the reciprocity theorem; one independent set, for example, is $Z_{12:12}$, $Z_{12:13}$, $Z_{12:34}$, $Z_{34:34}$, $Z_{34:13}$. Consequently the network may be completely specified by the addition of a single impedance; for the set illustrated, the impedance required is $Z_{13:13}$.

This impedance may be formulated as:

$$\begin{aligned} Z_{13:13} &= \sum Z_1^{(k)} - \sum (Z_1^{(k)} - Z_{12}^{(k)}) i_{k:13} \\ &= \sum Z_1^{(k)} - \nu \sum (Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}) i_{k:13}, \end{aligned} \tag{21}$$

where the summations extend as above from 1 to n.

Writing the equation around the loop used in deriving equation (15) it is found that:

$$\begin{aligned} &\sum (Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}) i_{k:13} \\ &\qquad = \sum (Z_1^{(k)} - Z_{12}^{(k)}) - Z_{13:12} + Z_{13:34} \\ &\qquad = \sum (Z_1^{(k)} - Z_{12}^{(k)}) - \nu (Z_{12:12} + Z_{34:34} - 2Z_{12:34}), \end{aligned} \tag{22}$$

the last step being made by use of the reciprocity theorem and the formulas already developed. Thus, finally:

$$Z_{13:13} = Z + \nu^2 (Z_{12:12} + Z_{34:34} - 2Z_{12:34}), \tag{23}$$

where

$$Z = \sum_{k=1}^{n} \frac{Z_1^{(k)} Z_2^{(k)} - (Z_{12}^{(k)})^2}{Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)}} .$$

As already mentioned, Z is the impedance between short-circuited terminals 1, 2 and 3, 4; this may be verified in a number of ways.

The remaining impedances follow by superposition, which can be carried out formally through the impedance notation in the manner suggested. There are 21 driving-point and transfer impedances between terminals which can be displayed in a triangular array similar to that shown on Fig. 2*B*. The additional measurable impedances at the terminals arise as follows: 36 from short-circuiting two terminals, 4 from short-circuiting three terminals and 3 from short-circuiting terminals in pairs.

Equation (22) may also be written in terms of currents $i_{k:12}$ and $i_{k:34}$, since $Z_{13:12}$ and $Z_{13:34}$ may be expressed in terms of the latter; this suggests the following relation:

$$i_{k:13} = \nu + \nu i_{k:12} - \nu i_{k:34}. \tag{24}$$

The relation is verified by substituting into the mesh equations for currents $i_{k:13}$; the typical equation is as follows:

$$\begin{aligned} - Z^{(k-1)} i_{k-1:13} + [Z_1^{(k)} + Z_2^{(k)} - 2Z_{12}^{(k)} + Z^{(k-1)} + Z^{(k)}] i_{k:13} \\ - Z^{(k)} i_{k+1:13} = Z_1^{(k)} - Z_{12}^{(k)}. \end{aligned}$$

The remaining current relations then follow by superposition, as follows:

$$\begin{aligned} i_{k:14} &= i_{k:13} + i_{k:34} = \nu + \nu i_{k:12} + (1 - \nu) i_{k:34}, \\ i_{k:23} &= i_{k:13} - i_{k:12} = \nu - (1 - \nu) i_{k:12} - \nu i_{k:34}, \\ i_{k:24} &= i_{k:14} - i_{k:12} = \nu - (1 - \nu) i_{k:12} + (1 - \nu) i_{k:34}. \end{aligned} \tag{25}$$

It will be observed that only three of the six currents $i_{k:12}$, $i_{k:13}$, $i_{k:14}$, $i_{k:23}$, $i_{k:24}$ and $i_{k:34}$ are independent; one independent set is $i_{k:12}$, $i_{k:34}$ and $i_{k:13}$. Hence any arbitrary set of currents I_1, I_2, I_3 and I_4 flowing out of the network at the terminals may be resolved into three flows, such as those illustrated in the independent set above, which leads to equation (2).

The first H network may be obtained in the following manner. The value of $Z_{12:13}$, namely, $\nu Z_{12:12} - \nu Z_{12:34}$, in conjunction with the condition $Z_1 + Z_2 = Z_{12:12}$, Z_1 and Z_2 being the impedances of branches to terminals 1 and 2, respectively, suggests the following values of branch self and mutual impedances:

$$\begin{aligned} Z_1 &= \nu Z_{12:12}, \\ Z_2 &= (1 - \nu) Z_{12:12}, \\ Z_{13} &= \nu Z_{12:34}. \end{aligned}$$

The value of Z_{13} is verified by inspection of $Z_{13:34}$ if

$$Z_3 = \nu Z_{34:34}.$$

Similarly, by inspection of $Z_{12:14}$ and $Z_{34:32}$, the values of Z_{24} and Z_4 may be tentatively set at

$$\begin{aligned} Z_{24} &= (1 - \nu) Z_{12:34}, \\ Z_4 &= (1 - \nu) Z_{34:34}. \end{aligned}$$

The impedance of the crossbar, say Z_0, may be found from any of the impedances $Z_{13:13}$, $Z_{14:14}$, $Z_{23:23}$, $Z_{24:24}$; e.g.,

$$\begin{aligned} Z_0 &= Z_{13:13} - (Z_1 + Z_3 - 2Z_{13}) \\ &= Z - \nu(1 - \nu)(Z_{12:12} + Z_{34:34} - 2Z_{12:34}). \end{aligned}$$

But the presence of seven elements, as already mentioned, suggests an arbitrariness which may be put into evidence by adding $\alpha Z_{12:34}$ to Z_1, which entails a similar addition to Z_3 and Z_{13}, and a similar subtraction from Z_2, Z_4 and Z_{24}.

Similar considerations apply to the derivation of the second H network.

The direct impedances may be found, in a well-known manner, by energizing the network between one terminal and the other three short-circuited terminals or by applying a formula due to G. A. Campbell.[8]

Acknowledgments

I am indebted to Mr. E. D. Sunde for a suggestion which brought this theorem into focus for me; Dr. H. M. Trueblood has aided in eliminating certain weak points in the proof; to Mr. R. M. Foster I owe a number of improvements, aside from that mentioned in the footnote, in the general spirit of the paper. My chief acknowledgment, however, as indicated by the footnotes referring to his work, is to Dr. G. A. Campbell; I should like the paper to be taken as an instance of the fertility and generality of the methods and results of network analysis he has introduced.

[8] "Direct Capacity Measurement," *Bell System Technical Journal*, I, 1, pp. 18–38 (July, 1922). The formula, given on page 34, requires modification only to the extent of substituting impedances for capacities; for four terminals the direct impedance, D_{ij}, between terminals i and j is given by:

$$D_{ij} = \frac{\Delta}{2\Delta_{ij}},$$

where Δ_{ij} is the co-factor of the element in row i, column j of the determinant:

$$\Delta = \begin{vmatrix} 0 & Z_{12} & Z_{13} & Z_{14} & 1 \\ Z_{12} & 0 & Z_{23} & Z_{24} & 1 \\ Z_{13} & Z_{23} & 0 & Z_{34} & 1 \\ Z_{14} & Z_{24} & Z_{34} & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{vmatrix}.$$

The elements of the determinant are the driving-point impedances between terminals indicated by the subscripts (all other terminals open), namely, $Z_{12:12}$, $Z_{13:13}$, etc., written for brevity Z_{12}, Z_{13}, etc.

Contemporary Advances in Physics, XXXI—Spinning Atoms and Spinning Electrons [1]

By KARL K. DARROW

No doubt you are all accustomed to thinking of atoms as objects—very small objects, of course—which are endowed with *weight*. I can say that with perfect safety to an audience of engineers and physicists; but indeed it can be said with safety to any audience—I mean, of course, any audience literate enough to attach any meaning at all to such a word as "atom." It may be that philosophers of the past have imagined weightless atoms—I am not historian enough to deny that, nor to affirm it; but if such have ever been invented, they have remained quite outside the currents of modern thought. For us, weight is a property which we attribute to the atom. Since this is, after all, a professional audience, I will now change over to that other word which many people have such difficulty in distinguishing from "weight": I will say that *mass is a property which we attribute to the atom*. In a way, that is a negative statement. It means that we do not hope to explain mass in terms of something more fundamental; it means that we accept mass as being itself so fundamental that even the elementary particles have it. When I say "elementary particles," I am still referring in part to the atoms, though it is a somewhat careless usage to do so; but I am referring also to electrons both positive and negative, to protons, to alpha-particles, to nuclei—to all the particles, in effect, of which the atoms are built up. Also I ought to include the corpuscles of light, but this lecture will be quite long enough if I leave them almost unmentioned. All of these particles, then, are endowed with mass; each of them has a characteristic mass of its own, which we do not attempt to explain, but which we do try to measure as closely as we can.

There is another property, familiar to you though not to everyone, which we accept as equally fundamental and equally unexplainable with mass: it is *electric charge*. We attribute it also to the elementary particles, though not, it is true, to all of them. We assign it to the electrons, of course, and to protons and alpha-particles and all of the hundreds of nuclei which distinguish the elements and the isotopes

[1] A lecture delivered before the American Physical Society at Chicago on November 27, 1936, and before the American Institute of Electrical Engineers at New York on May 6, 1937.

thereof from each other. When I say that we "assign" it, I do not for a moment mean that we are doing something arbitrary or untestable. We *know* that these elementary particles are charged, and indeed we have measured their charges. Atoms do normally appear to us uncharged, but we know that that is because the elementary particles of which they in their turn are made up are some of them positive, some of them negative, and the balance normally happening to be perfect. Some particles, the neutrons and the corpuscles of light, seem permanently chargeless; but perhaps some day we shall find it expedient to regard them as groups of smaller particles having charges which balance one another. Apart from these two cases, we may say with assurance that whenever we penetrate as far into the fine structure of substance as we are able to go, we find the elementary particles invested with mass and with charge.

And now I arrive at the subject of this talk, the *third* property of the elementary particles: the property which is called "angular momentum," or "spin" for short. Now of course I am speaking as to an audience of physicists, for if this were an audience of laymen it would certainly be frightened by such a term as "angular momentum." This is a misfortune, and perhaps a defect of general education; for angular momentum is about as important as mass or charge, not only on the scale of the elementary particles but also on the scale of the visible world. Think what it would mean if there were no such thing as the conservation of angular momentum! the earth might cease from turning, it might cease from providing us with the regular alternation of day and night, and with our standard of the flow of time; it might even cease from traversing its regular orbit, and fly off into space or fall into the sun. Well, I do not wish to scare you with any such dire imaginings—I only want to remark on the fact that the human race has been acquainted for a very long time indeed with angular momentum as something which is unvarying, imperturbable, incessant; for of all the unvarying and imperturbable and incessant things in the world, the rotation of the earth is the most obvious and the most striking. So striking it is, that you might reasonably expect that all the philosophers and all the physicists of the past would have conferred the property of spin on all the atoms which they have invented. Well, they did not; the notions of the spinning atom, the spinning electron, the spinning nucleus are among the newest in physics. I think that some of the reasons for the delay will be evident later on in this talk, but it remains partly mysterious, at least to me. Looking back on the situation with the well-known advantages of hindsight, I do feel a good deal of surprise that the spinning atom did not make an earlier entry upon the scientific stage. Perhaps some of you will remember hearing the words "vortex atom" and "vortex theory" which used to be so prominent in physics, and will take the spinning atom of today for a lineal descendant of those vortices of old. If this were correct, we could trace back the ancestry of the spinning atom for about three hundred years; but I think that it is not correct. The vortices of which Descartes and Malebranche were dreaming three centuries ago were more like whirlpools of streaming particles, and the vortices which were imagined by Helmholtz and Lord Kelvin some fifty years ago were also whirlpools, but they were whirls in an idealized continuous frictionless fluid. Let us pause for a moment to notice how the attitude of physicists has altered in these fifty years! Kelvin and Helmholtz began with the idea of an aethereal fluid pervading the whole of space, and valiantly tried to represent the atoms as whirlpools in that fluid; but we have long since discarded that aether, and our spinning atoms and other elementary particles are small delimited rotating bodies voyaging in a void.

It is nor therefore the vortex which I will introduce to you as the ancestor of the spinning atom, but rather the "Amperian whirl" as it still is sometimes called. You remember, of course, how Ampère in 1820 made a very great achievement which for the purposes of this talk I will divide into three. First, he discovered the fact that an electric current flowing in a circuit is equivalent to a magnet. Next, he worked out the mathematical laws whereby, given a current and the circuit in which that current is flowing, we may calculate the strength or the moment of the equivalent magnet. I will write down the formula for the case of a current i, flowing in a plane loop of area A: the magnetic moment of the equivalent magnet, μ, is given thus:

$$\mu = iA/c.$$

Here c is a factor of transformation which we are now obliged to employ because we habitually use, in atomic physics especially, a unit-system different from Ampère's. The third part of the great achievement was this: Ampère founded what remains to this day the theory of magnetism, by presuming that *the individual atoms of any magnetizable substance are themselves little magnets*, and that *the atoms are magnets because they have little whirls of current in them.*

This notion—the notion that atoms are magnets, and that they are magnets because they have internal circulating currents—is the true forerunner of our present conception of the spinning atom. It is, however, only a very primitive form of the modern conception, and there is much to be added to it. First of all and above all, there is the question of angular momentum. Is angular momentum an attribute of these whirling intra-atomic currents, or is it not? You may think that the answer to this question is self-evidently "yes!" but remember that for many generations of our forefathers electricity was an *imponderable* fluid. Weber, however, did consider the affirmative answer, and Maxwell even attempted to ask the question of Nature by experiment—vainly, as it turned out. Not till the electron was discovered did the mass of electricity become a prominent part of experience. A moment ago I divided Ampère's achievement into three parts; similarly I wish now to divide the discovery of the electron into three. Those who isolated and identified and measured the electron were proving three things: first, that negative electricity consists of corpuscles of a definite charge, e; second, that these corpuscles have a mass, m; and following from these two, the principle which I have called the third part of the discovery, *viz.* that an electron revolving in an orbit has an angular momentum.

I will designate angular momentum in general by the letter p, and now I will show you a formula for the ratio of μ to p in an atom in which an electron is running around in an orbit and constituting an Amperian whirl. The formula, like this other one, for μ, is valid for an orbit of any shape, but to get it quickly I will simplify by postulating a circular orbit. The radius of the circle being r, the area A is πr^2; the current around it is equal to the electron-charge e, multiplied by the number of times per second that the electron runs around the orbit; if I denote the velocity of the electron by v, this number is $v/2\pi r$; hence the product iA/c is equal to $evr/2c$. Now the angular momentum p of the electron, as you all know, is mvr; and hence for the ratio I derive:

$$\mu/p = e/2mc,$$

which is one of the most important formulae in the whole of atomic physics. You notice that it does not involve in any way the size or shape of the orbit or the frequency with which the electron travels around it. It is the same for any or all of the revolving electrons of any atom of any kind.

Now let us see how this formula may be tested. Imagine a rod of some highly magnetizable metal, iron for instance, and imagine it to be unmagnetized at the start. This means, that at the start the little atomic magnets are pointing at random in all directions; that is to say, the vectors which represent their magnetic moments are pointing every way, and so are the vectors which represent their angular momenta, the latter being parallel to the former. Since these atomic vectors of angular momentum are pointing every way at random, they add up to zero, and the rod as a whole possesses no resultant angular momentum; it is just standing still. Now let the rod be surrounded with a solenoid, and by means of a current in the solenoid let it be magnetized to saturation. Now all the arrows representing magnetic moments are pointing parallel to the axis of the rod. But so are all the arrows representing atomic angular momenta! their resultant is no longer zero—suddenly there has arisen a *resultant angular momentum*, belonging to the totality of all the atomic magnets, and quite large enough to be detected, instead of being tiny like the angular momentum of an individual atom. Unless our theory is fundamentally wrong somewhere, we should be able to observe this resultant angular momentum.

The experiment is done by hanging the rod vertically from a fine suspension, and sending the magnetizing current through the solenoid. At the instant of the magnetization, the rod turns sharply on its axis, twisting the suspending fibre. Thus it manifests the angular momentum of which I have just been speaking—though I ought to say that what we observe is of the nature of a recoil, or back-kick: when the totality of the little atomic magnets suddenly acquires its resultant angular momentum, the substance of the rod as a whole acquires an equal and opposite amount (so as to keep constant the total amount of angular momentum in the universe) and it is the latter which we detect. The experiment is quite a delicate one, but its technique has been developed to a remarkable degree since it was first attempted twenty years ago by Einstein and de Haas. What we measure is the ratio of the magnetization of the rod-as-a-whole to the angular momentum of the rod-as-a-whole; and this is just the same as the ratio of μ to p for the elementary atomic magnets. There are not many properties of matter of which we can say that the value measured on a large piece of matter is the same as the value for the individual atom; but there are a few, and this is one of them.

Now in giving you the result, let me first emphasize the general principle that here we have evidence of the spinning of elementary particles, and of the interrelation between spinning and magnetism. Next, I give you the numerical result itself. For iron and nearly all of the other ferromagnetic materials, we find:

$$\mu/p = e/mc$$

or *twice* the theoretical value which I quoted a moment ago.

This cannot be explained by assuming any peculiarity of size or shape or frequency of the electron-orbits in the atoms, for as I just said the theoretical formula is independent of all these things. We are obliged to make some more drastic assumption. If I had unlimited time before me, I might sketch the history of our assumption; but as I don't, I will come straight to the present situation. We assume first, that in the iron atoms in the rod the electron-orbits are so oriented with respect to each other that their magnetic moments kill one another off completely. We then assume that every electron has a magnetic moment and an angular momentum of its own, intrinsic to it and inherent in it, and altogether independent of whether or not the electron is revolving in an orbit. Just as the earth has a rotation of its own in addition to its elliptical course around the sun, so we imagine that the electron has a rotation of its own; this rotation has an angular momentum, and with it there is connected a magnetic moment. (You will remember doubtless that the earth also has a magnetic moment, but this is one of the analogies which it is better not to force too far.) When we magnetize the iron rod, it is the *electrons* which we are turning; the vectors which we cause to point all in the same direction are the magnetic moments and the angular momenta which are inherent in the electrons, and the value of their ratio is the value which is characteristic of the "spinning electron," as we call it. Therefore, amplifying the notation a little, I write:

$$\mu/p = g(e/2mc)\begin{cases} g = 1 \text{ for electron-orbits,} \\ g = 2 \text{ for spinning electron,} \end{cases}$$

and now I leave the spinning electron for a few minutes, in order to turn again to the theory of electrons revolving in their orbits.

You all realized, of course, that when I converted the Amperian whirl of current into an electron running around an orbit, I was adopting the atom-model known by the names of Rutherford and Bohr; for these were the original thinkers who impelled all the rest of us, following in their footsteps, to think of the atom as a positively-charged nucleus around which electrons are revolving like planets around the sun. This is an atom-model in which magnetism is inherent—*a Rutherford-Bohr atom is intrinsically a magnet.* Anyone who did not know the history of the model might well assume that it was designed expressly to account for magnetism, and any such person might also quite reasonably assume that all the physicists of the early nineteenth century thought of it simultaneously as soon as the electron was discovered. Well, it was *not* designed expressly to account for magnetism, and most of the physicists of the early nineteenth did *not* think of it—or if they did, they thought of it only to reject it. At that time, the atom-model with the orbital electrons seemed to be disqualified by a very potent reason; for according to the classical electromagnetic theory, an electron revolving in an orbit ought to radiate all of its energy in a very short time and fall into the nucleus. Bohr was the man who overrode this objection. He overrode it, not in order to construct a theory of magnetism in defiance of it, but in order to construct a theory of spectra in defiance of it. This theory has been extraordinarily successful. Our theory of magnetism is hardly more than a by-product of that theory of spectra; and this, in an odd sort of way, enhances its credit. A theory devised expressly for a certain purpose is always less impressive than one which follows incidentally from a successful theory devised for quite another purpose; and the contemporary theory of magnetism is a wonderful example of this latter and more impressive type.

The main element of Bohr's theory of spectra—if one can speak of one element as the main one, which is really not quite proper—is an assumption about the angular momentum of the electron in its orbit or, let me say, the angular momentum p of the electron-orbit. It was assumed that the electron may revolve, without radiating its energy, in any orbit of which the angular momentum is an integer multiple of $h/2\pi$, — h now standing, of course, for the famous quantum-constant of Planck which is the emblem of modern physics. I write this down as follows:

$$p = (1, 2, 3, 4, \cdots)(h/2\pi).$$

Bohr was thinking at first about the hydrogen atom; but hydrogen is an inconvenient example to use in talking about magnetism, and iron is a very complicated case indeed, so I will talk entirely about the sodium atom.

The sodium atom has a nucleus with a charge of $+ 11e$, and eleven electrons circulating in orbits around it. This certainly sounds formidably complex, but it happens—and I shall later remind you of this fact—that the orbits and also the spins of ten of the electrons are so oriented with respect to one another that their angular momenta and their magnetic moments completely neutralize each other. I shall therefore ask you to imagine these ten inner electrons as a sort of cloud. The eleventh electron of the sodium atom—known technically as the "valence" electron—cruises around this system; sometimes it is traveling in an orbit completely outside the cloud, sometimes in an orbit which cuts across the cloud, but never in an orbit which is entirely or even mainly inside the cloud. The ten inner electrons which constitute the cloud neutralize a part of the force with which the nucleus acts upon the valence-electron; but they make—I repeat —not the slightest contribution to the angular momentum or to the magnetic moment of the atom.

I have given above, the permitted values of the angular momentum of this orbit of the valence-electron. Now I point out that to each of these permitted values of p corresponds a permitted value of the magnetic moment μ, which I obtain by multiplying the former with $e/2mc$:

$$\mu = (1, 2, 3, 4, \cdots)(eh/4\pi mc).$$

However, only one (at most) of these values can be appropriate to the normal state of the sodium atom; all the rest must correspond to abnormal, unusual, or "excited" states. We are going to be interested primarily in the normal state, so we must identify the right one. In the early days of the Bohr theory, the right one was supposed to be the first which I have written down. However, the theory has been greatly remodelled and bettered since those days, with the aid of what is known as "quantum mechanics"; and it now seems quite certain that these lists of the permitted values of p and μ for electron-orbits are both incomplete. I must add to each of them the value *zero*, so that the two lists become

$$p = (0, 1, 2, 3, 4, \cdots)h/2\pi,$$
$$\mu = (0, 1, 2, 3, 4, \cdots)eh/4\pi mc.$$

Moreover, it is precisely this new value *zero* which belongs to the normal state of the sodium atom. So the theory, in this stage, quite definitely prescribes that the sodium atom in its normal state should have no magnetic moment and no angular momentum. But now let us look at the data.

It would do no good in this connection to make measurements on solid or on liquid sodium, for in those "condensed phases" the atoms are crowded so closely together as to be badly distorted. We can, however, experiment on free atoms of sodium, in their normal state, in the way illustrated by Fig. 1. In the upper portion, A represents an "oven," consisting of a small box heated electrically and containing some liquid sodium which is steadily being vaporized. There is a hole in the wall of the box through which free sodium atoms are steadily shooting in all directions, with the distribution-in-speed which we know from the theory of thermal agitation; and beyond, there is a sequence of diaphragms with slits in them which delimit a straight and narrow beam of these fast-moving atoms. Disregarding what the theory has just said, let us suppose that each of these atoms is a magnet—a bar magnet, with a north pole and a south pole. As they emerge from the oven, these atoms must surely be oriented at random in all directions.

Continuing to look at the upper part of Fig. 1, we have two large magnet-poles, and the beam travels between them, having no trouble with molecules of air as it shoots along, for the whole of this apparatus is enclosed in a highly-evacuated tube. It may seem natural to visualize these magnet-poles as the two broad flat extremities of a horseshoe-magnet, with a uniform magnetic field pervading all the space between them. Such an arrangement, however, would make the experiment futile.

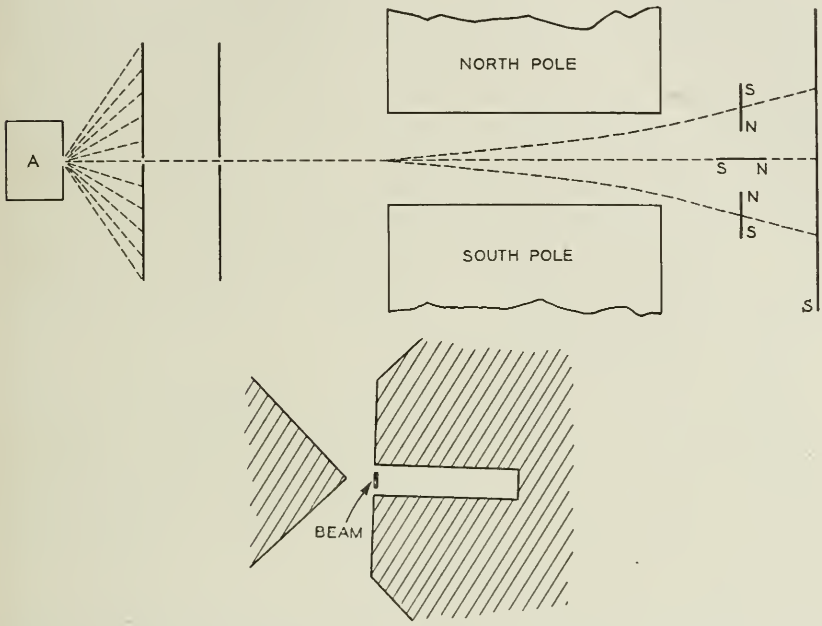


Fig. 1—Longitudinal section and cross-section of apparatus for the Gerlach-Stern experiment.

Nothing would happen to any of the atoms, for in the uniform field the north pole of each atomic magnet would be pressed downward just as hard as and no harder than the south pole is drawn upward, and the net force would be zero. The beam would go on unbroadened and undeflected, and make a small spot on the screen *S*, the spot being just opposite the slits in the diaphragms. Something else must be tried; and what we do—or rather, what Gerlach and Stern did in Hamburg some fourteen years ago—is, to shape one of the magnet-poles in the form of a wedge and hollow out the other, so that the field between the two shall no longer be uniform. The lower part of Fig. 1 represents the cross-section of such a pair of pole-pieces. The field-strength is now much greater near the wedge than near the opposite pole-piece, and there is a vertical gradient of field-strength which may be made fairly constant over most of the interspace. Rabi at Columbia achieves the same result more efficiently by peculiar arrangements of current-carrying wires instead of iron magnets.

Now consider (thinking classically!) a few of the atomic magnets as they shoot across this non-uniform field. Think of one which originally is oriented vertically, with north pole down and south pole up—the south pole will be in a stronger part of the field than its mate; it will be drawn upward harder than the north pole is pushed downward; the atom will sweep in a parabolic arc upward. Think of another which originally is oriented vertically with north pole up and south pole down—it will be swept in a parabolic arc downward. Think of another which originally has its axis pointing horizontally—it will shoot in a straight line across the field, as though the pole-pieces were not there.[2] Now think of all those which are oblique to the vertical; they will describe parabolic arcs of intermediate curvatures, upward or downward as the case may be. One infers that the beam must be spread out into a continuous fan, making a continuous vertical band upon the photographic plate.

Moreover, from the upper edge or from the lower edge of this continuous band, it should be possible to determine the magnetic moment μ of the atoms. For let us consider one of the vertically-oriented atoms, and call its pole-strength M and the length of its magnet r. The upward force on the north pole is MH,—H standing for the field-strength at the point where the north pole is. The downward force on the south pole is $M(H + dH/dz \cdot r)$. The net force is $Mr \cdot dH/dz$, which is $\mu(dH/dz)$, because Mr is the magnetic moment μ of the magnet by definition. The acceleration of the little atom is equal to this force divided by m_a, the mass of the sodium atom. The deflection is equal to half the acceleration by the square of the time during which the atom is exposed to the force. This time is equal to the distance D which the atoms traverse across the field, divided by v the speed which they have when they come out of the furnace. So finally, we have:

$$\text{Deflection} = \frac{1}{2}\frac{\mu(dH/dz)}{m_a}(D/v)^2.$$

We ascertain the deflection by looking at the end of the band on the photographic plate, and we can ascertain all the other things in the

[2] When thinking classically, we must not expect the atomic magnets to turn their axes toward the field-direction as they shoot across the field; the gyroscopic quality of these magnets, due to their angular momentum, inhibits this.

equation excepting μ,—v is the hardest to estimate accurately—and so we can solve the equation for μ.

When the experiment is done there appears, however, a very remarkable thing. Instead of there being a long band upon the plate, there are just two spots. Instead of the beam having been broadened out into a continuous fan, it has evidently been split into two separate pencils. It looks as though the field had acted first of all upon the magnets, by setting them all vertical,—half of them with north pole up, and half with north pole down. Not this phenomenon alone, but many others in Nature show us that this is just what happens. You may perhaps feel for the moment that it is intelligible, after all; the compass-needle turns to the north—why should not the little atomic magnets, as soon as they enter the field, turn their south poles toward the north pole of the magnet which attracts them? Well, this would not account for the magnets which constitute the beam which bends away from the wedge-shaped north pole, instead of toward it; and indeed it does not even account for the beam which bends toward the wedge-shaped pole. Classically the field should have no orienting effect whatsoever upon the atoms, and yet it evidently does.[2] This is one of the phenomena of the atomic world which we cannot properly visualize in terms of the behavior of objects large enough to be tangible and visible. All that I can do is to assert it, and to say that it justifies us in using this formula to calculate μ. When we use it, the value which we find for the magnetic moment of the sodium atom in its normal state is

$$eh/4\pi mc,$$

which happens to be one of the values in the sequence which I just wrote down.

I repeat that according to the theory in its present stage, the electron-orbits in the normal sodium atom have a net magnetic moment of zero. This value $eh/4\pi mc$ is, therefore, the magnetic moment due to the spin of the valence-electron—it is the magnetic moment of the spinning electron. I write it in the appropriate place, and then with the aid of the g-value derived from the gyromagnetic effect I write down the value of angular momentum which we assign to the spinning electron:

$$p = \tfrac{1}{2}(h/2\pi).$$

To this roster of three statements about the spinning electron I now make a final addition. The Gerlach-Stern experiment on sodium shows that a beam of sodium atoms—which for this purpose is the equivalent of a beam of spinning electrons—is divided into two by a magnetic field. I write down "2" to indicate this number of separated beams; but I will call it by preference the "number of orientations in the field," because that is the fundamental point. The spinning electron always sets itself in one or the other of two orientations, with respect to whatever field it happens to be traversing. We call them the "parallel" and the "anti-parallel" orientations, though according to quantum mechanics these terms are a little too strong. Here then is the list of the properties of the electron-spin: g equal to 2—angular momentum equal to $\frac{1}{2}(h/2\pi)$—magnetic moment equal to $eh/4\pi mc$—two permitted orientations in any field.

It has doubtless struck you as rather odd that I began by talking about the angular momenta and the magnetic moments of electron-orbits, and then carefully picked out a couple of special cases in which these neutralized each other altogether and there was nothing left over except what I ascribed to the electron-spin. Is there no point at all, then, in talking about the electron-orbits? Oh, very much so! Indeed there are cases in which the electron-spins neutralize each other altogether, and we have nothing left over except what is attributed to the orbits. To do this I may choose an atom like magnesium, which has a nuclear charge of $+ 12e$, a cloud of ten inner electrons which neutralize one another completely as to angular momentum and magnetic moment (just as in sodium), and *two* valence electrons instead of one. In some of the states of the magnesium atom—not in all of its states, but in *some* of them—the spins of the two valence electrons are oriented opposite to each other in the atom, and cancel each other out. When the atom is in a state of this kind, then nothing is left over except the angular momenta and the magnetic moments of the orbits of the two valence-electrons; and then, all the statements of the orbital theory (page 326) are applicable—g is equal to unity, the angular momentum takes one of the values $nh/2\pi$ and the magnetic moment takes one of the values $n(eh/4\pi mc)$. Moreover, there is another theorem derived from quantum mechanics which turns out to be valid: the number of orientations of such an atom in a field, the number of separated beams which appear in the Gerlach-Stern experiments, is chosen from among the members of this sequence: 1, 3, 5, 7. . . . (It is a most interesting historical fact, that Gerlach and Stern were moved to undertake their difficult experiment by the wish to test this remarkable assertion of quantal theory.) You notice that the number 2 does not appear in the sequence; were it not for the electron-spin, we never could obtain it; it is distinctive of the spinning electron.

I must, however, admit that all these cases of which I have been speaking are special, and comparatively rare. Both the cases in which the spins neutralize each other perfectly, and the cases in which the orbital moments neutralize each other perfectly—both types are unusual. Still more unusual, and yet occurring here and there, is the most special of all cases—that in which *all* of the moments and *all* the momenta, both those of the spins and those of the orbits, balance one another perfectly so that the sums are zero. An atom in such a state is completely devoid both of magnetism and of spin; such atoms are those of helium, of neon, of argon and the other noble gases, when in their normal states. Usually, however, we find ourselves confronted with some example of the general case, in which neither the spins nor the orbits are completely neutralized. The atom has an angular momentum which is a sort of composite or resultant of the angular momenta of the spins and the orbits, and it has a magnetic moment which also is a sort of composite or resultant.

If I were to embark on the description of the general case this lecture might go on interminably, and at its end you would probably not remember anything except what you had already known at its beginning. The laws of the composition of spins and orbits are so foreign to our customary ways of thinking, and the formulæ which express them are so curiously built, that to work once only through them is not sufficient: one has to memorize the derivations and the results alike, and go over them incessantly until they are imprinted on the brain. I think you will agree to this readily enough, when I remind you that this theory is none other than the general theory of spectra; for even quite outside the ranks of physicists, the theory of spectra is beginning to be notorious for its complexity. I shall not venture even to give the formulæ, much less their derivations; it must suffice to fill out the two lists of p-values and μ-values on which I have already begun, and the list of n-values or numbers-of-permitted-orientations.

The spinning atom is a congeries of electrons, all of them always possessing spin, most of them usually possessing orbital motions; and these motions are compounded with each other in such ways, that:

First, the angular momentum of the spinning atom has one of the values,

$$p = (0, \tfrac{1}{2}, 1, \tfrac{3}{2}, 2, \tfrac{5}{2}, 3 \cdots)h/2\pi.$$

Second, the number of beams in the Gerlach-Stern experiment, or the number of permitted orientations of the atom in the field, has one of the values,

$$n = 1, 2, 3, 4, \cdots.$$

Third—and now comes a surprise, for you will probably expect me to say of the magnetic moment that it is $eh/4\pi mc$ multiplied either by an integer or a half-integer; but this is not so. The actual state of affairs is described by a formula which is called the g-formula, because it gives g in terms of the moments, both spin and orbit, of the individual electrons. It was discovered by Landé and interpreted in terms of the spinning electron by Goudsmit. The g-formula gives unity, as of course it must, in the special cases where the electron-spins cancel each other and only the orbital moments are left over, and it gives 2 in the special cases where the orbital moments are neutralized with only the spins left over. In the other cases it may give any one of a large variety of values: mostly one gets simple-looking fractions such as 9/8 and 4/3 and 5/6. The magnetic moments are then computed by multiplying the appropriate g-values times $e/2mc$, into the existing values of angular momentum.[3]

I now turn to that component of the atom of which the spin remains to be discussed—to the *nucleus*.

As I have already intimated, the values of p and n and μ for the electron-family of any atom are mostly ascertained by analyzing their spectra and utilizing the great general theory of spectra. Such magnetic experiments as I used for my examples are relatively few, and feasible for relatively few substances. The reason for making this remark at this late moment is, that by analyzing spectra we may also learn something about the spins of nuclei; for nuclei also are invested with these properties of angular momentum and of magnetic moment. I take in particular the case of the *proton*—that lightest of all nuclei, the nucleus of the lightest known kind of atom which is ordinary hydrogen, so called to distinguish it from "heavy" hydrogen. Analysis of the spectrum of hydrogen shows us that the proton is capable of taking two permitted orientations in a field; thus, our first piece of information about the proton-spin is conveyed by writing $n = 2$.

Now that we have this piece of information, we deduce that as the spinning electron has an n-value of 2 and an angular momentum of $\frac{1}{2}(h/2\pi)$, so the proton with its n-value of 2 must have an angular momentum of $\frac{1}{2}(h/2\pi)$. Continuing along this line of thought, we are further tempted to infer that the proton should have a g-value of 2 and a magnetic moment of $(eh/4\pi mc)$. But what shall we suppose

[3] Should any reader intend to proceed from this article to a thorough study of atomic theory, he should be warned in advance that according to the latest form of quantal theory, the p-values and the μ-values here given are the values of the projections of these vectors upon the field-direction, the magnitudes of the vectors themselves being somewhat greater: I have given details as to this in "Contemporary Advances in Physics, XXIX, . . ." This *Journal*, **14**, pp. 293 ff.(1935).

about m? Formerly it represented the mass of the electron; now we are dealing with a different sort of particle, having a mass which (as many other kinds of experiments show us) is about 1835 times as great as the electron-mass. I denote this mass by M. It seems natural, then, to expect for the magnetic moment of the proton the value $eh/4\pi Mc$, or about 1/1835 of that of the spinning electron.

This is a formidably small magnetic moment to hope to measure, nay even to detect! yet Stern and his pupils undertook to measure it, and they succeeded. Of course, modifications had to be made in the technique which worked so well for sodium. Hydrogen being a gas at room-temperature, no heated oven was required; nevertheless they used an "oven," but it was refrigerated instead of being heated—a sort of super-ice-box; this was in order to obtain slow-moving atoms, for the slower the atoms traverse the field, the more accurately the experiment can be made. I just said "atoms"; but as most people know, the particles of gaseous hydrogen are not atoms, but diatomic molecules—systems composed of two protons and two electrons apiece. This is a circumstance which in many desirable tests of modern theoretical physics is a great inconvenience, for usually our simplest theoretical affirmations refer to hydrogen atoms and we should like to be able to experiment on them directly. Here, however, it turns out to be a great convenience, indeed perhaps the only thing that makes the experiment possible. For if we had an isolated hydrogen atom, the magnetic moment of its electron would so far exceed that of its proton that the latter would be indetectable. (Perhaps it is not superfluous to mention that bare protons could not be used in the experiment either, as the magnetic field would exert so large a force upon their moving charges that the forces upon their magnetic poles would be insignificant by comparison.)

But if in a single hydrogen atom the magnetic moment of the electron swamps that of the proton, how shall this fate be avoided for a system composed of two electrons and two protons? Here enters in, and in a very important and significant way, that law of the permitted orientations. Just as a spinning particle of angular momentum $\frac{1}{2}(h/2\pi)$ can take only two permitted orientations in a field, so it can take only two with respect to another particle of its kind—the parallel and the anti-parallel. It chances—or rather it does not chance, it follows from the underlying laws of Nature—that in the hydrogen molecule the two electrons are oriented anti-parallel to each other. Their magnetic moments cancel each other, and do not trouble the experimenter.

May not, however, the same thing happen in respect to the two protons, so rendering the experiment hopeless? It turns out that for these the spins are anti-parallel in some molecules but parallel in others. Molecules of the former type, which is called para-hydrogen, are indeed useless for the experiment; but molecules of the latter type, which is called ortho-hydrogen, are available, and in them the magnetic moments of the two protons collaborate so that the magnetic moment of the molecule-as-a-whole is twice as great as that of the single proton, a welcome assistance! In ordinary gaseous hydrogen at room temperature, about three-quarters of the molecules are ortho-hydrogen.

When the experiment was at last achieved by the school of Stern, it was found that the foregoing inference as to the μ-value of the proton is roughly but not exactly correct! The latest information is, that the magnetic moment of the proton is close to $2\frac{1}{2}$ times $eh/4\pi Mc$. Measurements with another method by Rabi and his school have confirmed these results; and we are definitively debarred from believing that for the proton and the electron, the magnetic moments stand in the inverse ratio of the masses. Perhaps this signifies that the proton is itself a composite particle, a notion for which there is some support from other sources.

I mention briefly the characteristics of a few other nuclei. After the proton, the next simplest is the deuteron or nucleus of the heavy-hydrogen atom. It is composed of a proton and a neutron, the latter being a neutral particle of about the same mass as the proton. We can observe free neutrons wandering about in space, but we cannot determine their spins nor their magnetic moments. The deuteron, however, has three permitted orientations (this we discern from the spectrum of heavy hydrogen) and consequently an angular momentum of $(h/2\pi)$. It is inferred that the neutron has $\frac{1}{2}(h/2\pi)$ for its angular momentum, and that in the deuteron these two constituent particles —proton and neutron—are oriented with their equal spins parallel to one another. The magnetic moment of the deuteron is less than that of the proton, and accordingly the neutron must have its magnetic moment oppositely directed to that of its companion in the system, even though their angular momenta be similarly directed—a strange complication!

After the deuteron, the next simplest among the nuclei (except two which are much too rare for investigation) is the alpha-particle or helium nucleus. It is composed of two neutrons and two protons. We find that its angular momentum and its magnetic moment are zero, a clear indication that the four spins of its components are cancelling each other two by two, and the four magnetic moments likewise. Next comes a nucleus composed of three protons and three neutrons, belonging to the element lithium. It is found that in angular momentum as in magnetic moment it is practically a duplicate of the deuteron, its six components having disposed themselves into a nearly normal deuteron attached to a nearly normal alpha-particle. I might proceed some distance farther along the list of the known nuclei after this fashion, were there space; but it is best to close this section with a general rule: *nuclei with an even number of constituent particles* (protons and neutrons) *have even spins, nuclei with an odd number of particles have odd spins.* "Even" and "odd" in this formulation mean that the angular momentum is an even or an odd integer multiple of $\frac{1}{2}(h/2\pi)$, respectively. One sees that if any two spins of magnitude $\frac{1}{2}(h/2\pi)$ are allowed to choose only between parallel and anti-parallel orientations, the rule follows inevitably; reversely, from the rule (which is based on experience with some fifty or sixty kinds of atom), we derive extra strength for that theorem about orientations.

Now to come to the conclusion and the climax. Although this property of angular momentum, of being allowed to take only a limited number of permitted orientations—although this strange and wonderful property of angular momentum was introduced in this lecture as though it pertained only to atoms subjected to applied external fields, yet it manifests itself far more broadly. Indeed, it manifests itself universally, and the stability and the character of the world are due to it. I have already mentioned in half-a-dozen places how it manifests itself within the molecule and within the atom: how in the atom, it is responsible for those laws of composition which determine the angular momentum and the magnetic moment of the electron-family—how in the hydrogen molecule, it establishes a difference between ortho-hydrogen and para-hydrogen—how in the nucleus it fixes the angular momenta and the magnetic moments of composite nuclei as sharply as those of their constituents the proton and the neutron. Perhaps these seem to be remote and unimportant qualities; but what has just been said of them may be said with equal truth and equal force of *all* the chemical and physical properties of all the elements, mass alone excepted (and even mass not fully excepted). If this feature of angular momentum did not prevail, there could not be the fixity of properties which characterizes each element by itself and the variety of properties which characterizes the totality of the elements. Gold would not be gold, lead would not be lead, oxygen would not be oxygen, helium would not be helium; for though it is commonly said that each element is distinguished by its nuclear charge and the number of electrons in its electron-family, this is not adequate. It is the law governing angular momentum which imprints upon these electron-families the characters which we recognize as the properties distinctive of the elements. It seems strange indeed that character should depend upon motion, and fixity upon the laws of whirling things; but however strange it may seem, there is no doubt about it. In the construction of houses the builder requires raw material in the form of brick and stone and wood and steel; but he requires also principles of architecture, whereby the raw materials may be parceled off and integrated into the general design. In the construction of the physical world, mass and charge fulfil the role of raw materials, and the laws of angular momentum furnish the principles of the architecture thereof.

A Multiple Unit Steerable Antenna for Short-Wave Reception *

By H. T. FRIIS and C. B. FELDMAN

This paper discusses a receiving system employing sharp vertical-plane directivity, capable of being steered to meet the varying angles at which short radio waves arrive at a receiving location. The system is the culmination of some four years effort to determine the degree to which receiving antenna directivity may be carried to increase the reliability of short-wave transatlantic telephone circuits. The system consists of an end-on array of antennas, of fixed directivity, whose outputs are combined in phase for the desired angle. The antenna outputs are conducted over coaxial transmission lines to the receiving building where the phasing is accomplished by means of rotatable phase shifters operating at intermediate frequency. These phase shifters, one for each antenna, are geared together, and the favored direction in the vertical plane may be steered by rotating the assembly. Several sets of these phase shifters are paralleled, each set constituting a separately steerable branch. One of these branches serves as an exploring or monitoring circuit for determining the angles at which waves are arriving. The remaining branches may then be set to receive at these angles. The several receiving branches have common automatic gain control and thus provide a diversity on an angle basis. To obtain the full benefit of the angular resolution afforded by the sharp directivity, the different transmission times, corresponding to the different angles, are equalized by audio delay networks, before combining in the final output.

The experimental system, located at the Bell Telephone Laboratories' field laboratory near Holmdel, New Jersey, is described. This system comprises six rhombic antennas extending three quarters of a mile along the direction to England. Two receiving branches, in addition to a monitoring branch, are provided. Experience obtained with this system since the spring of 1935 is discussed. The benefits ascribable to it are (1) a signal-to-noise improvement of seven to eight decibels, referred to one of the six antennas alone, and (2) a substantial quality improvement due jointly to the diversity action and the reduction of selective fading.

While a three-quarter-mile short-wave antenna system is an unusually long one, the steerability feature permits the employment of considerably more directivity, afforded by further increasing the length. A system two miles long is believed to be practicable and desirable. It could be expected to perform more consistently better than the three-quarter-mile trial installation, and should yield a signal-to-noise improvement of twelve to thirteen decibels

* Presented before Silver Anniversary Convention of the Institute of Radio Engineers, New York City, May 10, 1937. Published in *Proc. I. R. E.*, July, 1937.

referred to one rhombic antenna. With the object of predicting the performance of larger systems, the performance of the experimental system is examined in great detail and compared with theory.

I. Introduction

FOR more than a decade, point-to-point short-wave radio services have employed directional antennas both in transmitting and receiving. Transmitting antenna directivity results in increased field intensity at the receiving location and receiving antenna directivity discriminates against noise. Both directivities improve the signal-to-noise ratio of a given circuit and permit operation under more adverse transmission conditions. Arrays of simple antennas as well as extensive configurations of long wires have been used to produce these directivities in both the vertical and the horizontal planes.

Antennas in present use on the longer circuits, such as the New York-London telephone facilities, represent about the limit of fixed directivity. Further increase or "sharpening" of the directivity would seriously encroach upon the angular range of directions which are effective in the propagation of waves from transmitter to receiver. The vertical angle range useful in transmitting and receiving short waves is considerable. The horizontal range is appreciable although considerably less than the vertical range. To confine the principal antenna response to only a portion of these ranges penalizes the circuit when that portion is ineffective.

Much experience and considerable statistical data have been obtained which determine this useful range of directions for the New York-London circuits, and antennas have been designed in conformity with these results. However, too much weight must not be given to statistical results which indicate, for instance, that ninety per cent of the time the effective angles are, say, in the range from ten to twenty degrees. For, if the remaining ten per cent includes much of the time that has been lost with existing facilities, an antenna designed for a ten- to twenty-degree response may really be of no value, or even detrimental as a means of extending the usefulness of the circuit. Owing to the great variability in conditions on the north Atlantic path and to the relatively small amount of significant data which has been accumulated during times when gain is most needed it might be detrimental to carry fixed directivity further than present practice has adopted.[1]

[1] One way of attacking the problem of obtaining increased antenna gain has been proposed by John Stone Stone in U. S. Patent 1,954,898. This patent relates to fixed antennas but has certain features, such as delay equalization, in common with the system to be described in this paper.

If, however, the directivity can be varied or "steered" to meet the various conditions imposed by nature, a new field is opened in which a new order of antenna sharpness and gain is possible. In addition to the gain in signal-to-noise ratio afforded by directivity, a reduction in selective fading is possible if the sharpness is increased to the point where a separation of differently delayed waves is achieved. As early as 1927, Edmond Bruce[2, 3] found remarkable reductions in short-wave fading by using a receiving antenna having an extremely sharp directional pattern. The successful employment of sharp directivity is, of course, predicated upon considerable stability of wave directions. The experiments reported by R. K. Potter[4] in 1930 suggested that short waves are propagated in a more or less orderly manner and that stable wave directions might exist. Later experiments,[5] made in co-operation with the British Post Office, using pulse transmission to resolve angles in time, gave confirming data and demonstrated clearly the physical facts upon which is based the system to be described in the present paper. These fundamental facts, outlined in the paper describing the experiments just mentioned, are recapitulated here because a clear understanding of their nature and significance is an essential introduction to the subject in hand. In the pulse tests it was found that:

"1. To the extent that we have been able to resolve the propagation into separate (vertical) angles, the separate angles are found not to be erratic; they vary slowly.

"2. There appears to be at least a qualitative relation between angle and delay; the greater the delay the greater the angle above the horizontal.

"The existence of the many waves of different delay, which is known to make fading selective with respect to frequency, greatly impairs the quality of a short-wave radio telephone circuit. . . . The experimental facts, tentatively established, that individual wave angles are fairly stable and that waves of different delay invariably possess different vertical angles, make this problem hold considerable promise.

"The simple antennas described . . . are suitable for angle determination because of their ability to reject a single wave but they are not

[2] E. Bruce, "Developments in Short-Wave Directive Antennas," *Proc. I. R. E.*, vol. 19, pp. 1406–1433, August, 1931; *Bell Sys. Tech. Jour.*, vol. 10, pp. 656–683, October, 1931.

[3] E. Bruce and A. C. Beck, "Experiments with Directivity Steering for Fading Reduction," *Proc. I. R. E.*, vol. 23, pp. 357–371, April, 1935; *Bell Sys. Tech. Jour.*, vol. 14, pp. 195–210, April, 1935.

[4] R. K. Potter, "Transmission Characteristics of a Short-Wave Telephone Circuit," *Proc. I. R. E.*, vol. 18, pp. 581–648, April, 1930.

[5] Friis, Feldman, and Sharpless, "The Determination of the Direction of Arrival of Short Radio Waves," *Proc. I. R. E.*, vol. 22, pp. 47–78, January, 1934.

in general suitable for quality improvement. For such studies it would be preferable to construct a more elaborate antenna whose directional pattern has a single major lobe which is steerable in the vertical plane. Such an antenna would aim to select a narrow range of angles in which occur waves of substantially the same delay."

The present paper describes a steerable antenna receiving system of the general character suggested by the above quotation, and which has been in experimental operation at the Holmdel, New Jersey, field laboratory of the Bell Telephone Laboratories for the past two years. Certain other important features are incorporated in the system, notably an arrangement whereby individual wave groups arriving at different vertical angles are received separately and, after separate delay equalization, combined, thereby incorporating a unique form of diversity. Another important feature possessed by the system is its frequency range which permits operation on all of the frequencies used in short-wave transatlantic services.

II. Principles of Steering Antenna Directivity

An old and elemental type of steering of receiving antenna directivity is found in direction finders. The steering of a directional lobe as distinguished from the steering of a null has been accomplished in recent years. Schelleng [6] reported a moderate degree of horizontal plane steering, accomplished by means of phase shifters. Jansky [7] has obtained horizontal steering by bodily rotating an entire broadside array. Bruce and Beck [3] obtained vertical steering by varying the shape of a rhombic antenna by means of ropes, and demonstrated the value of steering in the reduction of selective fading. The present authors [5] have employed rotatable phase shifters to steer the nulls in the directional patterns of two spaced antennas. In that work the value of the rapid adjustments possible with phase shifters was very apparent. In the linear end-on MUSA [8] system to be described rotatable phase shifters are again employed to steer the vertical response.[9]

In Fig. 1 is shown a schematic representation of a linear end-on array of N equally spaced unit antennas in free space. The antennas are indicated by the numbered points. For simplicity it is assumed, in the following preliminary analysis, that the antennas are spaced far

[6] J. C. Schelleng, "Some Problems in Short-Wave Telephone Transmission," *Proc. I. R. E.*, vol. 18, pp. 913–938, June, 1930.

[7] K. G. Jansky, "Directional Studies of Atmospherics at High Frequencies," *Proc. I. R. E.*, vol. 20, pp. 1920–1932, December, 1932.

[8] The word MUSA is coined from the initial letters of "multiple unit steerable antenna."

[9] U. S. Patent No. 2,041,600.

enough to be substantially isolated from each other. Choosing antenna No. 1 for reference and considering a plane wave arriving at an angle δ with the axis of the array, it is clear that the output of No. 2 will add in phase with that of No. 1 if the phase advance ϕ is made equal to $2\pi ac/v - 2\pi a \cos \delta$, where c = velocity of light and v = the phase velocity of the transmission lines. Similarly, the output of No. 3 will add to that of No. 1 and No. 2 if its phase is advanced 2ϕ, etc. If the spacing, $a\lambda$, is sufficient there will be other angles for which the N outputs add in phase; at intermediate angles the outputs interfere with the result that zeros and minor maxima occur. By properly designing the unit antenna the undesired maxima may be suppressed.

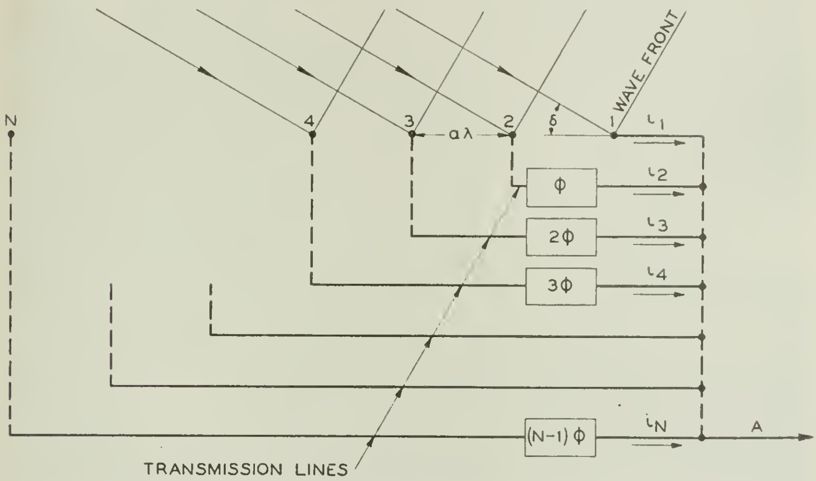


Fig. 1—A steerable antenna array using variable phase shifts ϕ, 2ϕ, 3ϕ, etc. The transmission lines indicated by broken lines are assumed to be of zero length; a is the spacing in free space wave-lengths.

In the Holmdel experimental system the unit antennas are of the rhombic type. An aerial view of the six antennas, which are located on the great circle through England, is shown in Fig. 2. These six antennas, combined as in Fig. 1, yield polar directional patterns such as those shown at the top of Fig. 3. The solid line pattern and the dashed line pattern correspond to different values of the phase shift ϕ. The multiple phase shifts of Fig. 1 are obtained by gearing the phase shifters to a common shaft which enables the directional pattern to be steered simply by rotating the shaft.

Thus far we have discussed the problem of sharp steerable directivity from the point of view of a single plane wave, whereas it is well known that multiple ionosphere reflections usually produce several more or less discrete waves, or bundles of waves, having different vertical angles and different transmission delays. To obtain the maximum advantage, however, requires that all of the several wave bundles be separately received and suitably combined after the transmission delays have been equalized. The achievement of this objective

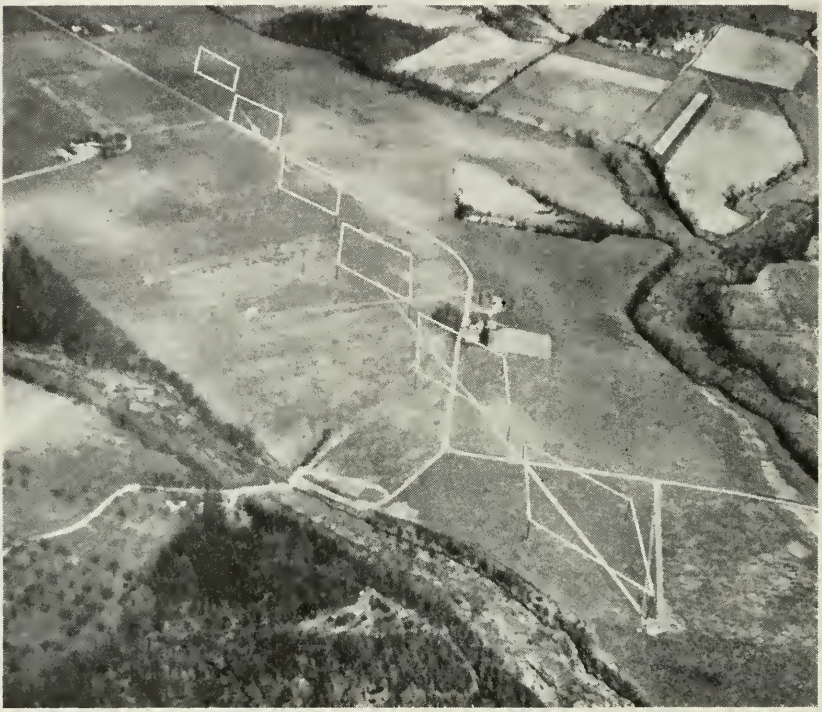


Fig. 2—Airplane view of the three-quarter-mile experimental MUSA on the receiving laboratory site located near Holmdel, New Jersey. The white line beneath the antennas is the newly filled trench in which coaxial transmission lines are buried. The building appearing in the right-hand foreground houses the receiving apparatus. The ground is flat to within $\pm$ 4 feet.

not only yields the ultimate gain in signal-to-noise ratio but at the same time *reduces the distortion associated with selective fading.*

The method of obtaining sharp steerable directivity by combining the output of fixed antennas through phase shifters makes it possible to use the same antennas and transmission lines to provide several separately steerable lobes each of which is in effect an independent MUSA.[10] In the experimental system, shown schematically in Fig. 3,

[10] R. K. Potter, U. S. Patent No. 2,030,181.

the antenna outputs are combined at intermediate frequency, and the separately steerable lobes are obtained by branching each of the six first detectors into three phase shifters and combining the outputs of

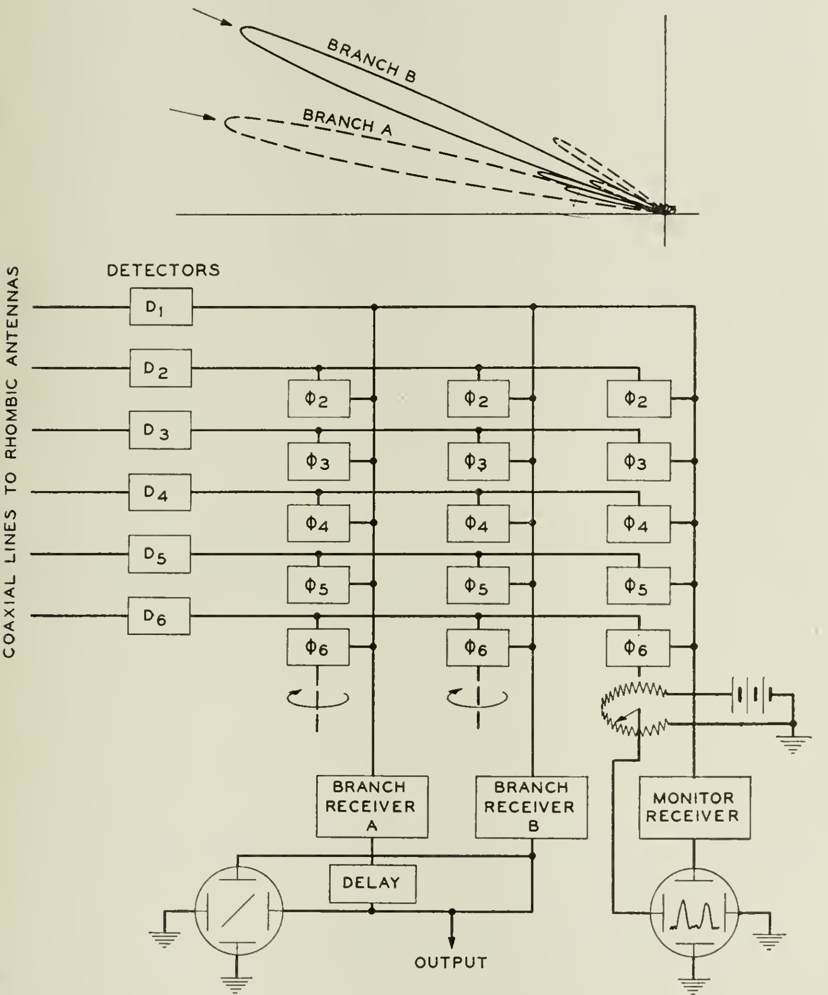


Fig. 3—Schematic diagram of the experimental MUSA receiver. The five phase shifters ϕ_2, ϕ_3, etc., of each branch, are geared to a shaft to provide the phase shifts ϕ, 2ϕ, 3ϕ, etc., of Fig. 1. The inset at the top shows the directional patterns of the two branches when steered at angles of 12 and 23 degrees, at a wave length of 25 meters.

the phase shifters to form three steerable branches. One branch is used continuously to explore the angle range to determine at which angles the waves are arriving. The other two branches are set accordingly and their outputs are "received" by conventional receivers, with common automatic gain control. The demodulated audio outputs are equalized for difference in transmission time and then combined. A cathode-ray oscilloscope displays the output of the exploring or monitoring branch. It plots amplitude (provided by a linear rectifier) as the ordinate, against phase shift ϕ_2 (corresponding to ϕ in Fig. 1). The screen of the oscilloscope is of the retentive type and thus displays several consecutive sweeps at once. A pattern corresponding to two waves is illustrated. The other cathode-ray oscilloscope is used in the adjustment which equalizes the delay of the two waves. Delay is added to the low angle branch until the oscilloscope shows a line (or compact elongated figure) which oscillates between the two axes as the two waves fade differently. This means that all of the audio frequencies of one branch are combining in phase with those of the other.

The above brief description was introduced to acquaint the reader with the essentially simple features of the MUSA system. Before describing the details and the results obtained with the experimental system, a more comprehensive analysis of steering principles will be given.

Returning to Fig. 1, it is assumed, of course, that the transmission lines are terminated in their characteristic impedance at the receiving terminal [11] (the phase shifters of Fig. 1) so that the phase is distributed linearly along the lines. Neglecting line loss (or equalizing it), the N currents, equal in magnitude and different in phase, are

$$\left.\begin{aligned} i_1 &= I\epsilon^{j\omega t} \\ i_2 &= I\epsilon^{j\{\omega t+\phi-2\pi a(v-\cos\delta)\}} \\ i_3 &= I\epsilon^{j\{\omega t+2[\phi-2\pi a(v-\cos\delta)]\}} \\ &\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot \\ i_N &= I\epsilon^{j\{\omega t+(N-1)\{\phi-2\pi a(v-\cos\delta)\}\}} \end{aligned}\right\} \qquad (1)$$

where i = instantaneous current in exponential notation
ω = angular frequency
N = total number of unit antennas
a = spacing in free space wave-lengths
$v = c/v$ = the ratio of the velocity of light to that of the transmission line.

The sum of the N currents is

$$A = I\epsilon^{j\omega t}\{1 + \epsilon^{j[\phi-2\pi a(v-\cos\delta]} + \cdots + \epsilon^{j(N-1)\{\phi-2\pi a(v-\cos\delta)\}}\}. \qquad (2)$$

[11] Non-characteristic terminations at the receiving ends of the lines are permissible if all terminations are identical and if the antennas are matched to the characteristic line impedance. Conversely, characteristic terminations at the receiving ends suffice if the antenna impedances are merely identical.

This exponential series may be evaluated with the aid of the identity [12]

$$1 + \epsilon^{j\theta} + \epsilon^{j2\theta} + \cdots + \epsilon^{j(n-1)\theta} \equiv \frac{\sin\dfrac{n\theta}{2}}{\sin\dfrac{\theta}{2}}\,\epsilon^{j\dfrac{(n-1)\theta}{2}}.$$

Using this summation we have

$$A = I\frac{\sin\dfrac{N}{2}[\phi - 2\pi a(v - \cos\delta)]}{\sin\dfrac{1}{2}[\phi - 2\pi a(v - \cos\delta)]}\,\epsilon^{j\{\omega t+(N-1)/2[\phi-2\pi a(v-\cos\delta)]\}}. \qquad (3)$$

The amplitude of A in (3) is the array directional pattern or array factor. It is zero when the numerator alone is zero, i.e., when

$$\frac{1}{2}[\phi - 2\pi a(v - \cos\delta)] \neq 0, \ \pm\pi, \ \pm 2\pi \cdots$$

and simultaneously

$$\frac{N}{2}[\phi - 2\pi a(v - \cos\delta)] = 0, \ \pm\pi, \ \pm 2\pi \cdots.$$

It attains its maximum value of NI when the denominator and numerator are zero simultaneously, i.e., when

$$\frac{1}{2}[\phi - 2\pi a(v - \cos\delta)] = 0, \ \pm\pi, \ \pm 2\pi \cdots$$

and

$$\frac{N}{2}[\phi - 2\pi a(v - \cos\delta)] = 0, \ \pm\pi, \ \pm 2\pi \cdots.$$

Plots of (3) for ten unit antennas ($N = 10$) spaced five wave-lengths ($a = 5$) are shown in Fig. 4 for two arbitrary values of ϕ. The same array used at twice the frequency ($N = 10$, $a = 10$) has the directional patterns shown in Fig. 5. The abscissas are labeled earth angle although nothing has been said thus far concerning the disposition of the N antennas with respect to the earth. In order that the simple multiple phase shifts of Fig. 1 shall suffice to steer the array, reflection from the ground must affect the phase of all antenna outputs identically. This is assured by constructing the array over, and parallel to, a flat expanse of ground. Since the angle δ measures the direction of

[12] This identity may be deduced by substituting $\epsilon^{j\theta}$ for r in the well-known formula for the sum of a geometrical progression

$$1 + r + r^2 + r^3 + \cdots + r^{n-1} \equiv \frac{r^n - 1}{r - 1}.$$

the wave referred to the direction of the array axis the array factor represents a surface of revolution. Figures 4 and 5 show merely axial cross sections which, for a horizontal array, may be considered vertical plane patterns.

Equation (3), as well as Figs. 4 and 5, shows that the sharpness of the principal lobe depends upon the total length of the array in wave-

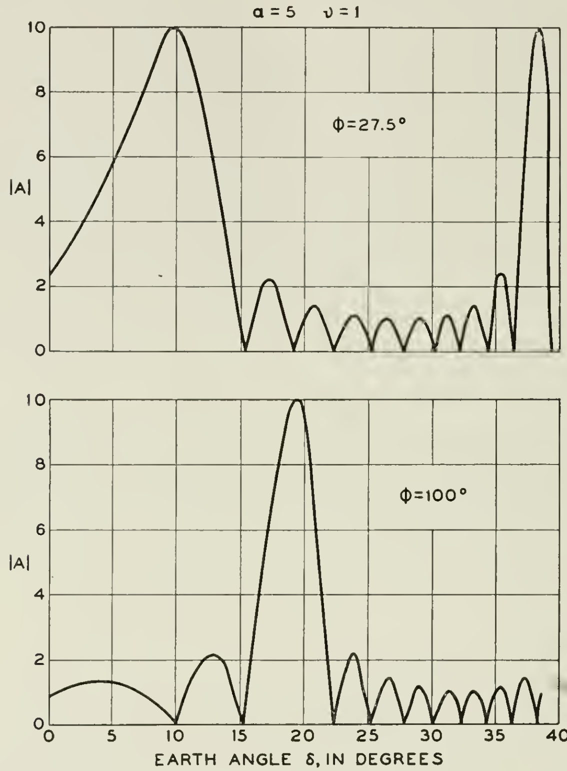


Fig. 4—Plots of the array factor for a 45-wave-length horizontal end-on array.

lengths, i.e., upon Na, while the angular spacing of adjacent principal lobes depends inversely upon the spacing "a." Thus, a single lobed pattern results if the array consists of a large number of closely spaced units.

A single lobed pattern is desirable, but to obtain it by using a large number of unit antennas [13] with separate transmission lines and phase

[13] The reader may observe that the reduction of the spacing would, if carried so far as to make "a" a fraction of a wave-length, violate the assumption that there is negligible reaction or coupling between unit antennas. As stated, this assumption is made in the interest of simplicity. It is theoretically possible to compensate for coupling between antennas so that (1), (2), and (3) still hold.

shifters would be a rather extensive undertaking. Provided a restricted range of steering is permissible, a simpler solution is to employ comparatively few large unit antennas and to let their directional pattern suppress the undesired principal lobes of the array pattern. Useful angles for transatlantic circuits are confined to the range from zero, or some low undetermined limit, to some higher limit. In what follows let δ_m represent an angle a little above the useful range so that a

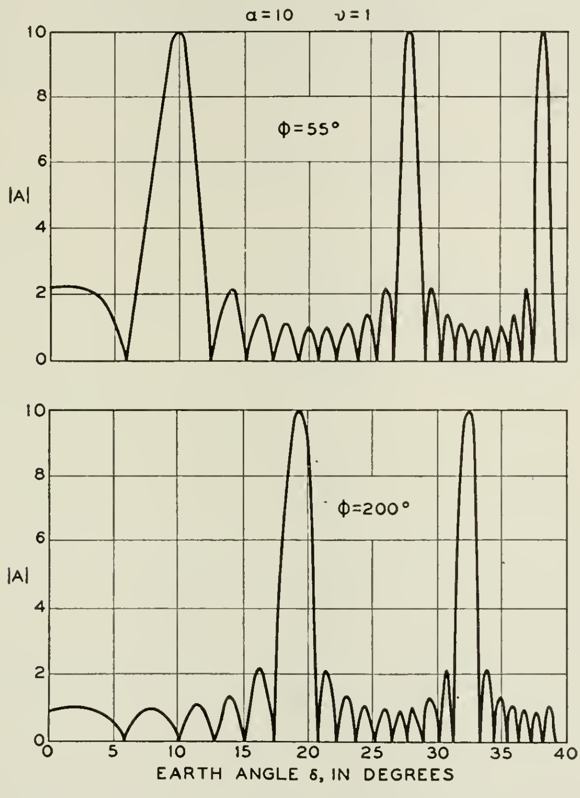


Fig. 5—Plots of the array factor for a 90-wave-length array; that of Fig. 4 used at twice the frequency.

null may be located at δ_m without imposing an excessive loss. The array may then be designed so that when the first principal lobe is steered at zero angle the second falls at δ_m or beyond. The question of whether the array design permits the construction of a suitable unit antenna in the length $a\lambda$ allotted to it is considered in the following paragraph. As a matter of fact, this analysis closely follows the actual steps in the development of the MUSA system.

Turning back to the ideal system comprising a very large number of closely spaced unit antennas, which yields the single lobed pattern, let us divide the antennas into N groups with n antennas in each group. Calling the group spacing "a" and the phase shift between adjacent antennas ϕ the application of (3) gives, dropping the exponential factor,

$$A = \frac{\sin \dfrac{nN}{2}\left[\phi - 2\pi \dfrac{a}{n}(v - \cos \delta)\right]}{\sin \dfrac{1}{2}\left[\phi - 2\pi \dfrac{a}{n}(v - \cos \delta)\right]}. \tag{4}$$

Multiplying numerator and denominator by

$$\sin \frac{n}{2}\left[\phi - 2\pi \frac{a}{n}(v - \cos \delta)\right]$$

results in

$$A' = \frac{\sin \dfrac{n}{2}\left[\phi - 2\pi \dfrac{a}{n}(v - \cos \delta)\right]}{\sin \dfrac{1}{2}\left[\phi - 2\pi \dfrac{a}{n}(v - \cos \delta)\right]} \times \frac{\sin \dfrac{N}{2}\left[n\phi - 2\pi a(v - \cos \delta)\right]}{\sin \dfrac{1}{2}\left[n\phi - 2\pi a(v - \cos \delta)\right]}. \tag{5}$$

Equation (5), which appears as the product of two array factors, is merely another way of writing the array factor for the large number (Nn) of unit antennas. The first factor represents an array of n "sub-unit" antennas of spacing a/n and phase shift ϕ. The second represents the array of these arrays with a spacing of $a\lambda$ and a phase shift $n\phi$. We now proceed to treat these two factors independently and assign the values ϕ_f and ϕ_v to replace ϕ and $n\phi$, respectively. Figure 6 depicts such an array of arrays. If now we regard the array of n sub-units as constituting a fixed unit antenna and adjust it to receive at zero angle (by putting $\phi_f = 2\pi a/n(v - 1)$) in accordance with the lower limit of the useful range, we obtain

$$A'' = \frac{\sin \dfrac{n}{2}\left[2\pi \dfrac{a}{n}(1 - \cos \delta)\right]}{\sin \dfrac{1}{2}\left[2\pi \dfrac{a}{n}(1 - \cos \delta)\right]} \times \frac{\sin \dfrac{N}{2}\left[\phi_v - 2\pi a(v - \cos \delta)\right]}{\sin \dfrac{1}{2}\left[\phi_v - 2\pi a(v - \cos \delta)\right]}. \tag{6}$$

The first factor in (6) represents the pattern of the unit antenna. It is a relatively broad single lobed pattern with maximum response at

$\delta = 0$. It drops to zero at $\delta = \cos^{-1}(1 - 1/a)$. For higher angles nothing but minor maxima occur since "a" is small and n is large. The second factor represents the steerable array pattern of the N unit antennas. With ϕ_v adjusted for maximum response at zero angle (this makes $\phi_v = n\phi_f$) this system is identical with the array of Nn subunits. In this case, all principal lobes of the array factor for the N units, excepting the first, coincide exactly with nulls of the array factor for the n subunits, and the familiar tapered distribution of minor maxima associated with the array of Nn subunits results. As ϕ_v is varied to steer for other angles than $\delta = 0$, the coincidence of nulls and undesired principal lobes no longer occurs. Since, however, the

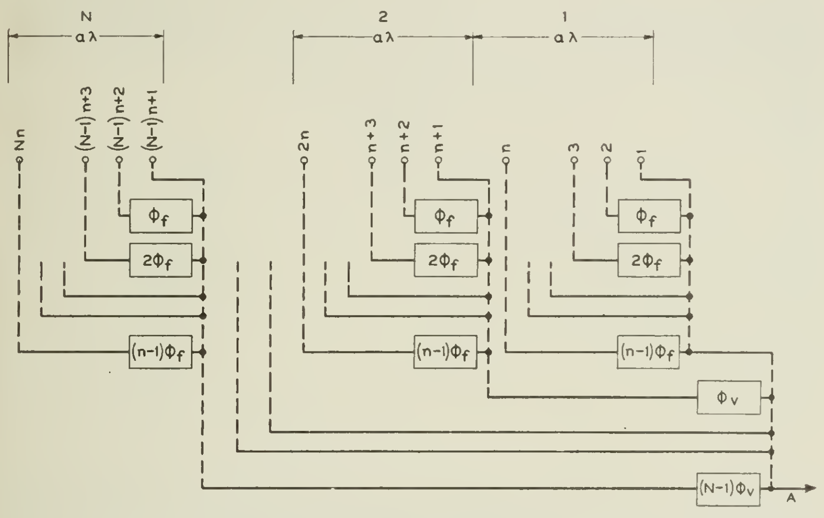


Fig. 6—A steerable array formed by dividing the antennas of Fig. 1 into N groups of n each. The subscripts "f" and "v" refer to fixed and variable phase shifts.

fixed unit antenna has only minor response beyond its first null, those undesired principal lobes are adequately suppressed, and the array may therefore be steered anywhere within the range from $\delta = 0$ to $\delta = \cos^{-1}(1 - 1/a)$, with single lobed response. As the principal lobe is steered away from $\delta = 0$ the maximum amplitude falls off in comparison with that of the array of Nn subunits. This represents a loss of signal-to-noise ratio and is to be regarded as a penalty for compromising to the extent of using fixed arrays as unit antennas. The loss is appreciable, however, only if the array is steered near the upper cutoff angle of the unit antenna. It remains but to select "a" so that $\cos^{-1}(1 - 1/a)$ represents the upper limit of the range, δ_m.

For a fixed physical spacing, "a" varies inversely with the wave-length, which results in an increasing steering range with increasing wave-length. Since the critical angle of reflection from the ionosphere increases with wave-length the upper limit of the range of useful angles can be expected to increase also. By selecting the proper spacing, the steering range and the critical angle can be made to agree satisfactorily. Figure 7 shows a plot of $\delta = \cos^{-1}(1 - 1/a)$ against wave-length for the unit antenna spacing of 200 meters which was

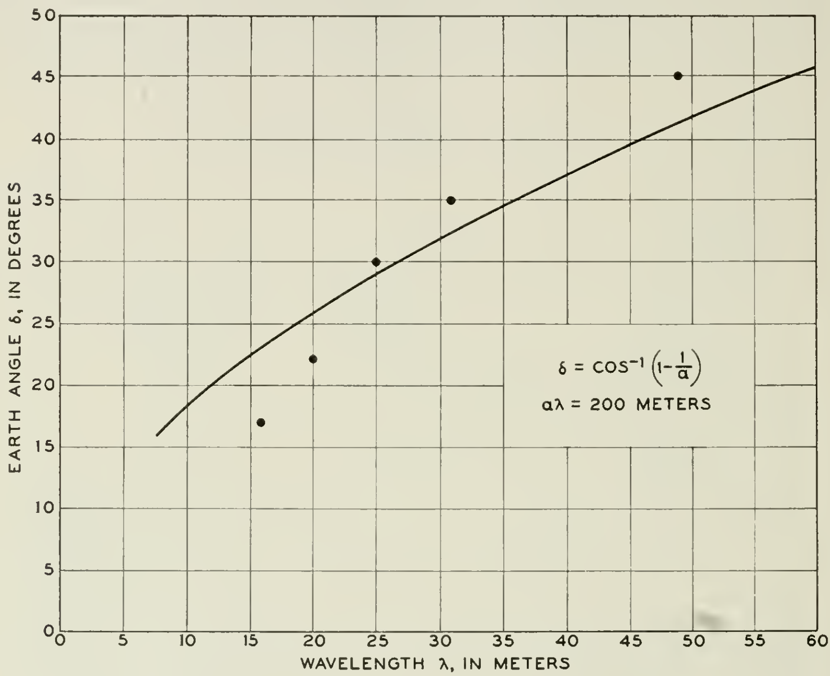


Fig. 7—Measured upper limits of vertical angles as a function of wave-length, compared with the upper limit of the array depicted in Fig. 6. The measured values represent the highest angles observed; usually stronger waves of lower angles predominate.

adopted for the experimental MUSA system to be described. The points denote upper limits of earth angles obtained from measurements made during the years 1933–1936 on signals from Rugby [5] and Daventry, England.

The foregoing analysis shows that

(1) A MUSA system may be so proportioned that the upper limit of its steering range follows, with fair accuracy, the upper limit of the range of useful angles, as the wave-length is varied.

(2) It is theoretically possible to construct a suitable unit antenna in the space provided for it when (1) is satisfied.

III. Description of the Experimental MUSA System

Antennas and Transmission Lines

Any type of unit antenna whose directional pattern suppresses the undesired principal lobes over the required wave-length range is basically suitable for use in a MUSA system. The rhombic antenna [14] does not fulfill this requirement as well as the linear array of subunits discussed in the preceding section. It was, however, selected on account of its advanced state of development. The manner in which it fits into the MUSA array factor will be discussed later.

The coupling or "cross talk" between antennas need not be of negligible magnitude in a MUSA system. For, to a first approximation,

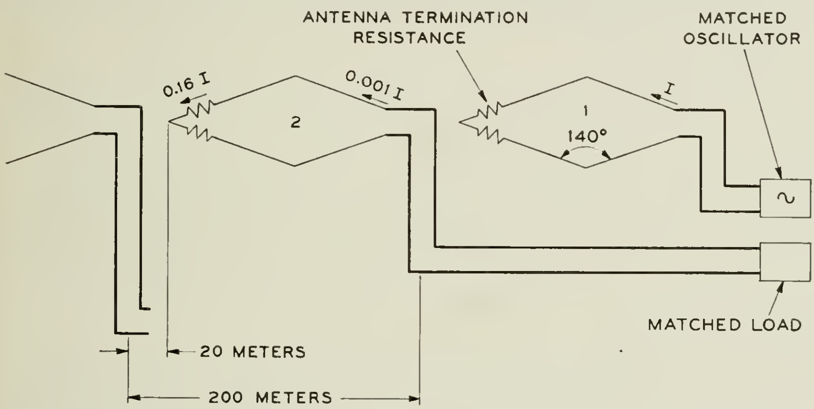


Fig. 8—Measurements of cross talk between adjacent antennas in the MUSA as made from the transmitting point of view.

the coupling is confined to adjacent antennas and is similar for all pairs so that only the *end* antennas could be expected to fail to combine properly with the others. At the ends, "dummy" antennas, not connected with the receiver but terminated like the others, could be erected to supply the coupling necessary to make all antennas alike. Measurements made on the experimental MUSA (Fig. 2) indicated that the cross talk is small enough to be neglected, however, so that dummy antennas ahead of or behind the six regular ones were considered unnecessary. The performance of the system in subsequent tests corroborates this conclusion.

The crosstalk measurements yielded the results indicated on Fig. 8. The small amount of crosstalk current ($0.001I$) measured at the trans-

[14] Bruce, Beck, and Lowry, "Horizontal Rhombic Antennas," *Proc. I. R. E.*, vol. 23, pp. 24–46, January, 1935; *Bell Sys. Tech. Jour.*, January, 1935.

mission line end of the forward antenna (No. 2) and the larger current ($0.16I$) at the other end reflect the fact that the rhombic antenna is "unidirectional." To a first approximation the current in such an aperiodic antenna accumulates progressively towards the output end. Therefore, the "effective" cross talk current is probably less than $(0.16I + 0.001I)/2 = 0.08I$; i.e., the effect upon the field radiated in the principal lobe will be altered by less than ten per cent due to the parasitic excitation of the antenna ahead. Antennas farther ahead as well as those behind contribute relatively nothing.

Since by the reciprocal theorem the directional pattern of any antenna is the same for transmitting and receiving, the crosstalk should likewise result in less than 10 per cent effect in the receiving case.

The measurements of Fig. 8 were made at 18 megacycles. At this frequency the rhombic antennas are proportioned to give maximum radiation approximately end-on. At lower frequencies the crosstalk is probably less.

The coaxial transmission lines are constructed of 60-foot lengths of one-inch copper plumbing pipe spliced with screw type plumbing unions. The inner conductor is one-fourth inch in diameter and is supported by isolantite insulators. The characteristic impedance of the lines is 78 ohms. The lines extend up the poles where they are connected to the antennas through balanced-to-unbalanced matching transformers.[14] At the receiving building the lines terminate on a special jack strip. Nitrogen pressure is maintained in all lines to exclude moisture.

In order to operate the MUSA system it is not essential that the velocity of the transmission lines be known. The velocity must be known accurately, however, in order to determine the angle of the waves as they are selected by the steerable lobe. Accordingly, the velocity was calculated (taking the insulators into account) and also measured. The calculated ratio of the line velocity to the velocity of light is 0.941; measurements yielded 0.933 ± 0.004. Using the value of 0.933, angles less than zero have occasionally been measured. A value of 0.937 would have made the lowest indicated angle just zero.

The longest line is about 1000 meters in length. Its impedance measured at one end when the other end is terminated by a resistance of 78 ohms shows some variation as the frequency is varied. In Fig. 9 are shown the results of impedance measurements made by substituting for the line an equivalent parallel combination of resistance and reactance. The two notable variations occurring at approximately 7.7 and 15.4 megacycles are believed to be caused by a slight irregularity at each joint, which adds a shunt capacitance of the order of 1.8 micromicrofarads. When spaced regularly at 60-foot intervals these capacitances have a somewhat cumulative effect at frequencies for which 60 feet (18.3 meters) is a multiple of the half wave-length. Sixty feet, when increased by the line velocity ratio, corresponds to 7.7 and 15.4 megacycles. Clearly, line sections which are not short compared with the shortest wave-length should be made unequal so that joint irregularities will not be harmful. The smaller variations of the order of $\pm$ 10 ohms may be due to random eccentricities produced by slight buckling of the inner conductor between insulators. With the possible

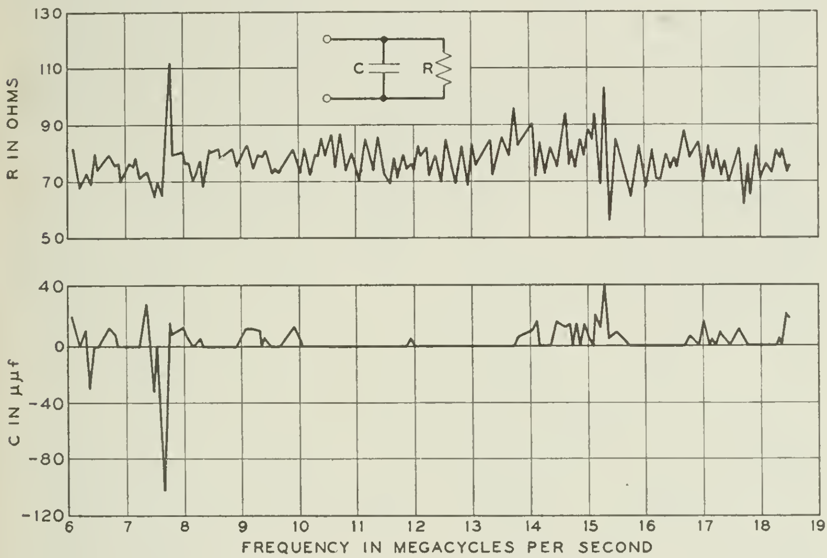


Fig. 9—Impedance measurements made upon the 1000-meter line terminated in a resistance of 78 ohms. The reactance is expressed as shunt capacitance, negative values meaning an inductive reactance numerically equal to the corresponding capacitive reactance.

exception of the two large variations this line is sufficiently smooth for use in a MUSA, as both theory and subsequent experience indicate.

Input Circuit and First Detectors

The MUSA system imposes requirements upon the input circuits and detectors which do not apply to conventional receivers. These requirements are as follows:

(1) The circuits must suppress standing waves on the transmission lines.[15]

[15] This requirement was more easily met than the alternate requirement mentioned in footnote (11).

(2) The phase shift from the transmission line to the phase shifter stage must be alike in all six circuits, independent of wave-length.

In order to simplify the experimental job it was decided to dispense with the selectivity afforded by high-frequency amplifiers and to use the simple circuits shown in Fig. 10. The capacitive coupling to the transmission line is a convenient means of matching the low-impedance lines to the high-impedance circuits. Plug-in coils (*L*) are used to cover the range from 4.5 to 22 megacycles.

The first detectors are of the two-tube balanced type which suppresses interference from two signals differing by the intermediate

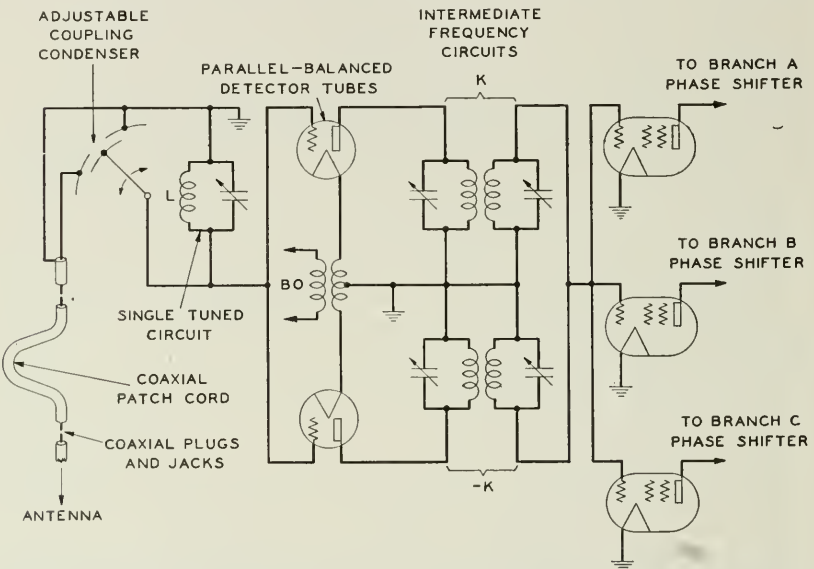


Fig. 10—Input circuit, first detector, and first intermediate-frequency tubes.

frequency and isolates the beating oscillator supply from the input circuits. The latter prevents crosstalk between the six inputs, and assures independence in the tuning of the input circuits. The beating oscillator voltage is introduced, at low impedance, between cathodes [16] by means of the distributing system of equal length coaxial lines shown in Fig. 11. This distributing system gives equiphase beating oscillator inputs to all detectors and makes requirement (2) attainable by having nominal similarity in the remaining parts of the six circuits.

Requirement (1) is met by feeding a test oscillator of 78 ohms impedance into the first circuit jack and adjusting the tuning condenser

[16] W. A. Harris, "Superheterodyne Frequency Conversion Systems," *Proc. I. R. E.*, vol. 22, pp. 279–294, April, 1935.

and the coupling condenser (Fig. 10) alternately until the maximum signal voltage appears on an indicating meter in one of the three intermediate-frequency branches. The three-terminal coupling condenser is an aid in this procedure since varying the coupling imposes only a slight variation in the capacitance across the coil. When the indicating instrument is a square-law vacuum tube voltmeter with the main

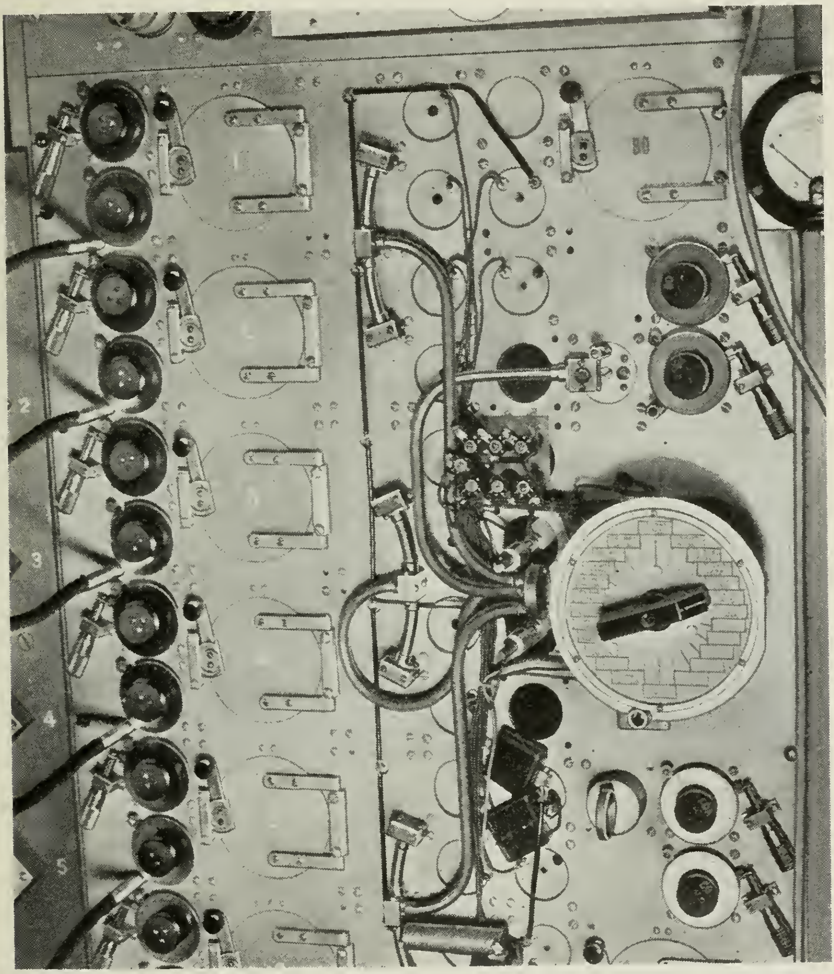


Fig. 11—Close-up view of high-frequency panel with cover removed. The beating oscillator supply line originates in the upper right-hand corner. It supplies the six detectors with equiphase and equiamplitude voltages. Plug-in coils fit into the compartments covered by the six circular doors. Micrometer heads which are used to adjust the six tuning condensers appear. The coaxial patch cords appear at the extreme left.

current balanced out and the remainder indicated by a 30-microampere meter, the sensitivity is more than sufficient to tune the circuits correctly.

The criterion of correct tune is the degree of suppression of standing waves on the transmission lines. To determine whether or not the maximizing adjustment insures an adequate standing wave suppression, a standing wave detector was incorporated in the experimental design. This is shown in Fig. 12. It consists of about 16 meters of 78-ohm coaxial line arranged in a coil and terminated by the

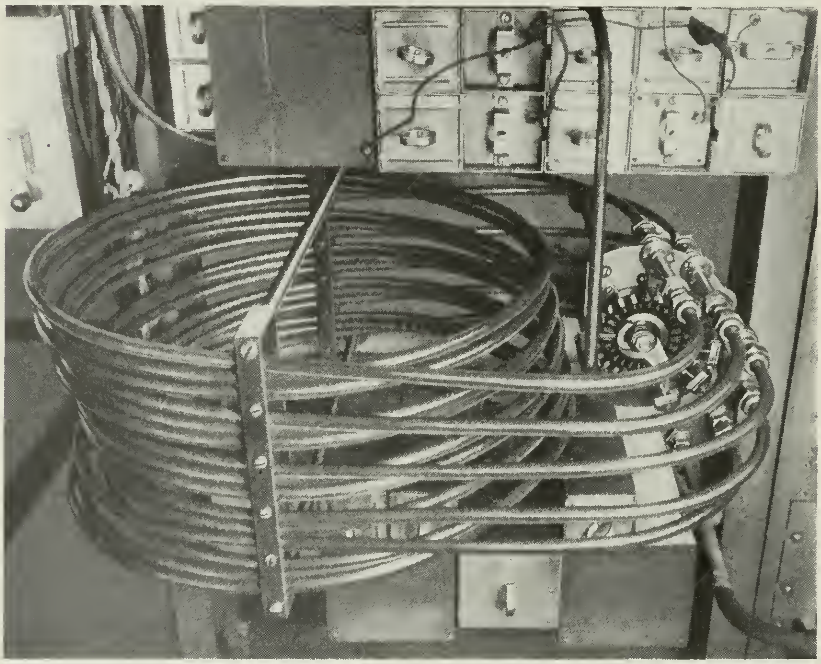


Fig. 12—The standing wave detector comprising 50 feet of 3/8-inch coaxial line, which may be used to test the correctness of the input circuit adjustment.

first circuit to be tested. It is fed at the other end by a test oscillator. Six capacitively coupled taps are brought to the low-capacitance switch shown in the photograph. The selector arm connects the taps to an auxiliary receiver with a high-input impedance. The absence of standing waves is shown by equal readings at the six positions. It was found that the maximizing adjustment results in a standing wave with less than ten per cent total variation, which represents nearly as much suppression as the smoothness of the line allows.

With nominally correct resistance termination standing waves of five per cent usually occur. For standing waves not exceeding ten per cent the accompanying phase distribution along the line does not depart more than a few degrees from the desired linear distribution. The use of the standing wave detector in routine operation was therefore not required.

Phase Shifters

Of the numerous methods of shifting phase the method [17] illustrated in Fig. 13 is the one chosen for the 18 circuits (3 branches, 6 antennas) of the experimental MUSA. Here points *a*, *b*, *d*, and *c* have voltages

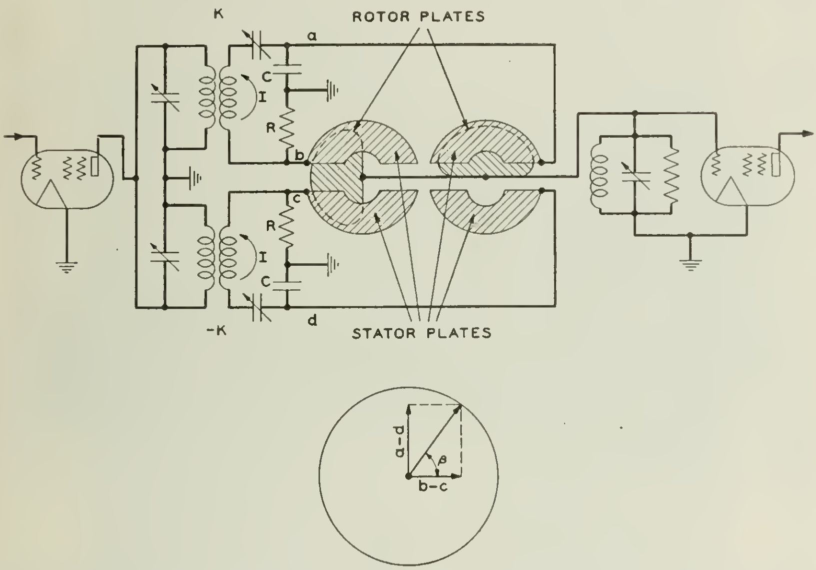


Fig. 13—Circuit diagram and vector diagram of the phase shifter. The rotor plates are especially designed to give a phase shift proportional to shaft angle.

to ground 90 degrees apart. The potential of point b is IR; that of c is $-IR$; that of a is $jI/\omega C$; that of d is $-jI/\omega C$. The resistance R and reactance $1/\omega C$ are made equal at the mid-band frequency so that four equal voltages, distributed equally over 360 degrees of phase, appear on the four stators of the special condenser. A photograph of this condenser appears in Fig. 14. Two specially shaped eccentric rotors mounted in quadrature to each other on the same shaft comprise the output terminal. It will be noted that voltages of opposite phase are connected to adjacent stators. Thus, with the rotors in

[17] L. A. Meacham, U. S. Patent No. 2,004,613.

the position shown dotted in Fig. 13 the output comes from point a since d is not coupled and b and c cancel each other. By shaping the two rotors so that the difference in exposure to opposite stator plates is proportional, respectively, to the sine and cosine of the angle of shaft rotation, the total current flowing from the two rotors will be constant and of phase proportional to the shaft angle. This is illustrated by the vector diagram in Fig. 13 in which β is the shaft angle and vectors $a - d$ and $b - c$ are the quadrature rotor outputs proportional to $\sin \beta$ and $\cos \beta$.

These phase shifters vary in output by less than $\pm$ 5 per cent as the shaft is rotated. The departure from linearity of phase shift is correspondingly small; i.e., less than $\pm$ 5 degrees.

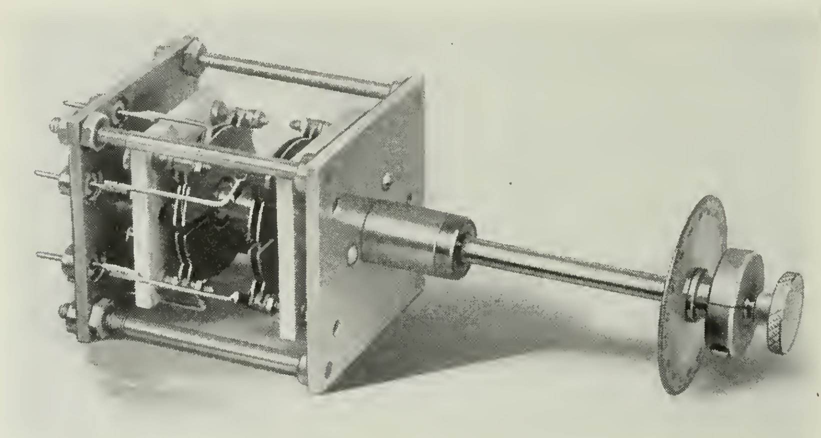


Fig. 14—The phase shifting condenser.

The useful band width of this type of phase shifter is fundamentally limited by the fact that $1/\omega C$ varies with frequency while R does not. However, this limitation does not appear in the Holmdel MUSA in which the percentage band width is small because the phase shifters operate at the intermediate frequency of 396 kilocycles.

The phase shifters are connected to the steering shaft with helical gears of multiple ratios as shown in Fig. 15. The phase shifter shafts may be slipped with respect to the main shaft. After they have been aligned so that locally supplied equiphase inputs to all detectors add in phase at the point where the phase shifter outputs are combined they are locked. This adjustment is independent of signal frequency. Provision is made for adjusting the gain of each of the six phase shifter circuits so that the differences in transmission-line loss may be compensated and any other desired amplitude adjustments made. The photograph of Fig. 15 shows the monitoring or exploring branch whose steering shaft is motor driven at one revolution per second.

Before leaving the subject of phase shifting it may be well to distinguish between phase shift and delay as here used. All electrical net-

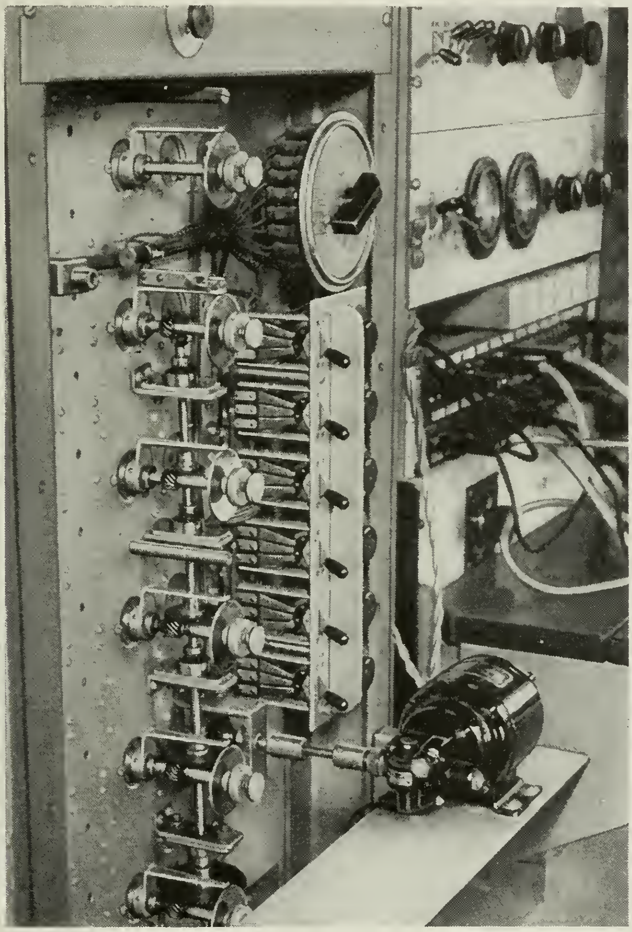


Fig. 15—Phase shifting panel of the monitoring branch. Only five of the six phase shifters are rotated for steering purposes. They are geared to the steering shaft in ratios of 1 : 1, 1 : 2, 1 : 3, 1 : 4, and 1 : 5.

works, except for certain highly distortive ones, possess a phase-frequency characteristic which is such that higher frequencies have their phases retarded with respect to lower frequencies. The ratio of the ncrement of phase retardation to the increment of frequency, i.e., the slope of the phase characteristic, is the delay. It is sometimes called the group delay or group transmission time as distinguished from the "phase time."[18] The delay is the only time which can be measured. It does not determine the phase shift of a particular frequency nor is it determined by the phase shift. A phase shifter applied to the network merely moves the phase curve intact up or down on the phase axis.

General Description of the System

The preceding paragraphs have described features which distinguish the MUSA system from conventional receiving systems. There remain to describe several auxiliary features and to present a unified picture of the whole.

The experimental system was designed for double side-band reception and all of the results reported in this paper refer to double side band. There has recently been completed equipment which may be substituted for the double side-band equipment for the reception of reduced carrier single side-band signals. The new equipment may also be used to select, with crystal filters, one side band of double side-band signals.

The delay to be inserted in the low-angle branch as indicated in Fig. 3 is obtained electrically from an audio-frequency delay network. The delay could theoretically be provided at the intermediate frequency but no advantage would result. The audio-frequency delay network is a special artificial line composed of forty sections and terminated by its characteristic impedance. Each section has a delay of 68 microseconds. A special switch is arranged to tap a high impedance output circuit across any desired section, thus providing a delay of 2.7 milliseconds variable in 0.068-millisecond steps. A special equalizing network[19] which makes the transmission loss the same for all steps and which also equalizes the frequency-loss characteristic so that the response is flat to 5000 cycles for all steps is automatically controlled by this switch. The forty delay sections appear in Fig. 16 just under the shelf on the right-hand bay. The maximum delay which has been required in actual operation is 2.5 milliseconds.

Both linear rectifiers and square-law detectors are provided for final demodulation and either may be switched into service as desired. The

[18] This distinction is brought out by J. C. Schelleng in a "Note on the Determination of the Ionization of the Upper Atmosphere," *Proc. I. R. E.*, vol. 16, pp. 1471–1476, November, 1928.

A general discussion of delay distortion (phase distortion) is to be found in three papers appearing in the *Bell Sys. Tech. Jour.*, vol. 9, July, 1930.

[19] This network and the delay sections were designed by P. H. Richardson of Bell Telephone Laboratories, Inc.

automatic gain control for use with either demodulator is obtained from linear rectifiers but a different diversity connection is made for each type of demodulator, in the interest of output volume constancy. A choice of time constants of 0.06, 0.5, and 4 seconds is provided.

Keys are provided, the ganged manipulation of which makes it possible, among other things, to compare (1) the MUSA output versus any one of the six antennas connected to one branch receiver, and (2) any pair of antennas in ordinary diversity using both branch receivers, versus one antenna using one receiver.

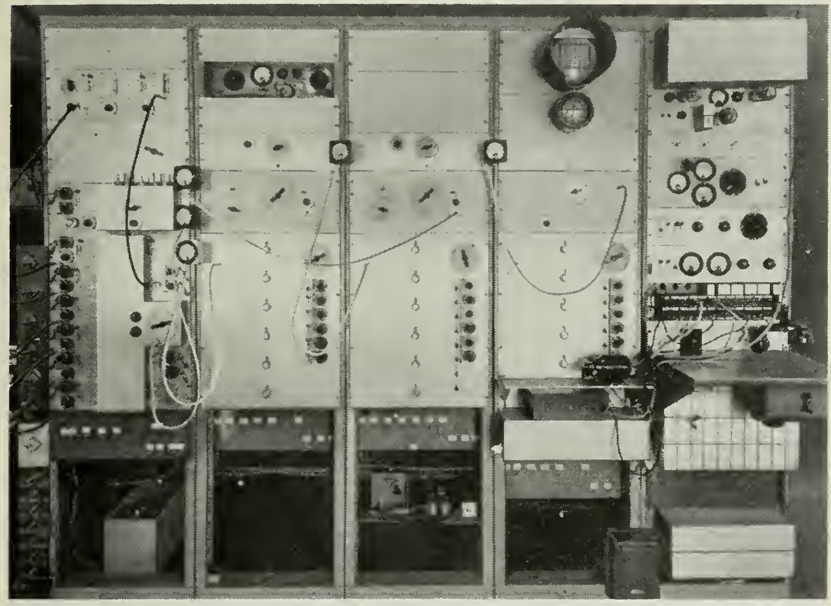


Fig. 16—Front view of the MUSA receiving equipment. The high-frequency bay is at the left and the audio-frequency bay at the right. The branch receivers are the panels directly above the phase shifting panels. The pulse receivers appear above these. At the top of the bay containing the monitoring branch equipment are the two oscilloscopes referred to in Fig. 3. The large tube with the ruled face is the monitoring oscilloscope.

In addition to the regular branch receivers with a 12-kilocycle band width and the monitoring branch receiver with a 2.5-kilocycle band width, two other receivers are provided in the experimental system. These receivers have a 30-kilocycle band width and are used for pulse reception. They are bridged across the inputs of the two regular branch receivers and are connected to a cathode-ray oscilloscope through a commutator.[5]

Various photographs of the MUSA receiver appear with explanatory captions in Figs. 16, 17, and 18.

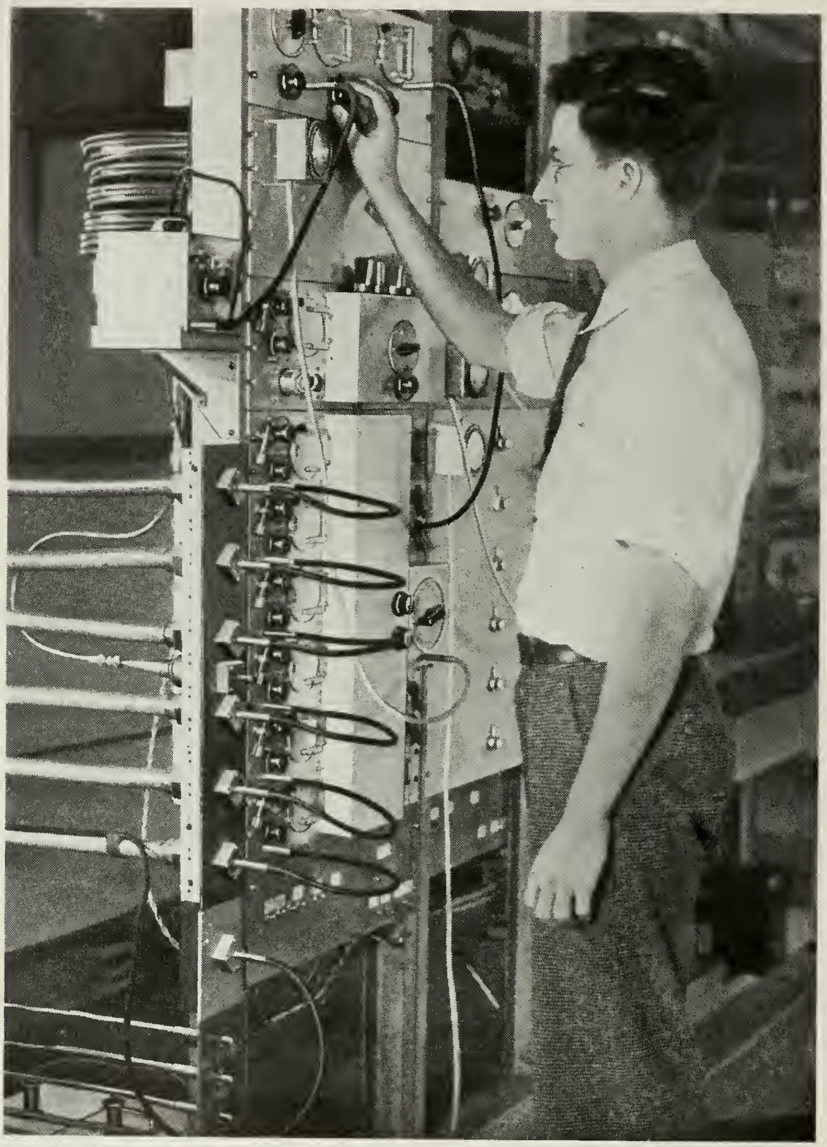


Fig. 17—View showing the six transmission lines and coaxial patch cords. The beating oscillator is mounted upon the shelf and is connected to the power amplifier (which is being adjusted by Mr. Edwards) at the top of the bay.

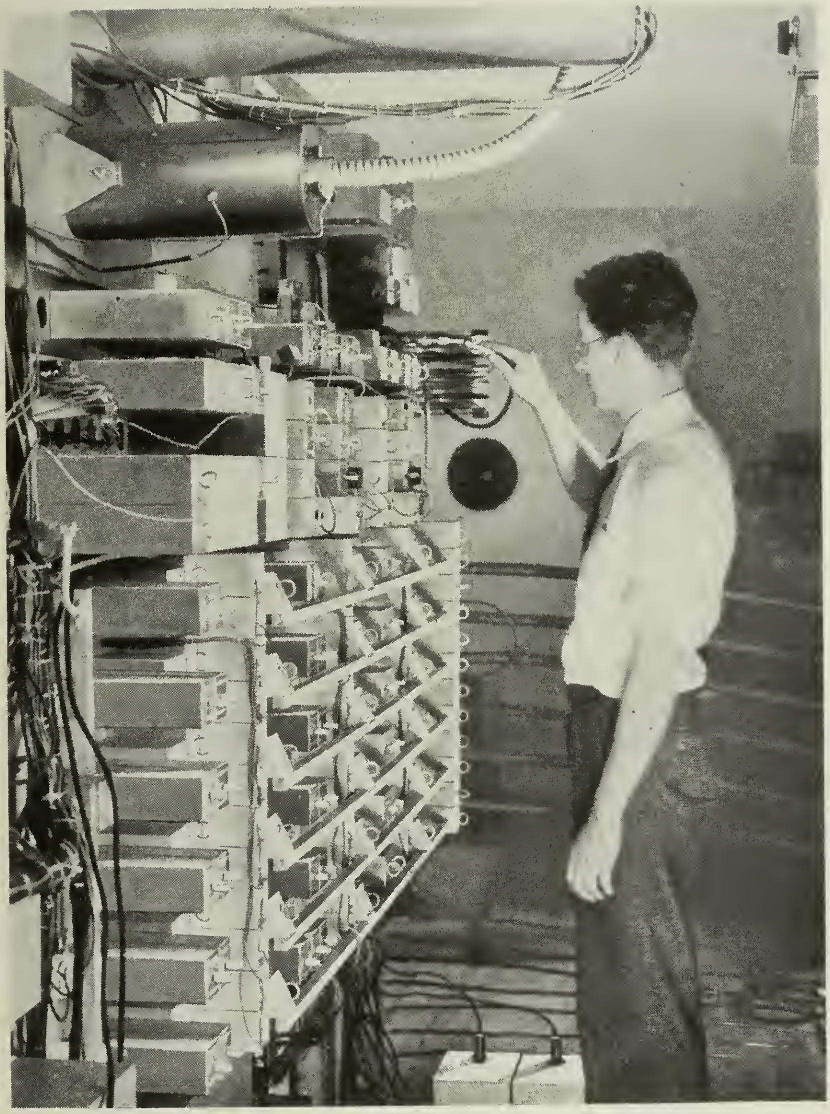


Fig. 18—Rear view of the receiving equipment. The six detector outputs feed the three branches via the square transmission lines.

A family of calculated directional patterns of the experimental MUSA is shown in Figs. 19 and 20. At the top of each column is shown the principal lobe of the vertical directional pattern of the unit rhombic antenna, calculated in the median plane. Beneath are shown six vertical patterns of the MUSA, which are obtained by multiplying the

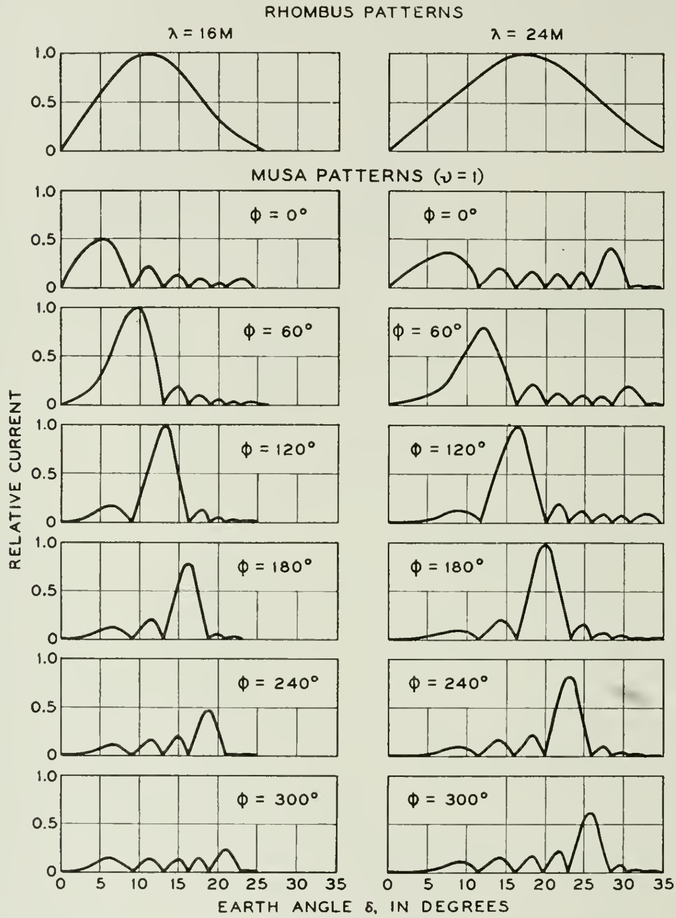


Fig. 19—Vertical directional patterns of the experimental MUSA.

array factor [20] by the unit antenna pattern. The upper pattern corresponds to phasing for zero angle. The remaining ones are plotted for increments of 60 degrees of phase.

These patterns fall short of the "ideal," which the reader may have visualized while reading Section II, in two ways. First, the unit an-

[20] Calculated from (3) putting $\upsilon = 1$.

tenna does not suppress the second lobe of the array factor as well as could be desired. By design, it does so for the short waves but inherently fails to do so for the longer waves. Second, the principal lobe of the unit antenna shifts bodily towards higher angles with increasing wave-length, whereas it is desirable to have only the upper cutoff move

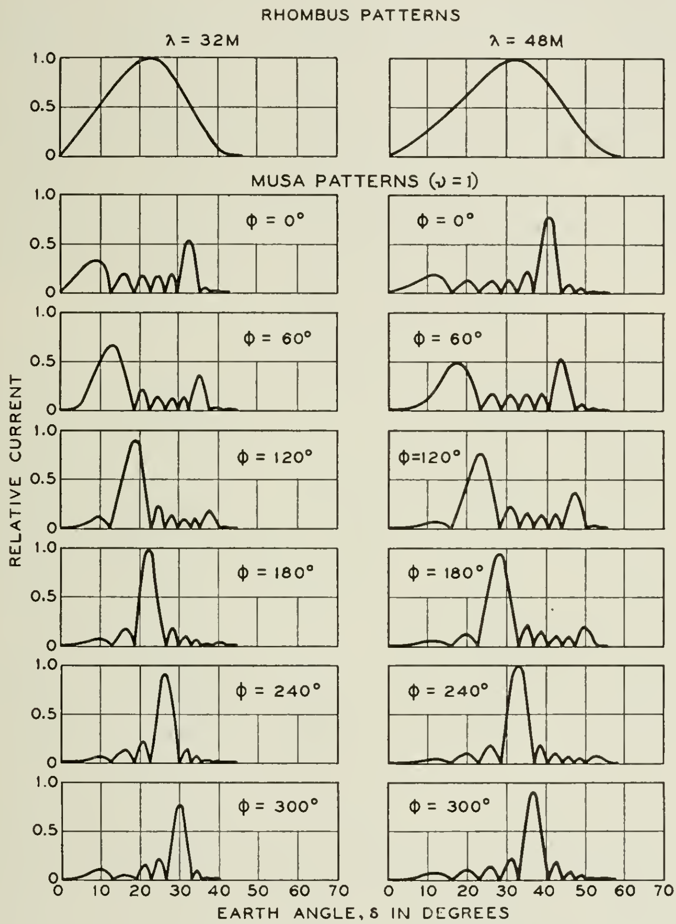


Fig. 20—Vertical directional patterns of the experimental MUSA. Note that the angle scale is half of that in Fig. 19.

upward while retaining low-angle response. Both of these shortcomings restrict the steering range through which the MUSA is essentially single lobed. The upward shift with wave-length of the unit antenna response is in fair agreement with the way in which the mean angle of arrival has been found to vary with wave-length and is, on that account, not altogether objectionable. The Holmdel MUSA employing such unit antennas represents, however, a considerable departure from present antennas of fixed directivity designed from statistical data, and approaches the ideal MUSA steerable over the entire useful angle range.

The curves as plotted assume that the differences in transmission line loss for the various line-lengths have been equalized in the intermediate-frequency circuits. By slightly tapering the amplitudes so that the antennas in the middle of the array contribute more than those near the ends a reduction of the minor lobes has been obtained at the cost of slightly widening the principal lobe. As a result of this, the directional discrimination of the experimental MUSA has been improved. All data and photographic records reported in this paper, however, were obtained before this improvement was introduced.

IV. Tests and General Evaluation [21]

Tests and Experience

Numerous experiments and tests had been carried out on the various parts of the MUSA system before it was first tuned to a transatlantic signal. Despite the fact that all tests concurred in predicting that the system would perform as designed, it was with considerable gratification that a pattern was observed on the monitoring oscilloscope, during one of the early trials, which was almost exactly as calculated for a single wave. Patterns corresponding to two or more waves in various degrees of resolution were observed from time to time. To increase the angle resolution, for test purposes, pulses were transmitted by the British Post Office on several occasions. Turning the steering shaft during these tests clearly showed the principal lobe sweeping through the angle range. When fairly discrete pulses were received the minor lobes could be readily identified. In Fig. 21 is shown a sample of motion picture oscillograms of pulse reception. Two principal waves or, more accurately, wave bundles occurred and were separated by the two MUSA branches as shown. For details of the pulse technique employed in these tests the reader is referred to a previous publication.[5]

Before exhibiting sample motion pictures of typical patterns displayed by the angle monitoring oscilloscope and the delay indicator oscilloscope, further discussion of the former is desirable. The photographs of Fig. 22 show the monitoring oscilloscope pattern with a locally produced equiphase, equiamplitude input supplied to each

[21] The theory and test results of the signal-to-noise advantage are considered together in Part V.

Fig. 21—This retouched plate shows pulses received with MUSA. On each frame time advances from left to right and is measured by the thousand-cycle timing wave. The center trace shows the output of one MUSA branch steered at 25 degrees. The bottom trace shows the output of the second MUSA branch steered at 32 degrees. Wide band amplifiers for pulse reception are bridged in parallel with the speech band intermediate-frequency amplifiers. The transmitted pulses are about 200 microseconds long. GCS (9020 kilocycles) Rugby, February 25, 1936, at 4:11 P.M., E.S.T.

Fig. 22—These five frames show the angle monitoring pattern when a local signal is used to simulate a wave. The bottom frame shows the ideal MUSA pattern for one wave. The remaining frames show the effect of reducing the number of antennas from six (1-2-3-4-5-6) to five (2-3-4-5-6) to four (3-4-5-6) to three (4-5-6) to two (5-6). These films were taken before the amplitude tapering was introduced. Tapered amplitudes reduce the minor lobes to about half of the amplitudes shown.

detector. This figure illustrates the manner in which the pattern is built from the six components after the manner of a Fourier synthesis. The vertical and horizontal axes visible on the monitoring oscilloscope in Fig. 16, but which do not appear naturally on the photographs, were drawn in Fig. 22. As mentioned previously, the oscilloscope sweep axis represents one revolution of the "fundamental" phase shifter so that the beginning and end of the sweep represent the same condition. The ends of the sweep are arbitrarily fixed to represent zero (or 360) degrees of phase shift referred to the output of the first antenna, whose phase is not varied. Consequently equiphase inputs result in a principal lobe half of which appears at each end. This would correspond to a wave of zero angle if the velocity of the transmission lines was equal to that of light. For a lesser velocity, zero angle may occur at any point on the phase axis, depending upon the wave-length. (See Fig. 28 for a sample angle calibration curve.) The principal lobe as well as the four minor lobes of the monitoring oscilloscope represents the output from one wave as the MUSA is steered through its entire range. The oscilloscope pattern, unlike the directional pattern, does not appear sharper for short wave-lengths than for long wave-lengths; the principal lobe is always 120 degrees wide and the minor ones 60 degrees wide on the phase axis. One degree of phase difference, however, represents a difference in steering angle which depends upon the wave-length and the earth angle.

The samples of motion picture film shown in Fig. 23 represent fairly typical "two-path" patterns. The camera was focused to include both oscilloscopes and was manipulated by means of a special step-by-step crank. The operator endeavored to expose each frame during one sweep of the monitoring tube. The delay indicator tube shows a continuous pattern produced by the audio frequencies. A correct delay setting is indicated by a straight line. Here, with the two branches steered at the indicated angles of 8.5 and 20.5 degrees, a delay of 950 microseconds was required to produce the straight line. The diversity action is apparent in the tilting of this line. When the low angle wave, which corresponds to the left-hand peak on the monitoring tube, is predominant the delay indicator line becomes horizontal and, conversely, when the high angle wave is predominant the line approaches the vertical axis. Automatic gain control is used on the branch receivers supplying the speech outputs but is not used on the monitoring branch.

Figure 24 shows, in samples 1 and 2, reception of two waves which are just separable by the directivity present in the Holmdel MUSA. The angles are 15 and 22 degrees and the wave-length is 31.6 meters. The

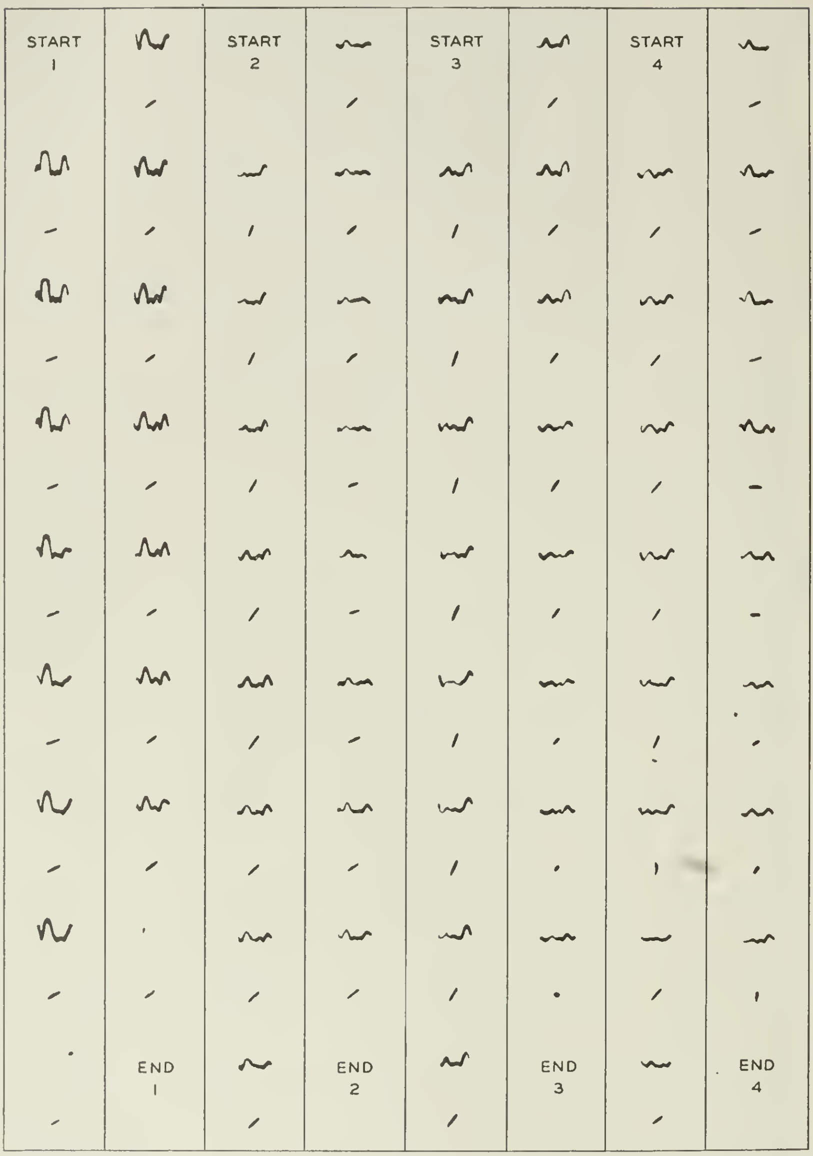


Fig. 23—Pictures of the angle monitoring oscilloscope and the delay indicator tube. The right-hand end of the sweep of the monitoring tube represents zero phase. The indicated angles are 8.5 and 20.5 degrees. The MUSA branches were set at these angles ($\phi_A = 240°$, $\phi_B = 30°$) using 950 microseconds delay. GSE (11,860 kilocycles) Daventry, February 21, 1936, 11:05 A.M., E.S.T. Musical program.

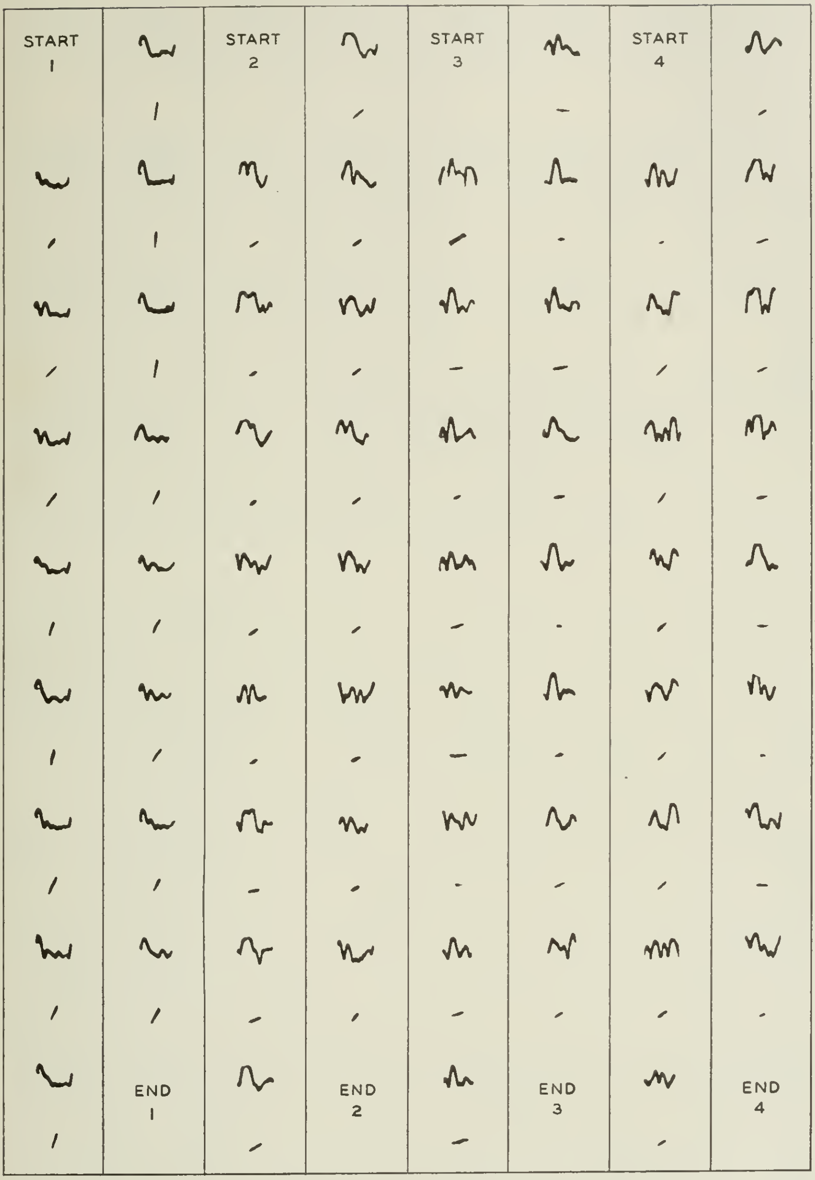


Fig. 24—Pictures of the angle monitoring oscilloscope and the delay indicator tube. In films 1 and 2 the indicated angles are 15 and 22 degrees. The MUSA branches were set at these angles ($\phi_A = 240°$, $\phi_B = 340°$) and a delay of 400 microseconds was used. The time was 3:20 P.M., E.S.T. Films 3 and 4 were taken at 3:15 when a third wave of 26 degrees was present. One branch was steered at this wave ($\phi_B = 50°$); the other at 16 degrees ($\phi_A = 250°$). A delay of 1000 microseconds was required. GSB (9510 kilocycles) Daventry, March 10, 1936.

delay used is 400 microseconds. In samples 3 and 4 a third wave of 26 degrees is present. One branch was steered at this wave; the other was steered at the 15-degree wave and a delay of 1000 microseconds was used.

It is of interest to compare these samples showing the manner in which the MUSA branch outputs combine, with the samples in Fig. 25 which were obtained with a two-antenna space diversity setup. Six antennas were retained in the monitoring branch but five were cut out of each receiving branch, leaving one antenna to supply each branch. In samples 1 and 2, antennas 1 and 6 (1000 meters apart) were retained. In samples 3 and 4, adjacent antennas (Nos. 1 and 2) 200 meters apart were used. These records were obtained about 15 minutes later than those of Fig. 23 and show the same two waves at 8.5 and 20.5 degrees. No delay was used. Note that the outputs combine in phase only when one wave predominates. Inserting delay in either branch is, of course, not effective in improving the audio combination. To do so would impair the addition when one wave is predominant and would not be beneficial when both waves are comparable.

Figure 26 shows, in samples 1, 2, 3, and 4, how the delay indicator tube pattern is affected by the delay adjustment. The two branches were steered at the same angle, thus making both branch outputs identical so that perfect delay adjustment occurs with zero delay. This is the condition depicted in sample 1. In samples 2, 3, and 4 the delays are 340, 680, and 2700 microseconds, respectively.

A number of tests were carried out with the cooperation of the British Post Office in which twelve tones were transmitted. These tones were nonharmonically related. They were separated at the output of the receiver by means of filters, and commutated to appear successively on an oscilloscope. The reader is referred to a paper [4] by R. K. Potter describing this technique. Figure 27 shows a sample of motion pictures made of the oscilloscope patterns. Two receiving systems are compared; the right-hand pattern shows the output of the MUSA while the left shows the output of a conventional receiver connected to a horizontal half-wave antenna. The tones trace the horizontal lines in sequence from top (425 cycles) to bottom (2125 cycles). After one pattern is executed the commutator switches from one receiver to the other. The twelfth tone is omitted to provide time for the switching. The complete double pattern is traced in about one-sixth of a second and the camera is operated at a speed which exposes each frame a little longer than one-sixth of a second.

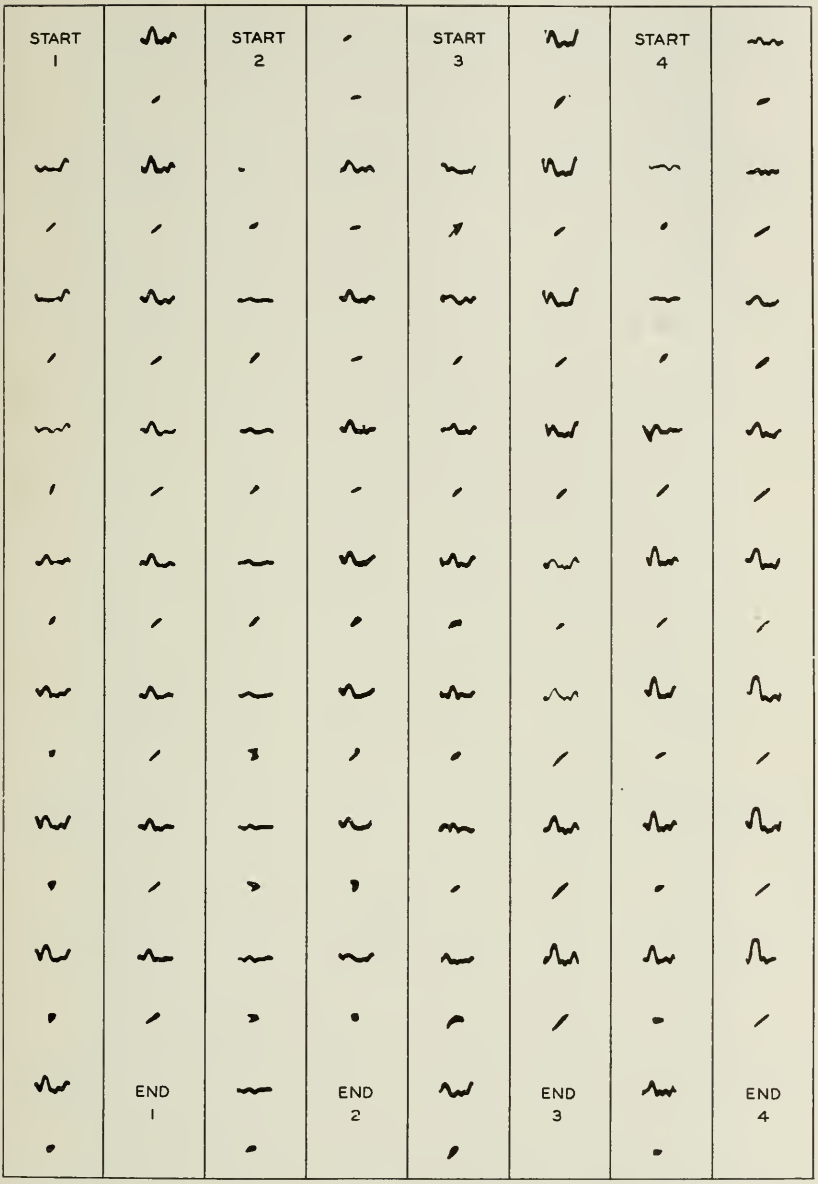


Fig. 25—This plate, made immediately following that of Fig. 23, shows for comparison the manner in which the audio outputs add in two-station space diversity. The angle monitor shows the 8.5- and 20.5-degree waves as before. Films 1 and 2 taken at 11:15 A.M., E.S.T., were obtained with rhombic antennas 1 and 6 (40 wave-lengths apart). Films 3 and 4 taken at 11:20 were obtained with antennas 1 and 2 (8 wave-lengths apart). Note the second harmonic in film 2 particularly. GSE (11,860 kilocycles) Daventry, February 21, 1936. Musical program. Zero delay.

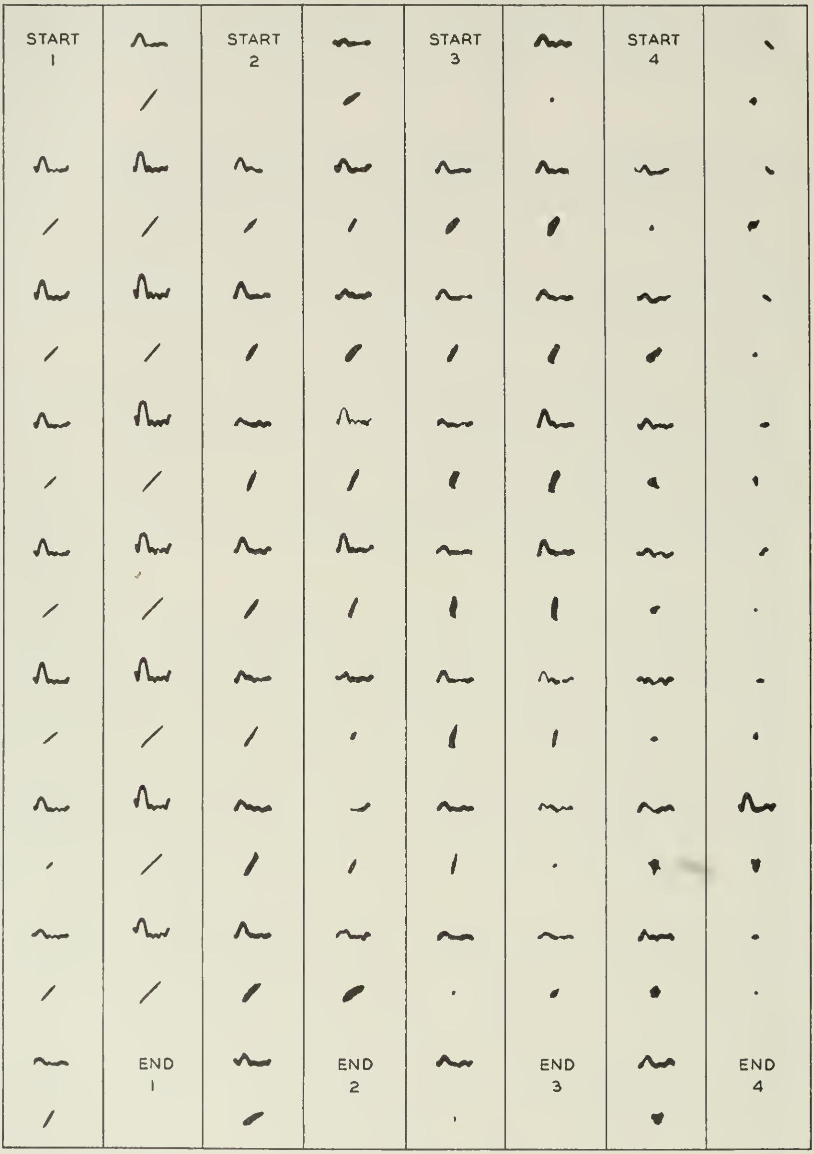


Fig. 26—Film showing the effect upon the audio addition of unequalized delay. Both MUSA branches were steered at the same angle,—that of the major wave shown. Film 1 shows no delay added and since each branch receives the same wave the audio outputs add perfectly. Films 2, 3, and 4 show the effect of adding 340, 680, and 2700 microseconds delay, respectively.

Fig. 27—This is a cathode-ray "multitone" record comparing the MUSA output with that of a horizontal half-wave antenna. The tones on the right-hand side of each frame are the MUSA output; those on the left are the output of the horizontal half-wave antenna. The two MUSA branches were steered at 15 and 22.5 degrees. A delay of 470 microseconds was used to equalize the transmission time. GCS (9020 kilocycles) Rugby, February 24, 1936, at 3:54 P.M., E.S.T.

In Fig. 27 the MUSA branches were steered at 15 and 22.5 degrees and employed an equalizing delay of 470 microseconds. While the MUSA output is not perfect it is vastly superior to that of the doublet. The tone frequencies and filters are such as to suppress harmonic distortion with the result that the patterns show mainly the selective fading of the fundamental audio frequencies. Note that the fundamental output nearly disappears in the doublet receiver. In practice this would correspond to violent harmonic distortion of speech or music.

In addition to the tests and experiments illustrated by the motion picture reproductions in the preceding paragraphs a series of experiments were conducted using broadcast transmission on 49 meters from a station at Halifax, Nova Scotia. In these experiments angles and delay differences were measured and compared with the multiple reflection theory. The agreement between measured and predicted values is not only interesting as a study of the ionosphere but constitutes a unique and valuable test of the performance of the MUSA system.

Observations on VE9HX, Halifax

During the course of reception experiments with GSL (BBC, Daventry, 6110 kilocycles) performed as a part of the routine operating program for the MUSA system, a broadcast station appeared on GSL's frequency. This station carried the programs of CHNS, Halifax, Nova Scotia, and was subsequently determined to be an experimental station with the call letters VE9HX located hear Halifax and nearly on the great-circle path from New York to London. The transmitting antenna is a half-wave horizontal, one-quarter wave above ground and oriented to radiate in the direction of New York.

The first experience with this station showed two stable transmission paths capable of being separated by the two branches of the MUSA. The delays could be accurately equalized and rather definite correlation was obtained with the multiple "hop" propagation picture. This fact and the additional reason that propagation from England on the same frequency might be compared with the simpler phenomena encountered with Halifax led to the measurements described in the following paragraphs.

About eleven hours of observation, distributed over fifteen days, are included. The log aimed to record all changes which occurred during an observation period. The procedure was as follows: The two branches of the receiver were steered at the angles indicated by the monitoring oscilloscope. Delay was added to the lower angle branch until the two audio outputs added. The delay setting was usually critical to one section of the network (67.5 microseconds) and always to two sections. The angles were determined from the calibration curve reproduced in Fig. 28. The phase readings observed on the monitoring oscilloscope were recorded to within ± 10 degrees and the earth angles determined by them are liable to be in error by one degree (possibly 1.5 degrees) apart from the ambiguity due to the multiple lobe characteristics of the MUSA. At this wave-length, the major lobe of the unit rhombic antennas is broad, the first null occurring at 58 degrees, so that two angles had to be considered possible.

The multiple hop picture is illustrated in Fig. 29. Here the delay

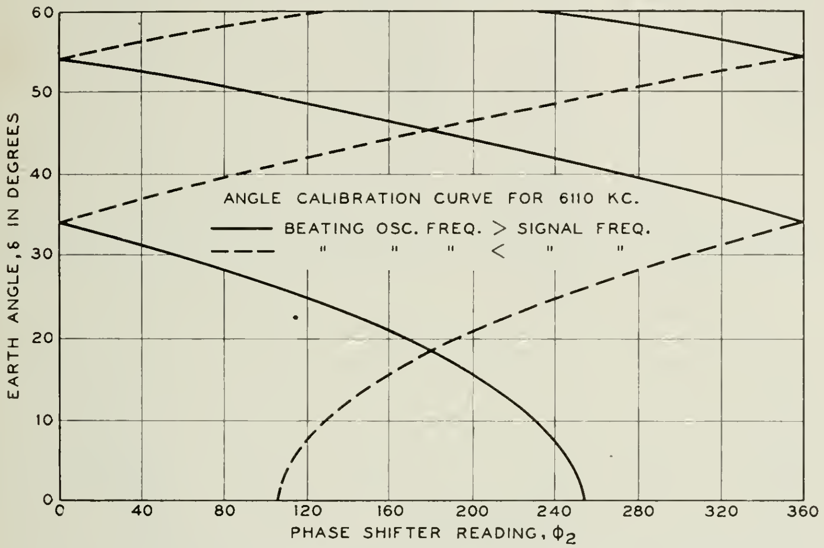


Fig. 28—Calibration curves of the Holmdel MUSA for 49.1 meters, giving the angle of the principal lobe as a function of phase advance ϕ_2 (Fig. 3). Note that the sense of the phase shift depends upon the beating oscillator frequency. The curves are calculated for a velocity ratio $v = 1/0.933$.

referred to the ground wave is expressed in terms of earth angles δ and n, the number of hops or ionosphere reflections. The height h and angle δ are also related through n as shown in Fig. 29. Using the first relation, the curves of Fig. 30 were drawn; using the second relation, points corresponding to various heights were located on the curves. For the Holmdel-Halifax circuit d is 643 miles (1030 kilometers) making $\beta = 9° 21'$. Corresponding to each measured angle there is a delay (referred arbitrarily to the ground wave which, of course, was not received) and a layer height, for each of the modes or orders. Both angles together yield a delay difference which is to be compared with the measured value.

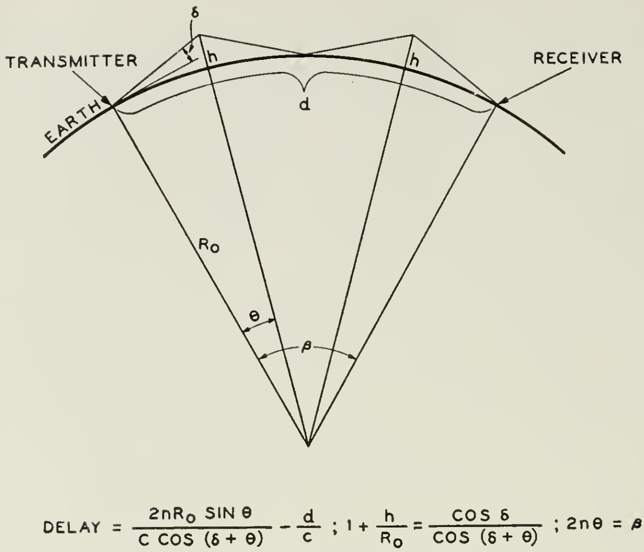


Fig. 29—Delay and angle relations for multiple reflection from a uniform reflecting surface. The number of ionosphere reflections is designated by *n*.

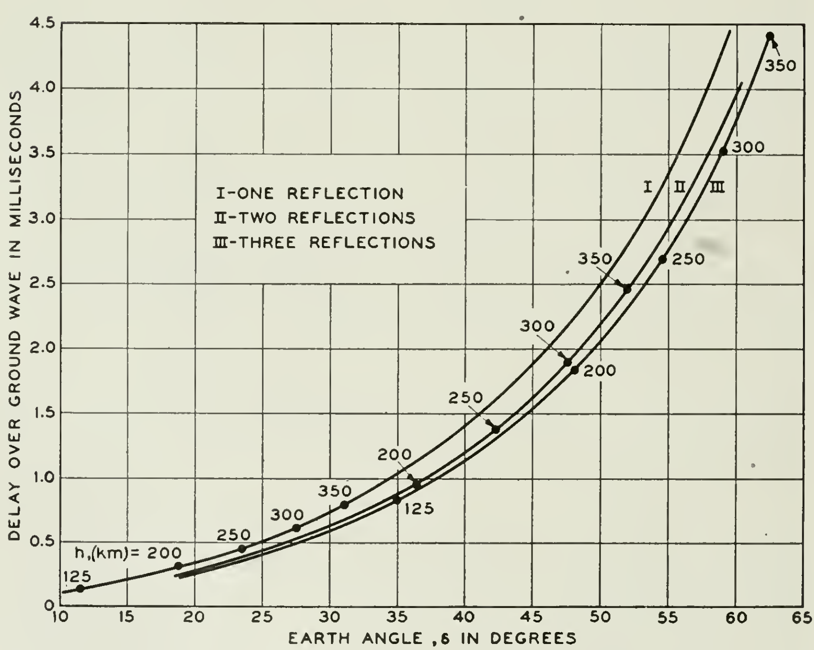


Fig. 30—Curves giving the delay-angle relations for multiple reflection on the Halifax-to-Holmdel path.

TABLE I
OBSERVATIONS ON VE9HX, HALIFAX, NOVA SCOTIA
6110 kilocycles 49.1 meters

	E.S.T.	Date	δ_1°	δ_2°	Relative Delay milliseconds		Virtual Height kilometers			Estimated		Field decibels above 1μv/m.
					meas.	calc.	1st	2nd	3rd	1st	2nd	
		1935										
	P.M. 5:00	11–25	18.2	38	1.01	0.77	195	215				
	P.M. 4:45	11–26	24.5	42	1.01	0.85	260	250				
	P.M. 4:40	12–17	26.4	44	0.95	0.97	290	270				27
	P.M. 4:41	12–17	24.5	43	0.95	0.95	265	255				
	P.M. 5:01	12–17	24.5	42	0.95	0.85	265	250				
	P.M. 4:20	12–18	25.5	44	0.95	1.05	275	270				27
	P.M. 4:25	12–18	24.0	42.5	0.95	0.90	250	250				
A	P.M. 4:40	12–18	23.0	42	—	0.90	245	250				
	P.M. 4:30	12–23	24.5	44	0.95	1.05	265	270				25
	A.M. 10:29	12–26	28.0	44	0.95	0.90	305	270				−14
	A.M. 10:46	12–26	28.0	44	0.81	0.90	305	270				
	A.M. 10:52	12–26	29.5	44	0.78	0.85	330	270				
	P.M. 4:33	12–26	24.5	44	0.88	1.05	265	270		242	245	18
	P.M. 4:49	12–26	24.5	42	0.95	0.85	265	250				
	P.M. 4:58	12–26	24.0	42.5	0.88	0.90	250	250				
	A.M. 10:07	12–27	25.5	43	0.95	0.95	275	255				−14
	A.M. 10:57	12–27	29.5	42	0.68	0.65	330	250		130	243	
	A.M. 11:03	12–27	31.0	42	0.88	0.70		250	100			
	A.M. 11:20	12–27	<8.0	45.5	1.28	1.6+	<120	280				
B						1.5+	<120		185			
	A.M. 11:48	12–27	8.0	47.5	1.28	1.8+	<120	300				
						1.7	<120		195			
	A.M. 11:56	12–27	8.0	45.5	1.28	1.6+	<120	280		130	247	
						1.5+	<120		185			
C	P.M. 4:32	12–27	25.5	44	0.88	1.05	275	270				14
	P.M. 4:59	12–27	27.0	43	0.88	0.85	295	255				
D	A.M. 10:45	12–31	31.0	42	0.50	0.55	350	250				0
	A.M. 11:45	12–31	8.0	35.5	0.71	0.8+	<120	195				
		1936										
	A.M. 10:30	1–2	24.5	42	0.88	0.85	265	250				−2
	P.M. 6:05	1–14	20.5	42	1.01	1.00	215	250				8
	P.M. 6:35	1–14	18.4	38	1.01	0.77	200	215				
E	P.M. 6:15	1–15	23.0	42	1.08	0.90	245	250				22
	P.M. 6:20	1–15	24.0	42	1.11	0.85	250	250				
	P.M. 6:40	1–16	25.5	42	1.08	0.85	275	250				14
	P.M. 7:21	1–16	24.5	43	1.18	0.95	265	255				
F	P.M. 8:39	1–16	31.5	37	0.27	0.20	355	205				
	P.M. 9:35	1–16	26.4	37	0.47	0.42	290	205				27
	P.M. 5:50	1–21	24.5	42	0.95	0.85	265	250		267	247	22
G	P.M. 6:10	1–21	26.4	44	1.01	0.97	290	270				
	P.M. 6:16	1–21	22.0	40	0.95	0.80	235	230		232	245	
	A.M. 10:40	1–22	24.5	43	0.95	0.95	265	255				−2
	A.M. 11:05	1–22	24.5	34	0.41	0.35	265	185				
						0.30	265		120			
H	A.M. 11:09	1–22	34.0	43	0.60	0.60	185	255				
						0.65		255	120			
	A.M. 11:30	1–22	24.5	43	0.95	0.95	265	255				
	A.M. 11:35	1–22	24.5	34	0.41	0.35	265	185				
						0.30	265		120			
I	P.M. 6:45	1–24	18.4	38	0.74	0.77	200	215				2

In Table I the virtual heights are deduced from the curves for the assumed hop orders. The calculated relative delay is the delay difference corresponding to these heights. All angles below 60 degrees were considered and all combinations of hop orders were considered for each angle, subject to the experimental knowledge of the sense of the delay. The values shown in the table are the ones which give the best agreement with the measured delay. In most instances there was no question concerning the interpretation; in a few doubtful cases two possibilities are presented (December 27 and January 22).

Examination of the table shows that except near noon, the propagation comprises the first and second reflections from the F region of the ionosphere. Groups A, C, E and G illustrate this. In the majority of instances the agreement is excellent; these cases constitute strong evidence that the MUSA performs correctly.

The discrepancies in the table between layer heights for the first and second hops and between measured and calculated delay are not entirely experimental error. Assuming errors in measured angles sufficient to make the delays agree will, in some cases, increase the discrepancy in heights. An interpretation one might make of this is that the ionosphere is not uniform over the circuit and the regular reflection basis of calculating is not strictly in accord with facts. However, there are other theoretical explanations for discrepancies in height. Under usual conditions, the second reflection height should be slightly greater than the first but for certain ionizations in the E region, the first F reflection may be retarded more than the second F reflection in passing through the E region. Thus the heights may differ in either direction without demanding horizontal non-uniformity. The discrepancies between measured and calculated delay may be explained by horizontal non-uniformity in the ionosphere. For an essentially non-dissipative atmosphere of ions having any vertical distribution but no horizontal gradient, and neglecting the earth's magnetic field, the group delay is identical with that calculated from triangular paths coinciding with the initial earth angles. Breit and Tuve showed this in their 1926 paper. With horizontal variations in the ionosphere such as tilting layers, no kind of agreement could be expected; the waves might even travel via other than great circle routes.

During three days of our observations W. M. Goodall made measurements of virtual height and of critical frequency which enabled him to predict the results we might be expected to observe. His estimates are shown in the next to the last column of the table.

The data for December 27 (B) are interesting in that after 11 o'clock the first F reflection apparently disappeared. Instead, a first reflection from the E layer is indicated. This was predicted by Mr. Goodall on the basis that the E region ionization at noon became so great that 24-degree waves should be reflected. For completeness the table shows an alternative interpretation of a first E reflection and a third reflection from a 185- to 195-kilometer height. The first E reflection and second F reflection are perhaps more likely. The 11:03 record is not explained.

Something similar appeared to happen on December 31 (D). On January 22 (H) normal first and second F reflections occurred with angles of 24.5 and 43 degrees. In addition a third wave of 34 degrees appeared. Two interpretations of this are shown but neither seems very plausible.

As a general rule propagation from Halifax is simpler than from Daventry on the same wave-length. In particular GSL waves received by the two MUSA branches are definitely less discrete and include sufficient delay differences in themselves to prevent the nicety of equalization possible with VE9HX. If multiple reflection takes place, which we have no reason to doubt, it is generally so distorted by non-uniformity over the path or by other factors as to be unrecognizable. In view of the occasional complexity of the Halifax circuit, only one-sixth as long, this is perhaps to be expected.

The absence from these observations on Halifax of any third reflections from the F layer is likely due to the fact that they would fall in the neighborhood of the first null of the rhombic antenna and would have to be much stronger in space in order to appear comparable with the second or first. There have been momentary appearances of waves which might have been third reflections but they did not persist long enough to work with.

When single waves were present, which was not unusual in the later evening hours, the angle more often corresponded with the first F reflection rather than the second.

Additional Numerical Data on Reception with the MUSA

The data shown in Fig. 31 are submitted to supplement the rather meager numerical data on transatlantic reception thus far presented. Here, relative delays and angles taken from the MUSA operating log are shown in plots A, B, and C. Only the end points of the lines are significant; they denote by their abscissas the angles at which the two receiving branches were set. The ordinates of the upper end points denote the equalizing delay. The lines merely connect coexistent points. The data shown were selected from the rather extensive log to present a fair cross section of conditions, omitting, however, all cases in which both branches were steered at the same wave bundle. They cover winter and summer and were obtained with frequencies appropriate to the time and season. Most of the observations were made on transmission from Daventry, the remainder on transmission from Rugby. In D are shown the results of pulse measurements made before the MUSA was in use. Here the angles were measured by the two antenna null method and the delays were observed directly on the oscilloscope time axis.[5] Although as many as five points, each denoting a wave bundle, are shown, generally not more than three were

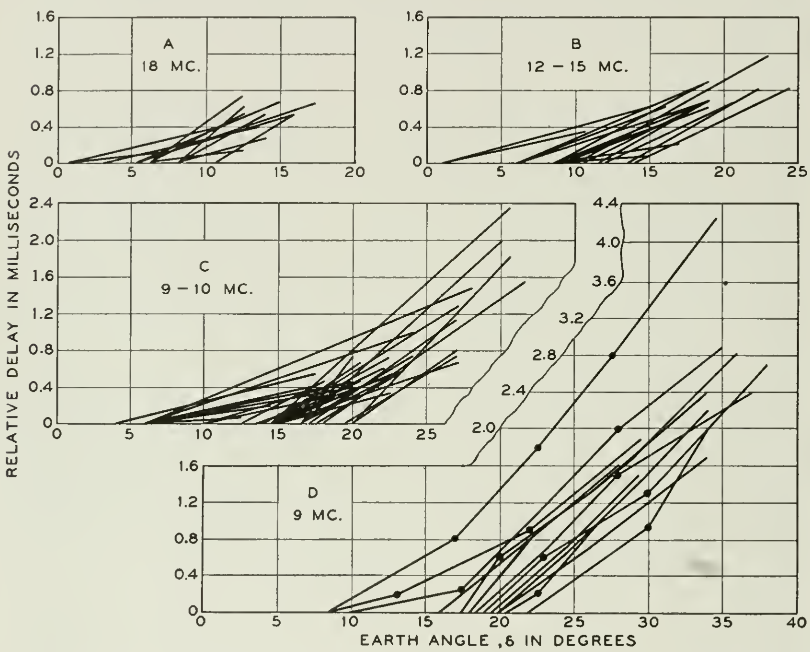


Fig. 31—Pairs of measured angles and relative delay denoted by the end points of the line segments. The data in *A*, *B*, and *C* were obtained with the MUSA; that of *D* was obtained by the use of pulses.

important at once. These measurements were made on transmission from Rugby.

It will be noticed that all four groups of data show that the relative delay per degree of angle difference is small at low angles and increases with the angle, roughly as the multiple reflection theory indicates. (This characteristic is distinctly favorable to the performance of the MUSA since its angle resolving power falls off at very low angles.) The scattering of the data indicates that an equalizing delay determined by the angle settings would not be successful; i.e., the delay must be capable of adjustment to meet the various transmission conditions.

Quality Improvement with the MUSA

The distortion of speech and musical quality which characterizes short-wave circuits is due entirely to the interference of differently delayed waves each of which individually is fundamentally free from all kinds of distortion except non-selective fading. This conclusion is almost self-evident and is corroborated by the results of several years of pulse investigation [5] made in cooperation with the British Post Office.

A MUSA system can be expected to select one out of several multiple reflections. However, these reflections suffer more or less scattering with the result that they appear as bundles of waves of various degrees of compactness. These bundles possess a small spread of both angle and delay. The delay interval included in a bundle of waves is rarely less than 100 microseconds. Double refraction or "magnetoionic splitting" occurring in the ionosphere doubtless accounts for the existence of a small minimum delay. A delay interval of fifty microseconds or so may be detected even in the unusually compact bundles represented by the pulses of Fig. 21. Transmission from Halifax appears to include a delay interval of this order, also. With transatlantic propagation it is not uncommon to have a bundle containing numerous weaker components extending over several hundred microseconds. On rare occasions these have extended over two milliseconds, masking any multiple reflections which may have been present.

The quality associated with one MUSA branch which selects one out of several bundles of waves is thus not perfect. The effect of a delay interval of a few hundred microseconds is scarcely noticeable, however, except during deep carrier fades. Therefore, if diversity action between two branches steered at the low and high angle parts of the same bundle is employed, deep fades are avoided to a large extent, and the quality is almost perfect. When more than one wave bundle is present diversity action between branches steered at the principal bundles accomplishes this escape from deep fades. It is desirable to utilize all of the principal bundles in diversity in order to preserve the discrimination of the MUSA. For, one of the bundles, if not provided with a branch to receive it, would cross talk into the other branches when it momentarily became strong and those provided with branch receivers became weak. Signal-to-noise ratio considerations discussed in Section V constitute an equally important (and related) reason for utilizing all principal bundles.

As distinguished from selective fading, which is greatly reduced by the rejection of all but one wave bundle, general fading is by no means eliminated. The reader may expect, however, that when the MUSA selects one wave bundle from several it restricts the waves accepted to those which have traveled more nearly a common path, and for a given degree of turbulence in the ionosphere, the fading should be slower, since only relative changes among the several waves result in interference fading. Such a tendency no doubt exists and has been noticed occasionally in the operation of the MUSA but rarely has there been a marked effect (excepting certain cases of flutter fading to be described later). This will be understood when it is recalled that even a fifty-microsecond delay interval means that a difference of 500 wave-lengths is involved for a wave-length of thirty meters. In order that the fading rate be sharply reduced it is required that the ionosphere shall preserve this difference, to within a half wave-length, more effectively than it does if larger differences are involved. Since a half wave-length is only 0.1 per cent of 500 wave-lengths a rather high degree of balance is thus required. Evidently, the turbulence of the ionosphere usually prevents such a balance.

Using broadcast signals (double side band) from Daventry a thousand or more comparisons were made of the MUSA versus a single antenna and receiver, using the switching arrangement mentioned in Section III. Remarkable improvements were sometimes observed and some improvement was almost always noted. The exceptions were the instances when distortion was not detectable using one antenna, and the rare occasions when particularly violent flutter fading occurred.

Space diversity reception using two antennas showed a substantial improvement, usually, but failed ever to show the order of improvement demonstrated by the two-branch MUSA when two or more wave bundles of comparable amplitude occurred. Figures 23 and 25 suggest, by the way in which the audio outputs are seen to combine, that the distortion with MUSA reception is slight compared to that with diversity reception.

The increased naturalness which results from reducing the distortion is, of course, pleasing to the ear and has some value in telephone circuits on account of the subscribers' satisfaction. In addition, it increases the intelligibility particularly when considerable noise is present. It is impossible to evaluate the increased intelligibility definitely but, in certain cases at least, it permits the signal-to-noise ratio to be two or three decibels lower. From the point of view of picking up short-wave broadcasts for rebroadcasting, a more substantial value can be attached to the MUSA quality improvement.

To a considerable extent, the magnitude of the quality improvement ascribed to the MUSA in the preceding paragraphs depends upon the fact that double side-band signals were employed. For, with double side-band signals the selective fading caused by the interference of the differently delayed waves results not only in selective fading of the audio output, but also produces non-linear distortion when the carrier fades selectively. This non-linear distortion sounds much like over-modulation, and when it occurs in its more violent forms it completely ruins the quality and intelligibility. With single side-band transmission it is possible to demodulate with such a strong carrier that non-linear distortion is virtually eliminated. The fading of the audio output is sometimes more selective than with double side-band but the resulting quality is substantially better.

Single side-band transatlantic signals were not available during the trial of the MUSA system. However, as mentioned in Section III, receiving equipment was available which rejects one side band and reduces the percentage modulation by a factor of ten or more. It was found that this equipment, applied to the one-antenna system, resulted in substantially reduced non-linear distortion and that the quality could be still further improved by the reduction of selective fading afforded by the MUSA. With MUSA reception there was apparently no quality improvement in going from double to single side band.

Summarizing Discussion

In this section the general performance of the experimental MUSA has been described in a necessarily qualitative manner. Motion picture oscillograms were shown to illustrate the performance under fairly typical transatlantic conditions. An investigation of propagation from Halifax in which the MUSA was employed to identify ionosphere reflections was included to supplement the rather fragmentary evidence available in motion picture oscillograms. The improved quality obtained with MUSA reception was discussed from several points of view. The evaluation of the MUSA has been general; it serves partly to introduce the following section which deals specifically with the signal-to-noise ratio evaluation.

Before closing this section it is appropriate to discuss conditions with which the experimental MUSA could not adequately cope.

On numerous occasions the fact that only two branches are provided has definitely handicapped the performance. More often, however, the need for greater angular discrimination or resolving power has been apparent. Except on infrequent occasions a MUSA two to three times the length of the experimental one and equipped with three branch receivers could be expected to perform as well as the experimental one now performs at its best. The occasions when it might not are the infrequent times when violent flutter fading occurs.

At least one type of flutter fading appears to be associated with a pronounced scattering which results in a kind of shower of erratic waves arriving over a wide range of directions. Receiving antenna directivity has been found definitely helpful in all except the most violent cases. Apparently when improvements due to directivity occur they occur principally by selecting a more or less normally propagated wave bundle and rejecting the shower of erratic scattered waves. When, in the most violent cases, no reduction of the flutter can be achieved the reason may be that the unit antenna accepts too wide a horizontal range to permit the MUSA to discriminate sufficiently against the shower. (It will be remembered that the MUSA array factor is of the form of a semiconical shell and thus the MUSA will, in general, accept as wide a horizontal range as the unit antenna permits.)

V. The Signal-to-Noise Improvement of the MUSA Receiving System

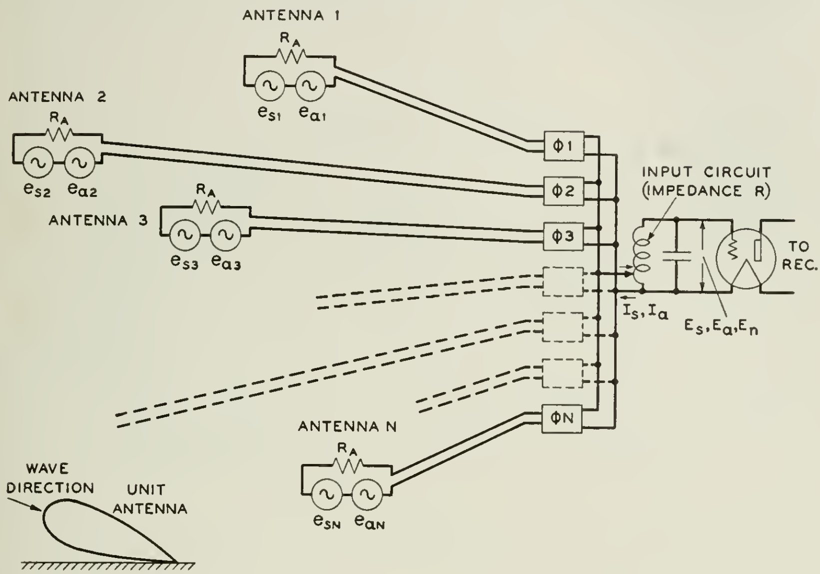
Because of the complicated nature of short-wave transmission and also because of the uncertain state of noise measuring technique, it is not a simple matter to give a satisfactory answer to the question: "What is the signal-to-noise improvement of a MUSA system?" In this section an attempt has been made to simplify the problem by separating the various factors involved. The section begins with an analysis of the problem assuming simple types of wave transmission. This is followed by experimental studies and discussions.

In discussing the signal-to-noise advantage of a MUSA it is understood that a reference receiving system must be adopted, and for this purpose one of the unit antennas connected to an automatic gain controlled receiver was chosen. Other types of antennas as, for instance, a simple vertical or horizontal doublet might have been used but other factors not significant to the MUSA would then have been involved.

Simple Analysis of the Signal-to-Noise Ratio Improvement

The MUSA differs from other directional antennas in that it is an array of antennas between which there is negligible electromagnetic coupling. This allows (but does not require) a different point of view, not explicitly involving directivity, in considering the signal-to-noise advantage of the array. The following analysis is made from this point of view. In Figs. 32 to 34 antennas are represented by signal generators, e_s, static generators, e_a, and resistances R_A. The input circuits of the receivers are matched to the antennas.

In Fig. 32, N spaced antennas are shown connected in parallel. The root-mean-square noise voltage, E_n, at the input to the receivers represents the thermal noise originating in the receiver input circuits.



FIRST ASSUMPTION:
ALL LINES MATCHED
(LINE IMPEDANCES = R_A
PHASE SHIFTER IMP. = R_A
INPUT CIRCUIT MATCHED
TO IMP. R_A/N)

$$I_s = \sum_{K=1}^{K=N} \frac{e_{sK}}{2R_A}, \quad E_s = I_s\sqrt{\frac{RR_A}{N}},$$

$$I_a = \sum_{K=1}^{K=N} \frac{e_{aK}}{2R_A}, \quad E_a = I_a\sqrt{\frac{RR_A}{N}},$$

$$E_n = \text{CONST} \times \sqrt{R}$$

SECOND ASSUMPTION:
SINGLE WAVE. SIMILAR ANT.
($e_{s1} = e_{s2} \cdots\cdots e_{sN} = e_s$
$e_{a1} = e_{a2} \cdots\cdots e_{aN} = e_a$)
SIGNAL CURR. PHASED
STATIC CURR. RANDOM

$$I_s = \frac{N}{2R_a}e_s, \quad E_s = \frac{1}{2}e_s\sqrt{N}\sqrt{\frac{R}{R_a}}$$

$$I_a = \frac{\sqrt{N}}{2R_a}e_a, \quad E_a = \frac{1}{2}e_s\sqrt{\frac{R}{R_A}}$$

$$E_n = \text{CONST.} \times \sqrt{R}$$

$$\frac{E_s}{E_a} = \sqrt{N}\frac{e_s}{e_a}$$

$$\frac{E_s}{E_n} = \sqrt{N}\frac{e_s}{\sqrt{R_A}}\,\text{CONST.}$$

Fig. 32—Simple signal-to-noise analysis of a system of N spaced antennas. Signal currents are phased and combined at the incoming frequency. The summation signs include addition on the power basis.

For the matched condition this noise is constant and independent of the number of antennas. A single wave is assumed and the signal outputs of the antennas are phased by means of the phase shifters ϕ. The maximum signal power obtainable from N antennas obviously is N times that obtainable from one antenna. In terms of receiver noise,

e_n, the improvement in signal-to-noise ratio is 10 log N decibels referred to one antenna. If, instead of receiver noise, static is the predominating noise, the signal power received is not significant but the same improvement is realized for the general case in which the static is distributed randomly among the N antennas.[22] In that case the N signals are phased to add on a current basis while the N noise sources

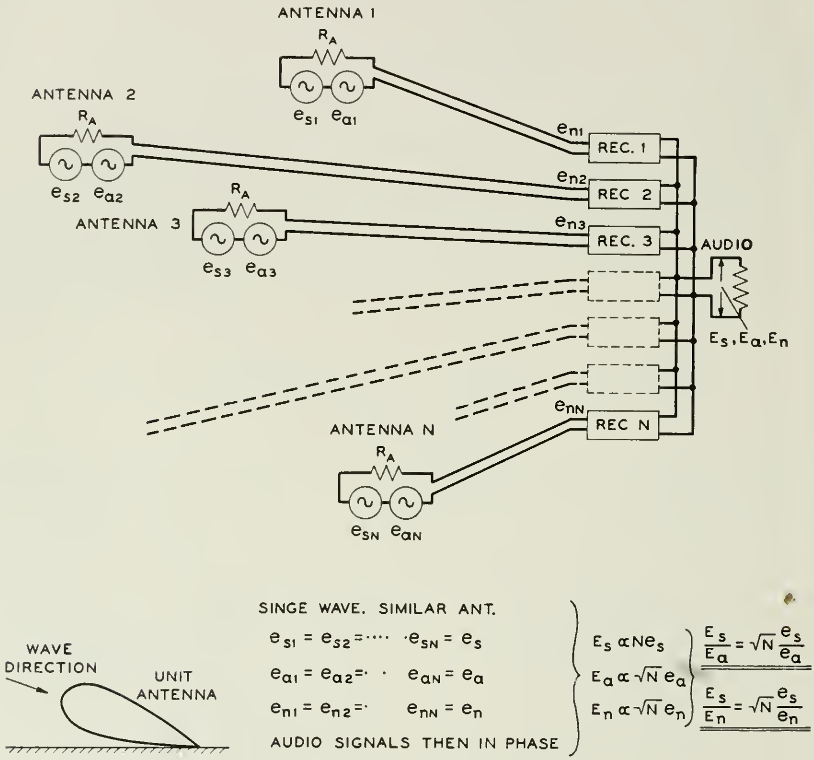


Fig. 33—Simple signal-to-noise analysis of a system of N spaced antennas. Signal currents are combined at audio frequency.

add on a power basis. Analogous arguments apply to a series connection of N antennas and result in the same improvement of 10 log N decibels.

[22] If static comes from all directions simultaneously, its distribution is random among the ideal unit antennas discussed in Section II. This is deduced from calculations which show that gain (signal-to-noise ratio) is proportional to the length of the system; i.e., to the number of unit antennas. The assumption of randomness requires that the spacing of unit antennas having a certain angular discrimination must be equal to or greater than the antenna length required to produce that discrimination in the simple linear end-on type of unit antenna.

That static is, on the average, distributed randomly among the rhombic antennas of the experimental MUSA is shown by measurements described later in this section.

The system described above has been shown mainly to introduce the system shown in Fig. 33. This diagram shows the audio addition of the outputs of N receivers fed by N antennas. Note that this system has no high-frequency phase shifters in the transmission lines. It is in fact similar to the diversity receiving system described by H. H. Beverage and H. O. Peterson.[23] For a *single wave* this is seen to be equivalent to the phased addition at carrier frequency shown in Fig. 32.

The signal-to-noise improvements shown on Figs. 32 and 33 were easily calculated because a single non-fading wave was assumed. In actual practice several fading waves are involved and it is then difficult, if not impossible, to make significant calculations. Later in this section, however, some of the general features of the system shown in Fig. 33 will be discussed from the point of view of several waves.

The MUSA system is characterized by the ability to separate waves and it is therefore possible to analyze it in a simple manner for cases of more than one wave. The arrangement in Fig. 34 corresponds to the Holmdel MUSA. The signals from the equally spaced antennas are here phased at the intermediate frequencies. Since random static and first circuit noise give identical results the analysis is given for static only.

As shown in Case I, if only one wave is present and both branches are phased for it the system functions as in Fig. 33 and it yields the same improvement of 10 log N decibels. If as shown in Case II the second branch is not phased for it (i.e., if the wave falls upon a minor lobe or a minimum of the MUSA directional pattern) less than the full improvement occurs. On the basis of linear audio detectors the reduction of improvement is 20 log x where x lies between 2 and $\sqrt{2}$. This quantity refers to the manner in which the noise from sources 1, 2, $\cdots$ N in Branch A adds with the noise *from the same sources* after having been phased differently and perhaps delayed differently in Branch B.[24] This involves the audio-frequency band width and method of noise measurement. As will be shown later x is usually not much different from $\sqrt{2}$. Taking $x = \sqrt{2}$ the loss in Case II is three decibels. If an audio detector is used which does not demodulate noise when the signal is absent (a square-law detector accomplishes this for practical purposes) this loss disappears, and branches may be phased for temporarily non-existent waves without incurring a penalty. Case III is the important one. It assumes two equal waves. Branch A is

[23] "Diversity Receiving System of RCA Communications, Inc.," *Proc. I. R. E.*, vol. 19, pp. 531–561, April, 1931.

[24] The case of $x = 2$ (in-phase addition) arises only when the phasing and delay of the two branches are alike.

phased for one; Branch B for the other. Again taking $x = \sqrt{2}$, the improvement referred to e_s/e_a is $10 \log N + 3$ decibels. Here e_s/e_a denotes the signal-to-noise ratio in each antenna due to one wave. Referring the improvement to the signal-to-noise ratio of one antenna

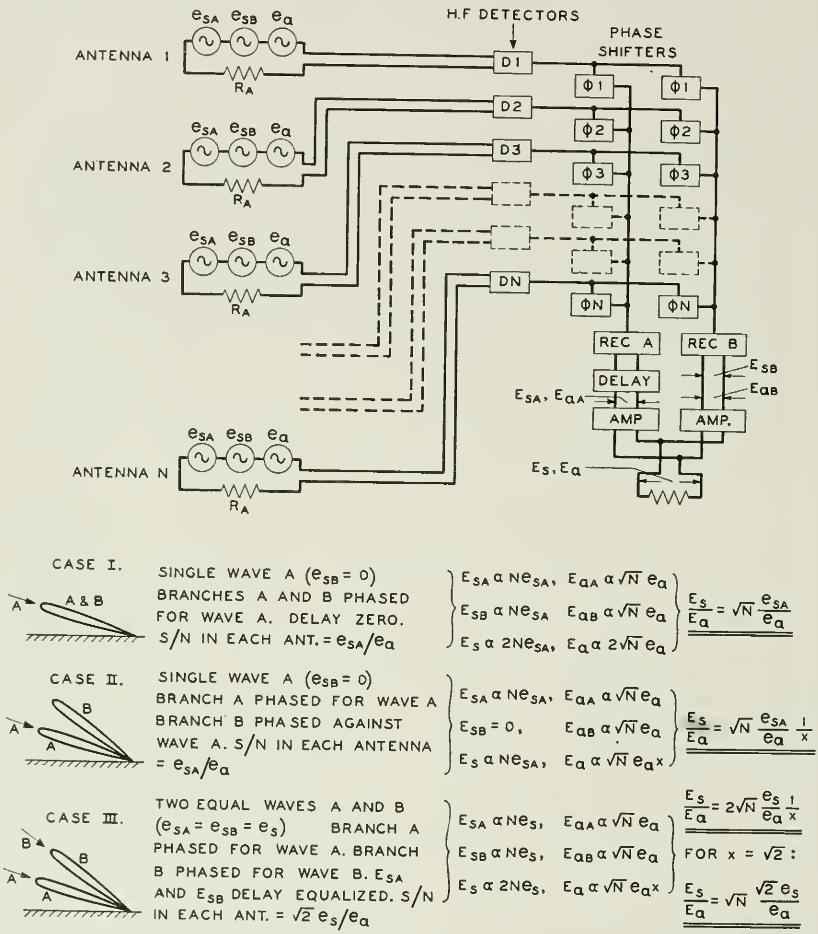


Fig. 34—Simple signal-to-noise analysis of the MUSA system.

receiving both waves, assumed to add randomly, increases the reference by three decibels and reduces the improvement to 10 log N decibels.

This analysis may be expanded to include K waves in which case K *branches* would be required to obtain the gain of *10 log N decibels* referred to one unit antenna.

It has been tacitly assumed in the foregoing analysis of Fig. 34 that the audio outputs of the several branches are delay equalized to add and that there is no diversity action (all of the waves are assumed to remain equal). The influence of fading is difficult to predict and will be discussed later in connection with experimental results.

Some readers, not concerned with details, may omit reading the following subsections and find it sufficient to read only the Summarizing Discussion of this section.

Test Method

From a practical point of view the best way of testing a MUSA system would seem to be to operate it on transatlantic telephone signals and compare its output with that of the reference system. Speech volume and noise could then be measured in the conventional manner. So far as the signal-to-noise improvement is concerned it would be a laborious and lengthy task to get satisfactory data because so often, during the test period,[25] static and receiver noise is masked by transmitted noise, interfering signals, and other man-made noise. To test the experimental MUSA, therefore, a different method was selected which gave significant data in a shorter time.

Since the success of the MUSA is related so fundamentally to the nature of the arriving signal the important thing to be determined by the measurements is how well the MUSA is able to cope with the various conditions of wave arrival. For instance, in the case of a single bundle of arriving waves how close does the actual signal-to-noise improvement come to the 10 log N decibel calculated for Case I (Fig. 34) in which a single non-fading wave was assumed? Likewise, for the case of two-wave bundles do the calculations of Case III agree with measurements?

For these purposes the signal-to-noise measurements would have to be free from directional static, interference, and transmitted noise; otherwise the measured improvements would be distorted. To insure uniform and desirable noise conditions it was decided to use thermal noise originating in the receiver input circuits [26] instead of whatever noise might be present on the radio channel. This was accomplished by inserting resistance pads in the antenna transmission lines to reduce the signal (and external noise) to a level where thermal noise greatly exceeded other noise. Signal-to-noise ratios in the range between fifteen and forty decibels were obtained in this manner, free of interference and directional static, and of transmitted noise.

[25] Transmission conditions during 1935 were comparatively undisturbed.

[26] A portion of the noise originates in the plate circuit of the first detector. For the present purposes this is equivalent to first circuit noise.

Substituting thermal noise for external static may at first seem far-fetched. Except for the fact that static is sometimes sufficiently directional to be received with different intensity as the MUSA is steered differently, the substitution is sound. In general, the static output does not vary with steering, as the measurements described later indicate but to avoid the distortion of results which would occur when this is not so, it was desired specifically to substitute non-directional noise. Studies of the characteristics of static and thermal noise have shown that both are alike so far as the effect of band width upon average and effective values is concerned, and have indicated that both consist of extremely short, randomly distributed pulses which overlap when received and detected by receivers of ordinary band widths. In a given band width, the envelope of the currents produced by static sources is highly irregular in comparison with that produced by thermal agitation. It appears, however, that the *character* of either envelope is not sensibly affected by the number of antennas combined nor by the manner in which the branch outputs are combined, so that both give the same improvement figures using any arbitrary noise measuring method.

There were several possibilities with respect to the signal to be employed in these tests. A single tone, a large number of tones distributed throughout the audio band and other special signals were considered. A simple method requiring no modulation was finally adopted. It consisted in alternately connecting the output of the antenna to be tested and that of the reference antenna to the same receiver. Assuming that the automatic gain control of this receiver would maintain a constant audio output level the signal-to-noise advantage is the ratio of the noise levels. The automatic gain control of the MUSA receiver did not, of course, hold the output level absolutely constant but a correction was easily made for the small variations in level.

The circuits of the measuring equipment are shown in Fig. 35. The rectified carrier appearing in the linear speech rectifier is taken to be proportional to signal and is measured simultaneously with the noise demodulated in the rectifier. When the keys are thrown to position 2 (by a gang arrangement) the signal meter shows the sum of the two rectified carriers and the noise meter reads the combined noise in the output of the diversity mixing amplifiers. Using the sum of the two rectifier currents to represent the signal implies that actual audio outputs from the two branches could be delay equalized to add arithmetically. As applied to a MUSA system this assumption is justified, in general. When the keys are switched to position 1, the rectified carrier of branch B alone appears on the signal meter and noise from branch B appears alone on the noise meter. At the same time the diversity connection is broken and all except one of the six-phase shifter amplifiers in branch B are biased to cutoff; i.e., only one unit antenna is used. The pad "L" is adjusted to give the same audio gain from rectifier B to the noise meter for connection 1 as for connection 2.

By manipulating the keys which control the cutoff biases on the phase shifter tubes the "1" to "2" switchover may also be used to compare one antenna (one receiver) with two antennas in ordinary space diversity or one antenna with all six in a single branch.

The use of receiver noise as a noise source depends upon (1) having the noise equal in all six circuits and (2) upon having it originate ahead of the point where the gain is varied. In well-designed receivers the noise should approach the thermal noise limit of the first circuit. It was found possible to have the signal-to-noise ratio, for a given signal level, of all six high-frequency input circuits equal to within $\pm$ 0.5 decibel and within a few decibels of the thermal limit.

The first tests were made with a local oscillator supplying the signal. They really constituted tests of the measuring set up. All six input circuits were fed simultaneously through 80-ohm pads giving equiphase and equiamplitude signals on each detector grid. This corresponds to receiving a single steady wave, and one branch was "steered" as if to receive such a wave. When the multiple switch was manipulated as to compare one antenna with the steered branch the indicated signal-to-noise improvement was usually between seven and eight decibels, compared with the theoretical value of 7.8 decibels (10 log 6).

Such a local test using the switchover with all associated equipment was made before and after every transatlantic test. Corrections based upon 7.8 decibels were made to the data in cases where the local tests showed a slightly different improvement factor. In all of the work the gains of the phase shifters were adjusted to equalize the difference in line loss. The effect of this is, however, trivial.

In measuring on transatlantic waves with automatic gain control the noise variation, corresponding nearly to the reciprocal of fading, rendered visual noise readings too rough to be suitable. A Weston high-speed db meter (copper-oxide bridge type) having a calibrated range of 16 decibels was used as a noise meter. To this instrument was added a fluxmeter (Fig. 35) of low restoring torque which automatically averaged the variations of the meter pointer over the 15-second periods of observation used in these tests. The fact that the noise meter rectifier is linear means that the noise current is averaged arithmetically

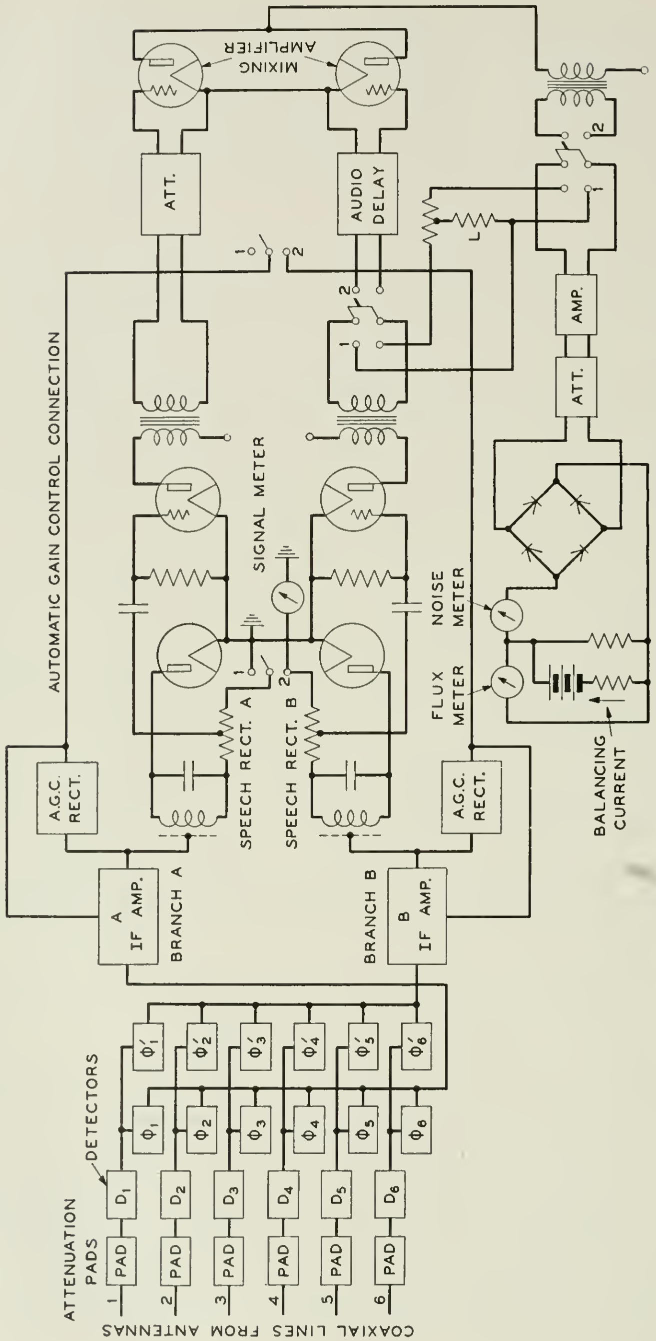


Fig. 35—Circuit used in measuring the signal-to-noise improvement of the experimental MUSA system.

along the time axis. (A discussion of other ways of averaging will be given later.) To permit the maximum time interval during which the restoring torque had negligible effect a balancing current was applied. A telechron motor marked time intervals of 15 seconds with a bell. The switchover between "1" and "2" (Fig. 35) was made and the fluxmeter restored to zero reading at the ring of the bell. The signal meter reading (calibrated in decibels) was maintained fairly steady with the automatic gain control, and could be averaged accurately by eye.

Test Results

The first measurements on transatlantic signals were made on 16 meters in September, 1935. Tests were made on the British Post Office Station GAU (18,620 kilocycles) between noon and 2 P.M., E.D.S.T., on September 16, 17, 18, and 19. At our request the transmitter operated on reduced power, presumably fifteen decibels less than normal. The propagation during these tests was characterized by low angles and slow fading. The angular discrimination of the MUSA for these low angles is so slight that it was decided to use only one branch and defer the question of diversity in these first tests.

Owing to the power reduction and to the fact that the period from September 15 to 18 inclusive was disturbed some of the data had to be obtained without antenna pads. Table II summarizes the results of the measurements.

When thermal noise is a contributing or predominating factor, the signal-to-noise ratio of the six-antenna branch must be compared with both No. 1 and No. 6 antennas in order that the line loss may be accounted for. With thermal noise predominating, the difference between the improvements referred to No. 1 and No. 6 should of course equal the line loss of No. 6. It may be shown that the arithmetic means of the two improvement ratios (voltage) should give very closely the improvement corresponding to random static (10 log 6 = 7.8 decibels). In Table II the arithmetic means of the improvement ratios are, therefore, called the equivalent improvement.

During these tests the indicated angle of arrival was from one to three degrees (such low angle determinations are not trustworthy within perhaps two degrees) and the receiving branch was set correspondingly and not altered during a test. The fading on No. 1 and No. 6 antennas was usually but not always unlike. Adjacent antennas always showed substantially synchronous fading.

A sample of the plots from which the figures in the table were obtained is shown in Fig. 36. This shows the noise readings reduced to a constant signal meter reading for Test 13. Each horizontal line segment represents one 15-second fluxmeter period (the actual period was 14 seconds, one second being required for switching). The arithmetic averages of the segments are shown by dashed lines. It is to be emphasized that the scattering of improvements taken from adjacent readings is not experimental error but is due to the fact that pairs of adjacent readings were obtained at different stages of the fading cycle. Two separate systems (receivers and antennas and measuring equipment) permitting simultaneous measurements would not help this

TABLE II

One Branch

GAU 18,620 kilocycles

Date	Test No.	Pads in Ant. db	Reference Antenna	Number of Readings	S/N Improvement db	Group Average db
1935						
9–16. . . .	1	20	1	19	3.7	
9–17. . . .	6	20	1	18	3.0	
9–17. . . .	10	20	1	16	3.0	
						3.3
9–16. . . .	2	20	6	20	8.8	
9–17. . . .	7	20	6	20	11.0	
						10.0
9–16. . . .	4	0	1	17	6.2	
9–17. . . .	9	0	1	11	6.5	
9–19. . . .	11	0	1	18	6.0	
9–19. . . .	13	0	1	28	5.6	
9–19. . . .	15	0	1	20	8.0	
						6.4
9–16. . . .	5	0	6	30	6.4	
9–17. . . .	8	0	6	17	7.9	
9–19. . . .	12	0	6	20	5.0	
9–19. . . .	14	0	6	29	6.1	
						6.4

The apparent line loss is 6.7 db (4.3 calculated).
The equivalent improvement obtained from the data with pads is 7.3 db; without pads, 6.4 db. Average 6.9 or 7 db.

situation, since fading would not be synchronous on two systems. Evidently a considerable number of readings must be taken to insure that all stages of fading are equally represented in both of the systems being compared.

Returning to Table II the discrepancy between the apparent line loss and the calculated value (the loss could really be five decibels perhaps) suggests that insufficient data were obtained with the pads. There is a possibility that the measurements without pads involve directional static. However, taking the data as they stand yields an average of seven decibels which is less than one decibel below the value 10 log N

= 7.8 decibels calculated for the non-fading wave of Case I, Fig. 34. (The fact that only one branch was used in the measurements instead of the two branches of Case I does not affect the situation since the idle branch was disconnected and so contributed no signal and no noise.) A reduction of the order of one decibel from the calculated improvement could be expected since all of the waves of one bundle

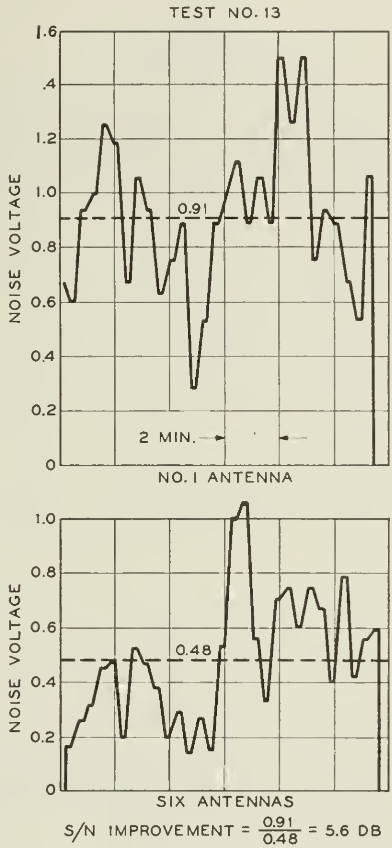


Fig. 36—Sample of measured noise plotted against time. GAU (18,620 kilocycles) Rugby, September 19, 1935.

cannot fall on the apex of the directional pattern. It was encouraging to find that no more loss occurred; i.e., to find that the waves in a single bundle may be phased so effectively.

Before leaving these tests, the results for September 18 should be mentioned. On this day the signal-to-noise ratio was so low, even without antenna pads, that measurements could not be made. The noise on this day was first taken to be thermal noise but was found during the course of experimentation to be external noise [27] some ten decibels higher than thermal noise, as received on a single rhombus. At the end of the test the operator at Rugby keyed the transmitter with tone, advising us that the schedule was completed and wishing us "good night." With one antenna the signal was hopelessly lost in noise; with the six antennas the code was readable.

TABLE III
Two Branches
GBW 14,440 kilocycles

Date	Test No.	Pads in Ant. db	Reference Antenna	Number of Readings	S/N Improvement db	Group Average db	Number of Wave Bundles
1935							
10–2 . . .	21	40	1	20	2.7		1
10–2 . . .	23	40	1	16	5.3		1
10–2 . . .	24	40	1	16	4.4		1
10–8 . . .	29	40	1	20	6.3		1
10–10 . . .	37	40	1	20	4.4		1
						4.7	
10–2 . . .	20	40	6	19	6.9		1
10–2 . . .	22	40	6	20	8.4		1
10–2 . . .	25	40	6	16	9.0		1
10–8 . . .	28	40	6	17	10.9		1
10–9 . . .	30	40	6	38	8.0		
						8.5	
9–30 . . .	17	40	1	22	3.4		2
9–30 . . .	19	40	1	14	4.1		2
10–2 . . .	26	40	1	16	4.6		2
10–10 . . .	33	40	1	19	2.0		2
						3.5	
9–30 . . .	16	40	6	17	8.0		2
9–30 . . .	18	40	6	17	7.9		2
10–2 . . .	27	40	6	20	6.7		2
10–10 . . .	34	40	6	20	7.3		2
						7.4	

The apparent line loss is 3.8 and 3.9 db for the one-wave and two-wave groups, respectively. The calculated loss is 3.8 db.

The equivalent improvement for one-wave group is 6.8 db to which may be added the later determined correction of 1.2 db for the effect of delay, giving *8.0 db.*

The equivalent improvement for two-wave groups is 5.7 db to which may be added 0.8 db for the effect of delay, giving *6.5 db.*

More comprehensive measurements were made on GBW (14,440 kilocycles) and a few on GCW (9790 kilocycles), using the two branches. Since an unmodulated carrier was used, rectified carrier being taken to represent signal, there was no criterion for setting the audio delay. Accordingly, it was kept at zero and a correction introduced later. The results are shown in Tables III and IV.

[27] This noise, which was directive to the extent that four-decibel variation occurred with steering the MUSA, was doubtless a sample of the "star static." It was encountered also on 31 meters in October. See footnote (32).

TABLE IV

Two Branches

GCW 9790 kilocycles

Date	Test No.	Pads in Ant. db	Reference Antenna	Number of Readings	S/N Improvement db	Group Average db	Number of Wave Bundles
1935							
12–13...	39	40	1	15	4.8		2
12–13...	40	40	1	14	4.7		2
						4.7	
12–13...	38	40	6	14	7.9		2
12–13...	41	40	6	13	6.7		2
						7.3	

The apparent line loss is 2.6 db. The calculated loss is 3.1 db.

The equivalent improvement is 6.1 db to which may be added 0.9 db for the effect of delay, giving *7.0 db.*

The data in the tables are classified roughly according to whether two bundles or one bundle of waves was present. In the latter case the two branches were steered, one on each side of the bundle, a few degrees apart. During these tests slight adjustments in steering were made when indicated by the angle monitoring tube, as in normal operation of the system. The large amount of data taken with GBW makes the results in Table III particularly reliable. This is reflected in the close agreement between measured and calculated line loss. Before discussing the results further the effect of delay needs to be analyzed.

Correction Due to Delay

The effect upon the noise, of delaying one audio output, is shown in Fig. 37. The curves were obtained with the circuit shown in Fig. 35

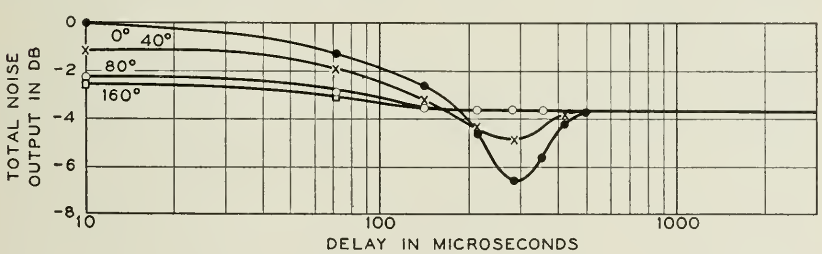


Fig. 37—The effect of delay and phasing upon noise output of the MUSA receiver.

using thermal noise. The equiphase, equiamplitude signal source mentioned previously was used to supply the inputs. Branch A was kept phased so that the six signals added, and branch B was varied. The curves show the effect of delay upon noise in a 250- to 2750-cycle audio band, as measured with the Weston db meter, for various differences in steering. The curves are labeled in terms of the difference in the phase settings of the two branches. The 40-degree or 80-degree curves correspond in practice to steering on each side of a single bundle of waves. The 160-degree curve typifies steering at two separate wave bundles. The use of 100 microseconds (or more) delay is generally advantageous for the audio addition when steering at one bundle. Since this amount of delay makes the audio noise addition nearly random and for widely different steering the addition is also random the assumption that the noise from the two branches adds on a power basis, made in reference to Fig. 34 ($x = \sqrt{2}$), is justified.

The effect of delay is to produce an interference pattern in the audio noise spectrum. This accounts for the dip in the noise curve for a delay of 300 microseconds which locates the first interference minimum at about the center of the audio band. The asymptotic approach to 3.5-decibel reduction corresponds more nearly to a reduction ratio of $2/\pi$ than to $1/\sqrt{2}$ due to the fact that the Weston db meter is nearer linear than square law in response.

In obtaining these curves it was desired to simulate the reception of two waves for which the corresponding branches were phased to add. It was not convenient to set up locally such a two-wave case but the single wave input should give identically the same results provided phases were avoided which resulted in a signal at the second detector too low to demodulate the noise. A signal level so high that further increase did not affect the noise output was used for all points. The real purpose of the signal was to insure that the demodulated noise was not dependent upon the intermediate-frequency bands and that the results would be unaffected by possible differences in intermediate-frequency bands.[28]

As mentioned, noise has been measured with an unweighted 250- to 2750-cycle frequency band. Had a weighting network [29] which emphasizes frequencies in the vicinity of 1000 cycles been used the dips in the curves marked 0° and 40° would have been deeper and would have occurred in the vicinity of 500 microseconds delay.

Returning now to Tables III and IV the measured improvements were corrected to correspond to the effect of the delay which would probably have been used to obtain the best audio addition for the signal. The 1.2-decibel correction for the one-bundle case represents

[28] This precaution was subsequently found to be unnecessary; i.e., similar results were obtained with no input signal.

[29] Barstow, Blye, and Kent, "Measurement of Telephone Noise and Power Wave Shape," *Elec. Engg.*, vol. 54, pp. 1307–1315, December, 1935. Technical Digest published in *Bell Sys. Tech. Jour.*, January, 1936.

the reduction of noise obtained with 60-degree phase difference with the addition of 100 or 150 microseconds delay. The 0.8- or 0.9-decibel correction for the two-bundle case corresponds to any large delay and a large phase difference, say 160 degrees. These corrections would not have been much different if a weighting network had been used.

The measured difference of 1.5 decibels, shown in Table III, between the one-wave bundle and the two-wave bundle measurements is probably real and due to the fact that when the branches are steered at two separated bundles some loss is incurred when one wave disappears for a few minutes. This loss could have been at least partly recovered by using square-law detectors. Tests showing the advantage of square-law detectors over linear detectors are described later in this section. Employing square-law detectors would justify a correction of about one decibel to be added to the two-bundle improvement measurements in Tables III and IV. Applying this correction we summarize the results in Table V.

TABLE V

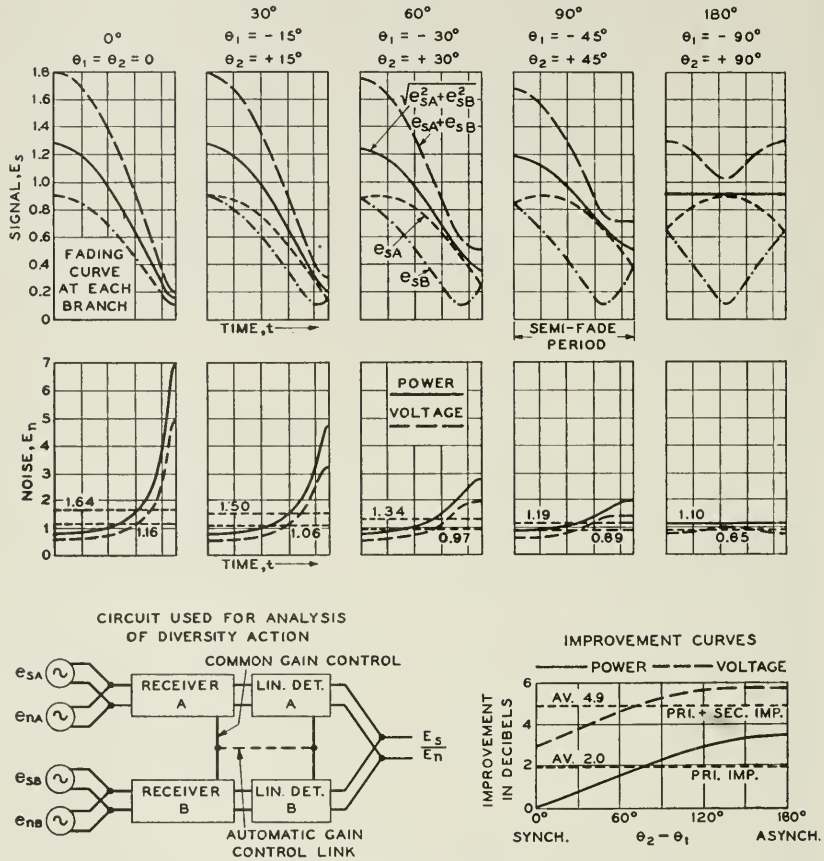
Summary of Signal-to-Noise Measurements

One Bundle of Waves		*Two Bundles of Waves*
One Branch only	*Two Branches*	*Two Branches*
7 db (Table II)	8 db (Table III)	7.5–8 db (Tables III and IV)

These improvement figures for two branches as they stand are approximately equal to 10 log 6 = 7.8 decibels as calculated for non-fading waves, and leave nothing to be ascribed to diversity action. Since a loss of perhaps one decibel occurs in the case of one branch (Table II), the recovery of that one decibel with two branches is to be ascribed to diversity action. Originally, considerably more was expected of the angle diversity. It appears however from theoretical and experimental evidence that one decibel is about what should be expected for the case in hand. It seems appropriate to include this study of diversity here.

Diversity Action

The first attempt to analyze diversity action was made with a graphical approach to the problem. On Fig. 38 is shown a schematic diagram of the system to be analyzed. Two receivers A and B with linear audio detectors may be regarded as fed from two angle branches of a MUSA. The noise generators e_{nA} and e_{nB} are assumed to be of equal power but of random phase. The signal generators e_{sA} and e_{sB} are assumed to fade according to the equations shown beneath the diagram. They represent the carrier amplitude but since fading is assumed to be essentially non-selective in each branch they also repre-



$$e_{sA} = \sqrt{1 + 2 \times 0.8 \cos(2\pi t/T + \theta_1) + (0.8)^2}$$

$$e_{sB} = \sqrt{1 + 2 \times 0.8 \cos(2\pi t/T + \theta_2) + (0.8)^2}$$

Fig. 38—Graphical analysis of diversity action as it relates to signal-to-noise ratio.

sent the side bands. This type of fading might result from interference between two waves of small relative delay whose amplitude ratio is 0.8, such a pair being received by each branch. The automatic gain control is assumed to be perfect; i.e., the audio output, E_S, is maintained constant.

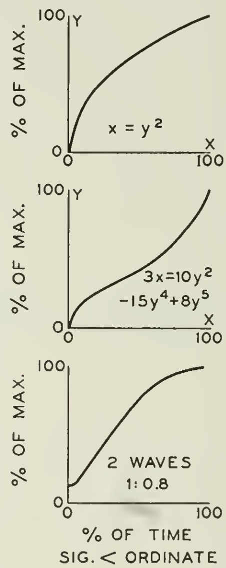
By definition, diversity in fading occurs when the fading of the two branches is not synchronous. If all degrees of asynchronism are equally probable the diversity is random. This is the case considered. In Fig. 38 five stages in the cycle of variation from synchronous to asynchronous fading are used for calculation. In each of these, fading curves corresponding to the two assumed waves are shown displaced from each other by 0, 30, 60, 90, and 180 degrees. With "ideal" automatic gain control, the two receiver gains, always equal, will be proportional to the reciprocal of the resultant of e_{SA} and e_{SB}. For "ideal" linear detectors the noise output of the receivers will be proportional only to the gain. The noise curves plot the noise variation on this basis. Two cases of signal addition are considered—voltage and power addition. The corresponding noise curves differ only as the reciprocals of the resultant signal curves differ. These noise curves are averaged (with a planimeter) and the resulting average signal-to-noise ratios are used to plot the improvement curves shown in the figure. The improvement curve for power addition of signal is located on the improvement axis so that zero improvement is shown for synchronous fading. The curve for voltage addition of signal is located three decibels higher at synchronous fading. These curves are again averaged over the cycle from synchronism to asynchronism (by averaging noise voltage). The improvements are 2.0 decibels and 4.9 decibels.

Power addition of the two signals corresponds in practice to the case in which the delay is unequalized and sufficient to cause the audio outputs of each branch to combine on a power basis like noise. The two-decibel improvement might appropriately be called the primary improvement since it is due solely to the diversified fading. The additional improvement of 2.9 decibels found with voltage addition of the signals is due to favorable discrimination in the addition of signal and of noise and might be called the secondary improvement. The secondary improvement occurs in reception with the MUSA; it has already been included in the 10 log N decibel improvement calculation.

In practice, it would be undesirable to use the "ideal" automatic gain control assumed in the above analysis; the action must be smoothed out with, for instance, a capacitance-resistance network. The effect of this is to reduce the primary gain since the noise peaks, whose avoidance by diversity action results in the primary improvement, are reduced. An analysis of diversity action without automatic gain control was made. In this case the signal was averaged while the noise remained constant. The results are included in the table shown in Fig. 39 which is introduced later.

This treatment of diversity action has been made from the point of view of MUSA reception but is applicable to ordinary space diversity using two stations or antennas. In the case of a single bundle of waves no modification of the analysis need be made; the signal generators then represent the spaced antenna outputs which are fading randomly. In this case voltage addition of the audio outputs may be

NO. OF STATIONS OR BRANCHES	$E_s = \sqrt{e_{sA}^2 + \cdots e_{sK}^2}$ $E_n = \sqrt{e_{nA}^2 + \cdots e_{nK}^2}$ PRIMARY IMPROVEMENT		$E_s = e_{sA} + \cdots e_{sK}$ $E_n = \sqrt{e_{nA}^2 + \cdots e_{nK}^2}$ PRI. + SEC. IMPROVEMENT		ASSUMED DISTRIBUTION CURVES
K	NO AGC	AGC*	NO AGC	AGC	
QUADRATIC DISTRIBUTION					
1	2.5	0	2.5	0	
2	2.8	2.2	5.5	4.8	
3	2.9	2.5	7.3	6.8	
FOUR WAVE DISTRIBUTION					
1	3.4	0	3.4	0	
2	—	—	6.4	5.2	
3	—	—	8.2	7.5	
TWO WAVE NEAR-SINUSOIDAL DISTRIBUTION 19 DECIBEL FADE					
1	2.8	0	2.8	0	
2	3.2	2 0	5.8	4.9	
3	—	—	—	—	
AVERAGED RESULTS					
1	3	0	3	0	
2	3	2	6	5	
3	3	2.5	8	7	



* AUTOMATIC GAIN CONTROL

Fig. 39—Summary of results of diversity analysis.

expected to occur since a single bundle will typically include only a small delay interval. In the general case of two or more wave bundles the signal generators must be interpreted to represent not the carrier but the side-band average, for fading will then be essentially selective and the audio output will not be proportional to carrier as was assumed in the analysis. If this interpretation is made the signal-to-noise ratio in each receiver becomes proportional to the generator amplitudes e_{sA}

and e_{sB}, and the analysis of Fig. 38 is applicable. In this case voltage addition does not occur since the audio outputs are essentially different owing to the selective fading.[30] They add, in general, to a value intermediate between the power and voltage sum, although for the more complicated conditions they combine on a power basis.

The above analysis has been based upon simple two-wave interference and the results might not be applicable to the more complicated and changing conditions of actual transmission. Accordingly, R. L. Dietzold has made a statistical analysis for other types of fading and for three stations as well as for two. The results appear in Fig. 39 together with the results of the above graphical analysis for two-wave interference fading. The time sequence of amplitude in the more complicated types of fading encountered in practice is not significant; the percentage distribution determines the results. The "four-wave" distribution curve corresponds to four equal waves of random phase. The quadratic distribution curve was deduced experimentally by R. S. Ohl. Except that these different distributions were assumed, the assumptions were the same as those of Fig. 38. The improvements are expressed in decibels referred to the signal-to-noise ratio for one station or branch with ideal automatic gain control.

The small effect upon the results of assuming different time distributions lends significance to these calculations. The averaged round numbers are probably about right.

With no automatic gain control (or with one which acts slowly compared with the fading) there is little or no primary gain. With infinitely fast and stiff gain control action there is a 2- and 2.5-decibel primary gain for two and three stations, respectively.

A few measurements were made at Holmdel on two-station diversity (antennas 1 and 6 of the MUSA). The thermal noise and rectified carrier technique was used. The results appear in Table VI.

The measuring technique was exactly the same as used in obtaining the data for Tables III and IV in which voltage addition of the signal was assumed. A 3-decibel improvement is therefore included in the 3.6-decibel figure. This leaves only 0.6 decibel (possibly one decibel or even 1.5 decibels since the measurements are too meager to be reliable to better than one decibel) for primary gain compared with a possible 2.0 decibels. We are inclined to use about one decibel for primary gain. The time constant on the automatic gain control was of the order of 0.06 second in this and all signal-to-noise comparisons. That this time constant was not fast enough to produce the high noise

[30] This refers to speech signals; in the case of telegraph signals the frequency band is so narrow that fading is always essentially non-selective, and voltage addition occurs.

peaks corresponding to the inverse of fading is shown by the transcribed motion picture record of the signal and noise meter variations, shown in Fig. 40. Note the signal fades.

A secondary improvement of 3 decibels is too high for two-station space diversity; i.e., the signals do not add on a voltage basis.[30]

TABLE VI

ANTENNAS 1 AND 6 IN DIVERSITY
GBW 14,440 kilocycles

Date	Test No.	Pads in Antennas db	Reference Antenna	No. of Readings	S/N Improvement db	Average
1935						
10–9 . . .	32	40	1	19	1.7	
10–10 . . .	35	40	1	20	2.2	
						2.0
10–9 . . .	31	40	6	14	4.9	
10–10 . . .	36	40	6	20	5.2	
						5.1

The apparent line loss is 3.1 db. The calculated loss is 3.8 db.
The equivalent improvement is 3.6 db.

Oscilloscope observations of the diversity combinations of the audio outputs of two spaced antennas (No. 1 and No. 6 of the Holmdel MUSA) indicate, however, that on the average the secondary improvement is appreciable and probably about 2 decibels for two antennas and 3 decibels for three antennas. This improvement depends upon

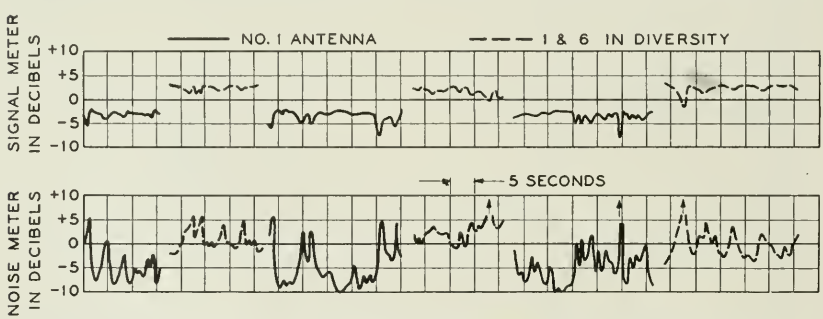


Fig. 40—Sample of signal and noise variations occurring in diversity tests. The arrows indicate noise levels beyond the scale of the meter. GBW (14,440 kilocycles) Rugby, October 10, 1935. 1920 G.M.T.

the number of wave bundles, their angular separation, their relative delays, the spacing of the antennas and the frequency band occupied by the signal. For a single compact wave bundle the secondary improvements will be nearly 3 decibels and 4.8 decibels for two- and three-antenna systems, respectively, but for several bundles of large relative delay the secondary improvement may disappear.

The results of some recent tests of a three-antenna diversity system on trial at Netcong, N. J., carried out under the direction of F. A. Polkinghorn, showed a signal-to-noise improvement of 3 to 3.5 decibels. Assuming a 3-decibel secondary improvement there remains something of the order of 0.5 decibel for the primary improvement. This is plausible in view of the time constant of one second used on the automatic gain controls. Linear audio detectors were used in these tests. As will be discussed later the employment of square-law detectors could be expected to add 0.5 decibel to this figure. Linear detectors are to be preferred, however, on the basis of quality distortion. Table VII is based upon the theoretical and experimental study of diversity action and gives typical results for space diversity systems.

TABLE VII

Summary of Space Diversity Improvements

Number of Antennas	Primary Improvement in db Automatic Gain Control		Secondary Improvements in db Number of Wave Bundles		
	0.06 Sec.	1 Sec.	1	2	3–5
2	1	0.5	3	2	1
3	1	0.5	4.5	3	1.5

Add 0.5 db to primary improvement when square-law detectors are used.

Table VII shows that on the average the secondary improvement is larger than the primary improvement. In other words, the advantage which accrues from the similarity of the antenna outputs exceeds that which accrues from their diversification. This result had not been expected.

It should be emphasized here that the improvements summarized in Table VII for space diversity systems and in Table V for a MUSA system refer to signal-to-noise ratios only; i.e., quality improvement is not included.

An important advantage of a MUSA system over a space diversity system is its ability to maintain its improvement when more than one wave bundle occur, and since two or more bundles are common, the advantage is distinctly real. A further advantage not discussed thus far relates to interfering signals as distinguished from static. Unless the interfering signals fall upon the principal lobe of the MUSA array pattern when it is steered to receive the desired signal, important directional discrimination against the interference occurs. Little or no discrimination against interference can occur in a space diversity system since it lacks the phasing which produces directional discrimination.

The Time Constant of the Automatic Gain Control

Thus far no comments have been made on the improvement figures relating to "no automatic gain control" shown in the table of Fig. 39. This table shows that the signal-to-noise ratio for one antenna ($K = 1$) is from 2.5 to 3.5 decibels higher when no automatic gain control is used; i.e., perfect automatic gain control penalizes the signal-to-noise ratio to that extent.[31] The advantage of automatic gain control is a constant output volume. In practice, a compromise is effected by retarding the action of the control. A time constant of 0.5 or one second is usually used. (This compromise is influenced by quality considerations as well as noise considerations.)

In the MUSA system signal-to-noise ratio measurements the time constant of the automatic gain control circuit (0.06 second) was not changed during the switchover from the MUSA to the single antenna. If a time constant of 0.5 second had been used with the MUSA and a one-second time constant with the reference receiver, the measured improvement would probably have been reduced by a little less than one decibel.

Method of Averaging Noise

In all of the signal-to-noise measurements and in the diversity analysis noise voltage has been averaged arithmetically along the time axis. Owing to a rather marked reduction of noise peaks with the MUSA compared with a unit antenna different improvements would result if different ways of measuring it had been adopted. To investigate this, motion pictures were made of the signal meter and noise meter variations for the MUSA and for the single antenna. The transcribed records appear in Fig. 41. Some calculations have been carried out for the noise distributions marked A, B, and C in Fig. 41. If the noise ratio of B/A measured by arithmetically averaging noise voltage is called 0 decibels, it becomes $+ 2.4$ decibels by averaging power arithmetically. The corresponding figures for B/C are 0 and $+ 2.7$ decibels. Thus, if noise power is averaged instead of noise voltage the measured primary diversity improvement is substantially increased.

[31] The action of the automatic gain control does not change the instantaneous signal-to-noise ratio. Interpreting signal-to-noise ratio as *average signal* divided by *average noise* rather than the average of the *signal divided by the noise* results in this difference.

From the point of view of the interfering effect upon speech it is not clear which method of averaging is more significant. This matter is probably related too closely to the distortion which incidentally accompanies the noise peaks to be considered alone.

In the light of the discussion presented in the preceding pages it appears that the signal-to-noise improvement of the experimental MUSA can be expressed as 8 ± 1 decibels.

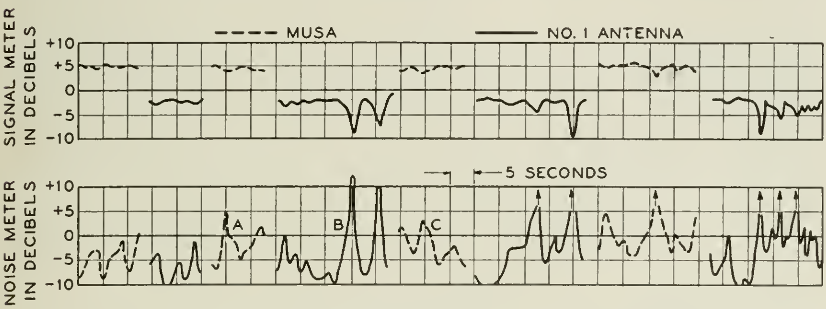


Fig. 41—Sample of signal and noise variations occurring in MUSA tests. The arrows indicate noise levels beyond the scale of the meter. This record was obtained five minutes before that in Fig. 40.

Square-Law Detectors

In the discussion of the signal-to-noise measurements it was stated that the measured improvements would have been higher had square-law audio detectors been used instead of linear rectifiers. For the two-bundle MUSA measurements one decibel was allowed for this and for the three-antenna diversity measurements at Netcong 0.5 decibel was allowed. These figures are based upon tests to be described in the following paragraphs. First, however, an analysis will be made of the effect upon the signal-to-noise ratio of various types of detectors in a MUSA system. Figure 42 is a schematic representation of the system to be analyzed, comprising K branches. The K signal generators e_{sA}, e_{sB}, $\cdots e_{sK}$ represent the various wave bundles as received by the steerable branches. The noise generators e_{nA}, e_{nB}, $\cdots e_{nK}$ are equal in amplitude but random in phase. The detectors are generalized to the extent that the audio output is proportional to the u power of the input.

Assuming that $e_s \gg e_n$ the audio outputs are proportional to $e^u{}_{sA}$, $e^u{}_{sB}$, $\cdots e^u{}_{sK}$. The noise outputs are then proportional to $e_{nA}e_{sA}{}^{u-1}$, $e_{nB}e_{sB}{}^{u-1}$, $\cdots e_{nK}e_{sK}{}^{u-1}$ since the signal-to-noise ratio in each branch must be independent of u. Assuming the signals to be delay equalized,

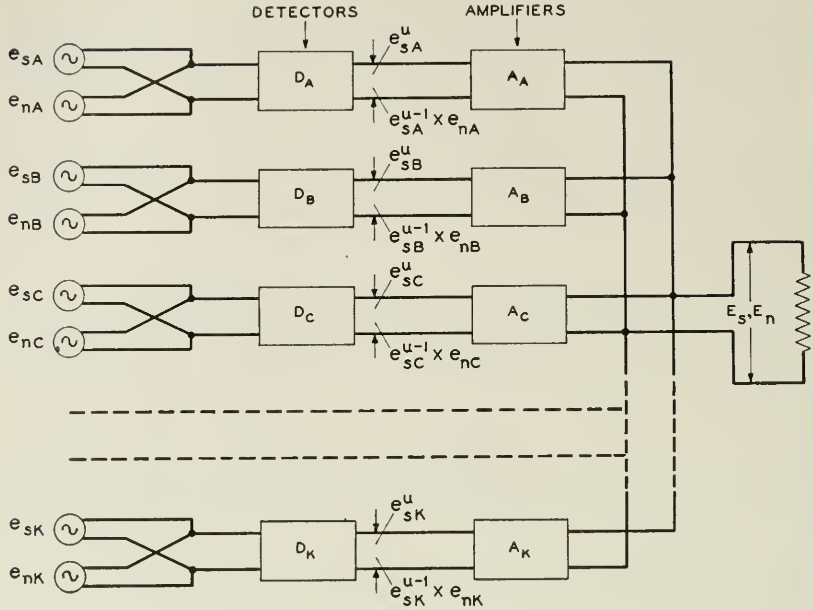


Fig. 42—Circuit employed in the analysis of the effect of detector characteristics upon signal-to-noise ratio.

the signal-to-noise ratio of the final output is

$$\frac{E_s}{E_n} = \frac{e^u{}_{sA} + e^u{}_{sB} + \cdots + e^u{}_{sK}}{e_n\sqrt{e_{sA}{}^{2u-2} + e_{sB}{}^{2u-2} + \cdots + e_{sK}{}^{2u-2}}}. \tag{7}$$

Maximizing this expression with respect to u shows that the maximum occurs for $u = 2$ (square law) and is

$$\underset{(u\,=\,2)}{\frac{E_s}{E_n}} = \frac{\sqrt{e^2{}_{sA} + e^2{}_{sB} + \cdots + e^2{}_{sK}}}{e_n}. \tag{8}$$

That this expression represents the maximum signal-to-noise ratio may also be concluded by observing that it is proportional to the square root of the total energy.

For linear detectors $u = 1$ and the signal-to-noise ratio becomes

$$\underset{(u\,=\,1)}{\frac{E_s}{E_n}} = \frac{e_{sA} + e_{sB} + \cdots + e_{sK}}{e_n\sqrt{K}}. \tag{9}$$

If the branch signals are all equal, i.e., if $e_{sA} = e_{sB} = \cdots = e_{sK}$, (8) and (9) give the same result, but for unequal amplitudes there is an advantage in using square-law detectors.

This analysis shows that square-law detection introduces just the correct amount of emphasis upon the stronger waves and that any additional expansion or contraction of the differences among the waves is detrimental. This means that the gains in all branches should be equal. It also indicates that any arrangement in which the stronger of the several waves is automatically switched in and the remaining ones switched out is inferior.

The experimental MUSA receiver is equipped with both linear and square-law detectors, and some signal-to-noise ratio comparisons were made using locally generated signals. Figure 43 shows schematically the essential parts of the test circuit. The noise generators represent the thermal noise originating in the receiver input circuits. The input signal e_s was modulated with a tone. The calculated curves shown in the figure are obtained from (8) and (9) which reduce to

$$\frac{E_s}{E_n} \propto \sqrt{1 + \left(\frac{e_{sB}}{e_{sA}}\right)^2} \qquad (10)$$
$$u = 2$$

and to

$$\frac{E_s}{E_n} \propto \left(1 + \frac{e_{sB}}{e_{sA}}\right). \qquad (11)$$
$$u = 1$$

The equation for the square-law detector is sound and was verified by the measurements. The equation for the linear detector should apply only over a certain range of signal and noise levels. The measurements indicate this.

Automatic gain control was not used in these tests since the two gain controls could not be relied upon to "track" sufficiently well. To make the measurements significant manual gain control was used to maintain the receiver gains equal and the output normal. It may be pointed out here that in receiving actual radio signals with linear detectors accurate equality of gains is not required. Moderate differences in gain (of a few decibels) can be depended upon to be beneficial as often as detrimental. With square-law detectors no departure from equality can be beneficial.

The curves of Fig. 43 show that 10- and 20-decibel differences in signal level give the square-law detector an advantage of one and two decibels, respectively. In receiving two bundles of waves the branch outputs commonly fade in and out with the result that their average ratio is of the order of 10 decibels. Such were the conditions, as well as could be estimated, during the two-bundle tests in Tables III and IV. The one-decibel correction applied to those results in Table V is therefore justified. In the one-wave bundle measurements, the percentage of time during which the branch outputs were substantially different was so small that no correction was applied in Table V. In the case of space diversity the correction is also small; about 0.5 decibel seems reasonable.

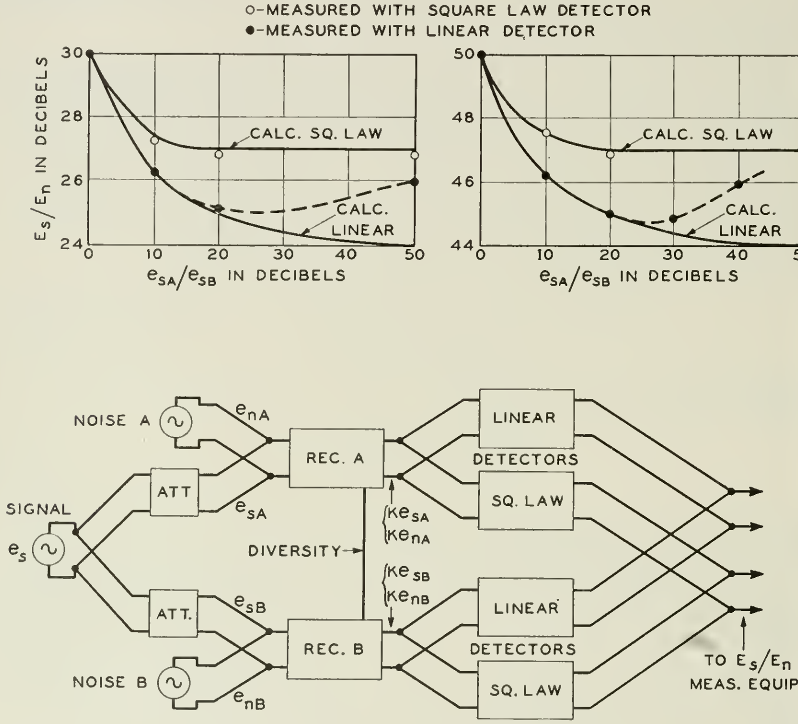


Fig. 43—Test circuit and experimental results of the study of detector characteristics.

From the point of view of distortion in receiving double side-band signals linear detectors are superior to square-law detectors and this compensates their inferiority in signal-to-noise ratio. The principal reason for using linear detection in the signal-to-noise ratio tests of the MUSA was, however, their experimental advantage in simplicity and accuracy (Fig. 35).

Random Addition of Static

In analyzing the spaced antenna systems at the beginning of this section it was assumed that the static outputs of the antennas add on a power basis. An experimental study of this was made by measuring the static output of one unit antenna and comparing it with the static output of the six antennas combined as one MUSA branch. The circuit shown in Fig. 35 was used for these experiments. The results are tabulated in Table VIII.

TABLE VIII

Date	GMT	f_{mc}	Type of Static	Addition Max. db	Addition Min. db	Thermal Noise db	QRM	Method
1935								
9–19	1530	18.6	star	8.5		−12	light	Rdg. DB Meter
10–15	1500	9.51	star	8.0				Rdg. DB Meter
10–16		9.51	distant	7.5		− 6		Rdg. DB Meter
10–22	1500	9.51	crash	8		−20	none	Vary I-F gain
10–23	1500	9.51	distant	8		− 9	light	Vary I-F gain
	1820	11.86	distant	8.5			light	Vary I-F gain
10–24	1500	9.51	star	11.4	5.4	−12	light	Vary I-F gain
	1510	9.51	star	11.0	6.0	−12	light	Vary I-F gain
	2045	9.51	crash	7.5		−30	light	Vary I-F gain
11–1	1450	9.51	distant	9.0		− 8	none	Vary I-F gain
	1830	9.51	distant	8.0		− 7	—	Vary I-F gain
1936								
1–7	1500	9.51	distant	7.5		− 8	light	Vary I-F gain
	1505	9.51	distant	8.0		− 8	light	Vary I-F gain
1–14	0300	4.82	crash	8.2	7.3	−20	none	Fluxmeter
1–15	0300	4.82	crash	6.8	3.0	−30	none	Fluxmeter
		Average		8.0				

The column headed f_{mc} indicates the frequency to which the receiver was tuned. The MUSA was tuned to a desired station which possessed a comparatively clear channel, and, following the sign-off of the station, static was measured without a demodulating carrier. The receiver was, of course, operated with manual instead of automatic gain control. Care was taken to insure that overloading did not occur. The column headed "addition" gives the ratio in decibels of the static output of the six-antenna branch to that of one antenna. Where figures are entered in the middle of the column no effect of steering was noticed. (Effects of the order of one decibel could have been overlooked, however.) The column headed thermal noise gives the ratio in decibels of receiver noise originating in one of the first circuits (measured with a resistance replacing the antenna) to the total noise measured with the unit antenna connected. It shows that thermal noise was negligible.

The star static [32] was steady and therefore accurately measurable. The crash static on 4.82 megacycles was so intermittent that it required the use of the fluxmeter to obtain a satisfactory measurement.

[32] K. G. Jansky, "Electrical Disturbances Apparently of Extraterrestrial Origin," *Proc. I. R. E.*, vol. 21, pp. 1387–1398, October, 1935.

The average of all those measurements showing no effect of steering is 8.0 decibels compared with the theoretical figure of 10 log 6 = 7.8 decibels, for random addition. The random assumption employed in the analysis is thus justified on the average.

Summarizing Discussion

The aim of the signal-to-noise study described in this section has been not so much to evaluate the intrinsic merit of the experimental MUSA system as to compare its behavior with a simple theory. The element of research has been to find out how well the transatlantic waves fit into the background of the simple theory. To this end somewhat artificial devices, thermal noise and rectified carrier, were substituted for static and speech signals.

The study included an analysis of diversity action in which the effect upon the signal-to-noise ratio of (1) delay equalization, (2) detector characteristics, and (3) automatic gain control action was displayed prominently. Although those effects, taken together, are important, they are individually small and could be separately evaluated only by locally controlled test methods.

We propose now to review the results of the tests and studies. The slight decrease in improvement which would have occurred had the speed of the automatic gain control been reduced, and the increase in improvement which would have resulted had noise *power* been averaged do not affect the fundamental considerations and will be neglected here.

One wave bundle received with one branch (Table V) yielded an improvement of 7 decibels which is about one decibel less than 10 log N = 7.8 decibels calculated by simple theory (Fig. 34, Case I). Comments relating to this have been given.

One wave bundle received with two branches steered on each side of the center of the bundle yielded eight decibels improvement (Table V). Of this, one decibel is due to primary diversity action and three decibels are due to the secondary diversity gain. This three-decibel gain was assumed in the simple theory although it was not then designated as secondary diversity gain. It accrues by virtue of voltage addition of the signals and power addition of the noise. Both of these conditions are satisfied in Table V. There remain four decibels which represent the signal-to-noise improvement in each branch, referred to one antenna. This indicates a loss of three decibels as compared with a single branch steered at the center of the bundle, which gives seven decibels; this is reasonable when it is remembered that the branches were steered apart by a phase difference of about 60 degrees ($\phi_A - \phi_B$

$= 60°$). A loss of three decibels is about what one would estimate upon inspecting the directional pattern for $\phi = 60$ and 120, say, on Fig. 20. The procedure in which two branches are steered at one bundle as in the above is frequently employed and is an important factor in the operation of a MUSA.

The case of two wave bundles tabulated in Table V also yields an improvement of 7.5 to 8 decibels. Of this, one decibel and three decibels are due to primary and secondary diversity action, respectively, as in the case of one bundle. This leaves 3.5 to 4 decibels for the signal-to-noise improvement in each branch referred to a unit antenna. But the unit antenna has the advantage of two bundles, whereas the MUSA branch excludes one of them, a three-decibel difference. In comparison with a unit antenna receiving only one bundle, the improvement to be ascribed to one branch thus is increased to 6.5 or 7 decibels. This result compares favorably with the seven decibels yielded by one branch steered at one bundle. It is in this case of two bundles that square-law detectors are most important. Their advantage, amounting to an estimated one decibel, has already been included in Table V, it will be remembered.

The measurements which have permitted the above analysis of the MUSA signal-to-noise improvement were of course supplemented by aural observations made over the course of a year and a half. The listening tests corroborate the analytical results as well as can be expected of such observations. Not infrequently they showed somewhat less than the full eight-decibel improvement. The indications are, however, that a larger MUSA with three (or possibly four) branches would have yielded more nearly its full gain of 10 log N decibels. For, a MUSA receiving system does not perform its functions properly unless it is sharp enough to separate the waves sufficiently to permit effective delay equalization; also, to obtain the full gain, enough branches must be provided to utilize all of the important wave bundles. The Holmdel experimental MUSA is really a conservative approach to the field of steerable directivity. There is, of course, an upper limit to the size of a MUSA, beyond which (1) technical difficulties in phasing, etc., will occur, (2) the cost of the improvement may be less if introduced at the transmitter, and (3) the directional sharpness becomes too great to permit practical operation with waves of the stability encountered in transatlantic transmission. At present, a system about three times the length of the experimental MUSA comprising eighteen antennas and equipped with three branches seems practical. It should yield an improvement of $10 \log 18 = 12.5$

decibels more consistently than the present MUSA yields eight decibels.

It may be worth while here to point out that as the number of antennas in a MUSA system is increased there is no tendency for static to become subordinate to thermal noise (set noise) or vice versa when static, like thermal noise, adds on a power basis. Only to the extent that transmission-line loss increases with the number of antennas will the ratio of thermal noise to static increase.

A type of transmission sometimes occurs for which the experimental MUSA gives only small signal-to-noise improvement. We refer to the highly scattered propagation associated with flutter fading, discussed at the close of Section IV. In such cases signal-to-noise improvement is not highly significant, however, since at least in the worst cases, the distortion renders the circuit worthless. Thus, increasing the transmitting power is likewise ineffective. On the other hand the experimental MUSA can accomplish something by rejecting some of the scattered waves which appear to be responsible for the flutter fading. This is accomplished without a corresponding *loss* of signal-to-noise ratio since, of course, noise is rejected, too. Fortunately, flutter fading does not seem to be associated prominently with greatly depressed field intensity so the failure to secure signal-to-noise improvement with flutter fading does not appreciably penalize the MUSA as a means of extending operation through periods of depressed field conditions.

VI. Recapitulation

The MUSA receiving system described in this paper is the culmination of some four years effort to determine the extent to which receiving antenna directivity may be carried to increase the reliability of short-wave transatlantic telephone circuits.[33] Fundamental experimental studies of wave propagation were made with particular emphasis upon how the waves arrive. Based upon the results of these studies a system was evolved in which a new technique of phasings was required. The result is a steerable antenna whose signal-to-noise advantage is seven to eight decibels compared with the largest fixed antenna that can be employed effectively. By analyzing this improvement and comparing the various contributing factors with theory, it is possible to estimate that a system three times larger than the experimental one will yield an additional four to five decibels, and will perform better consistently. In addition to the signal-to-noise improvement a

[33] Potter and Peterson, "The Reliability of Short-Wave Radio Telephone Circuits," *Bell Sys. Tech. Jour.*, vol. 15, pp. 181–196, April, 1936.

substantial improvement in quality is obtained by reducing the distortion associated with selective fading. It is both interesting and important to note that whereas so often one advantage is gained only at the expense of another, in the MUSA system the best quality improvement and the greatest signal-to-noise advantage are obtained together, without compromising.

The system developed is expensive and might be thought to illustrate the law of diminishing returns. As a part of a point-to-point radiotelephone system, however, it has certain compensating features not mentioned thus far. One of these is the broad frequency band feature.

With essentially aperiodic unit antennas the MUSA possesses a broad frequency range; i.e., the directional pattern, despite its sharpness, is substantially the same over a band of a hundred or more kilocycles provided the terminal equipment is made sufficiently broad. (See Appendix I.) The broad-band feature is important for its possibilities in multiplexed operation of telephone circuits; i.e., it makes possible, insofar as the antenna system is concerned, the adaptation of some of the carrier telephone methods to radio circuits. It is to be expected that, excepting certain critical cases, fairly large percentage frequency bands will follow virtually the same paths. This assumption was verified by a few experiments in which pulses were received simultaneously from GBS (Rugby, 12,150 kilocycles) and GBU (Rugby, 12,290 kilocycles) 140 kilocycles apart. These tests showed that, although the pulse fading was, of course, not synchronous, the angles involved were alike.

Another compensating feature of the MUSA receiving system is that, with suitable terminal equipment, reception may be carried on from several points at once provided they lie within the horizontal angular range of the unit antenna. Some sacrifice in vertical angular selectivity occurs but this is confined to low angles where it is least important.

Certain features of the system make for economies in plant cost. The fact that a great many components are identical permits manufacturing economies. Also, spare units need be provided only for a few vital functions, since the failure of one of the many similar parts does not disrupt service.

The development of steerable directivity has thus far been concerned with receiving antennas. In receiving, one has the obvious advantage of having, in the monitoring branch, a criterion to dictate the steering adjustments. The lack of such a direct criterion for adjusting transmitting directivity does not, however, rule out the possibility, at some future time, of a MUSA transmitting system. That horizontal steering of transmitting directivity may be decidedly important is strongly suggested by observations made on transmissions from Daventry in which significant effects upon flutter fading have been found to be associated with the orientation of the directional transmitting antennas.

Acknowledgment

The experiments described in this paper necessarily involved the coordinated effort of many individuals, in both the British Post Office and the Bell System, and their help has been appreciated. Mr. E. Bruce had charge of the design of the rhombic antennas and transmission lines, and Messrs. L. R. Lowry and W. M. Sharpless had important parts in the various phases of the work.

The authors are particularly indebted to Mr. R. K. Potter who contributed much through his keen interest throughout the entire work.

Appendix I

Broad-Band Characteristic of the MUSA

The frequency characteristic of the MUSA may be calculated from (3). Frequency and angle appear only in the form $2\pi a(\upsilon - \cos\delta)$ where a is inversely proportional to frequency. By writing the equation

$$\frac{2\pi Df}{c}[\upsilon - \cos\delta] = \frac{2\pi D(f + \Delta f)}{c}[\upsilon - \cos(\delta + \Delta\delta)]$$

we express the angular shift, from δ to $(\delta + \Delta\delta)$, of a given point on the directional pattern as the frequency is varied from f to $(f + \Delta f)$. This equation may be rewritten as

$$1 + \frac{\Delta f}{f} = \frac{\upsilon - \cos\delta}{\upsilon - \cos(\delta + \Delta\delta)}.$$

As an example consider $\Delta f = 200$ kilocycles, $f = 10$ megacycles, $\delta = 30$ degrees, and $\upsilon = 1.05$. Then $\Delta\delta = -0.4$ degree. For lower values of δ, $\Delta\delta$ becomes still smaller.

The frequency characteristic expressed in terms of percentage band and angular shift given by the above equation is independent of the size of the MUSA. It relates to the over-all length of the system, however, by the fact that for greater lengths a given angular shift has more effect.

The broad band of the MUSA reflects the fact that, with the terminal at the "leeward" end as assumed heretofore, the delay of the space paths is nearly the same as that of the transmission-line paths so that if the antenna outputs are phased to add at one frequency they will nearly add at other frequencies. If the terminal is located at the center of the MUSA to economize on transmission line, the frequency range is greatly reduced. The broad band may be regained, however, by delays introduced in the receiving equipment. With a center location, the antennas in the forward and rearward sections of the MUSA must have their phases shifted oppositely, and, unless certain other compensating networks are provided, the two phase shifts must be coupled in different phase relations for different wave-lengths.

Abstracts of Technical Articles from Bell System Sources

Modern Theater Loud Speakers and their Development.[1] C. Flannagan, R. Wolf, and W. C. Jones. Although many of the basic ideas involved in the operation of present-day loud speakers were conceived during the early stages of the development of the telephone, it was not until the advent of the vacuum tube amplifier that these principles were applied to the design of structures capable of delivering sufficient acoustical power to be audible throughout a room or auditorium. Having reached this stage, however, the developments that culminated in the sound reproducing systems employed with present-day sound pictures came in rapid succession. These developments have embraced all phases of loud speaker design, with the result that systems are now available that convert from 25 to 50 per cent of the electrical input into acoustical output, and maintain conversion efficiencies of this order of magnitude over a frequency range of 50 to 10,000 cps. These systems are so designed as to be capable of reproducing the recorded sound at intensities that not only greatly enhance the dramatic effect of the presentation in the theater, but also open entirely new fields in recording. All these improvements have been attained with a reduction in distortion and improved fidelity of the reproduced sound. The directional properties of the loud speakers also have been markedly improved, with the result that the better quality of reproduction achieved is available throughout the entire seating area and the undesirable beam effects previously experienced have been eliminated.

Power System Faults to Ground—Part I: Characteristics.[2] C. L. Gilkeson, P. A. Jeanne and J. C. Davenport, Jr. The results of an extensive oscillographic study of power-system faults to ground are presented herewith. While this study was made primarily to obtain data useful in inductive coordination problems, the results are believed to be of general interest as well. They provide data on such items as frequency of occurrence of ground-current disturbances, their monthly distribution, duration, cause, method of clearance, and wave-trace characteristics. Data on fault resistance are given in part II, a companion paper.

[1] *Jour. S. M. P. E.*, March 1937.
[2] *Elec. Engg.*, April 1937.

Direct Recording and Reproducing Materials for Disk Recording.[3] A. C. KELLER. Recently materials for direct recording and reproducing work have been improved so that they are now suitable for many uses. These materials, as they are available on the market, are classified chemically into five groups and measurements are given of frequency characteristic, surface noise, life, distortion, etc. These data have been taken with both lateral and vertical recording.

[3] *Jour. Acous. Soc. Amer.*, April 1937.

Contributors to this Issue

Edwin H. Colpitts was introduced to readers of the *Journal*—if it can be said that he needed an introduction—in the April issue.

Karl K. Darrow, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

C. B. Feldman, B.Sc., University of Minnesota, 1926; Teaching Fellow, University of Minnesota, 1926–28; M.Sc., University of Minnesota, 1928. Bell Telephone Laboratories, 1928–. Mr. Feldman has been engaged in short-wave radio receiving. His work has been mainly on transmission lines, antennas, and wave propagation problems.

H. T. Friis, E.E., Royal Technical College in Copenhagen, 1916; Columbia University, 1919–20. Research Department, Western Electric Company, 1920–24; Bell Telephone Laboratories, 1925–. Mr. Friis' work has been largely in connection with radio reception methods and measurements. He has published papers on vacuum tubes as generators, radio transmission measurements and static interference. As Radio Research Engineer he now directs studies of new methods of short-wave reception.

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.

John Riordan, B.S., Sheffield Scientific School, Yale University, 1923. American Telephone and Telegraph Company, Department of Development and Research, 1926–34; Bell Telephone Laboratories, 1934–. Mr. Riordan's work has been mainly on problems associated with inductive effects of electrified railways.

R. A. Sykes, Massachusetts Institute of Technology, B.S. 1929; M.S. 1930. Columbia University, 1931–33. Bell Telephone Laboratories, Research Department, 1930–. Mr. Sykes has been engaged in the application of piezoelectric crystals to selective networks, and more recently in the use of coaxial lines as filter elements.

VOLUME XVI OCTOBER, 1937 NUMBER 4

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50c per Copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are fifty cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

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

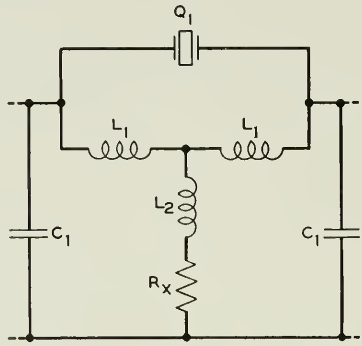


Fig. 1—A bridge T quartz crystal filter.

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

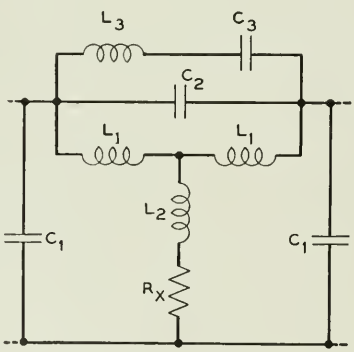


Fig. 2—Electrical network equivalent to crystal filter of Fig. 1.

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.

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 sub-

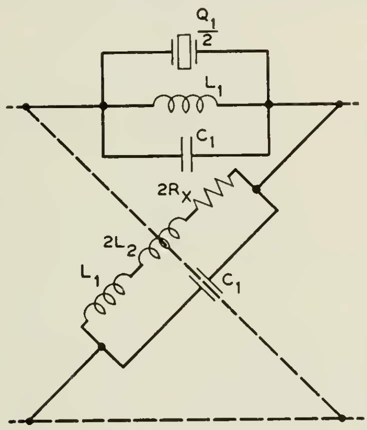


Fig. 3—Lattice equivalent of crystal filter of Fig. 1.

stantially 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

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

[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.

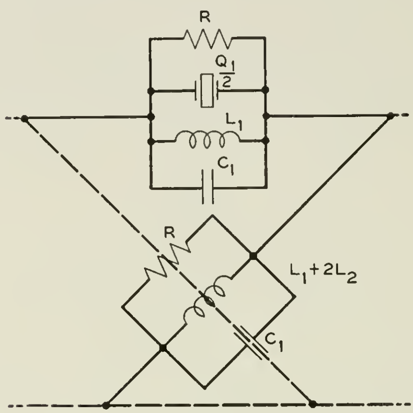


Fig. 4—Filter showing paralleling resistance.

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

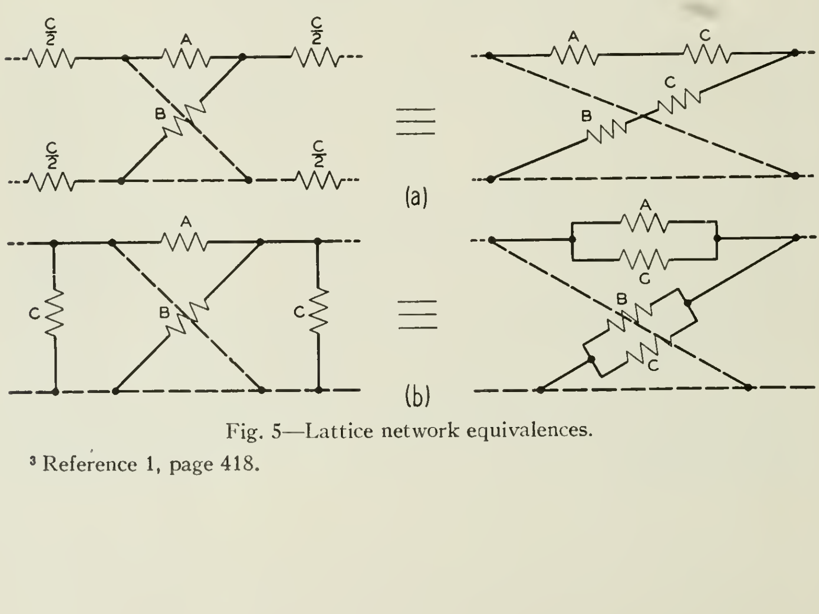


Fig. 5—Lattice network equivalences.

[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

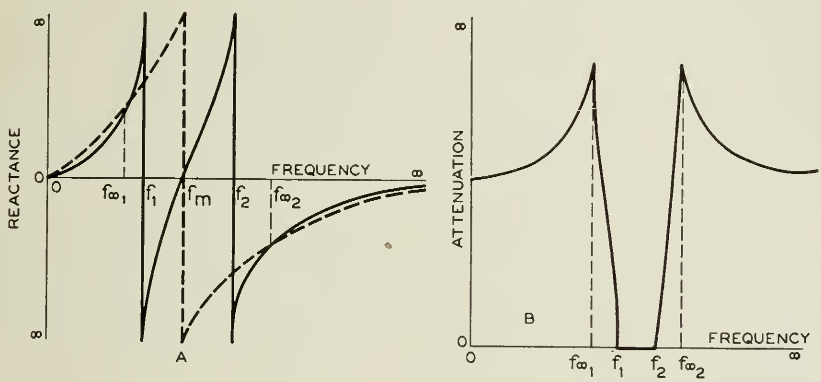


Fig. 6—Characteristics obtainable with the crystal filter of Fig. 1.

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

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

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

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.

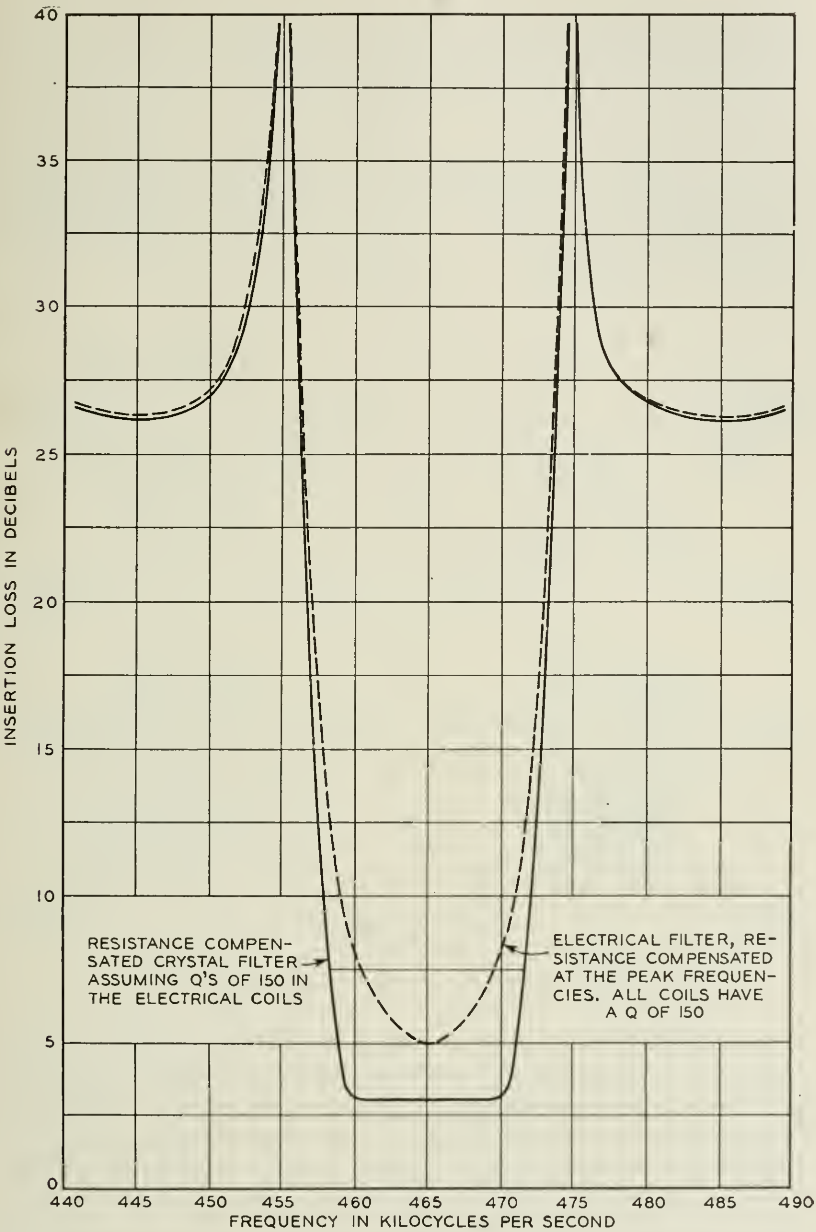


Fig. 7—Numerical comparison between the loss characteristics of a crystal filter and a coil and condenser filter.

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.

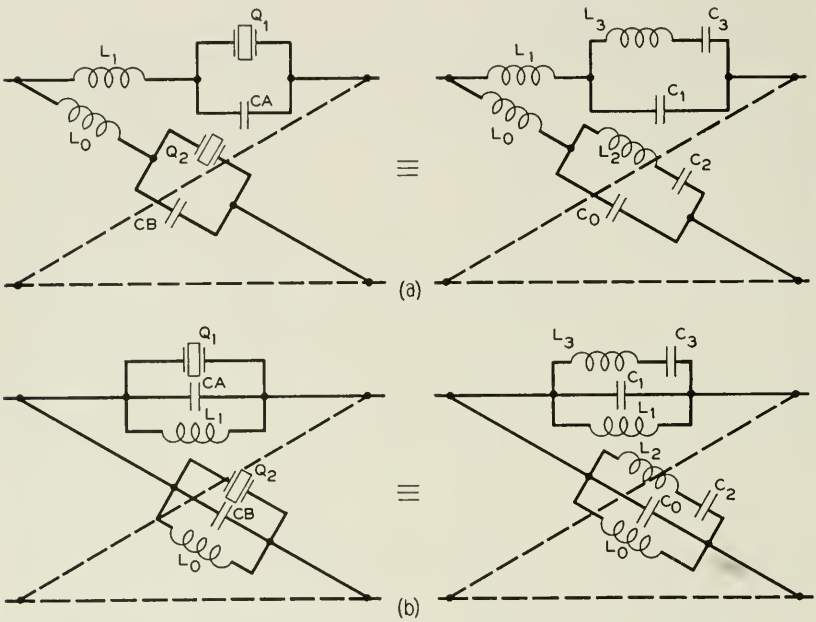


Fig. 8—Wide-band lattice crystal filters.

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-

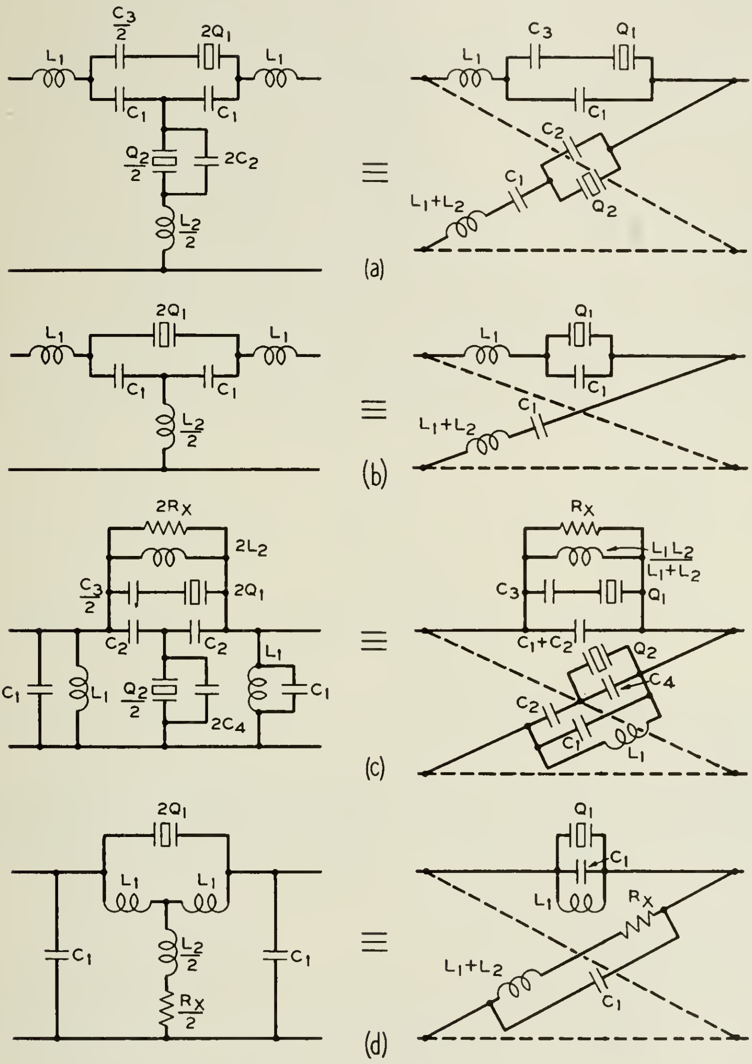


Fig. 9—Wide-band bridge T crystal filters.

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

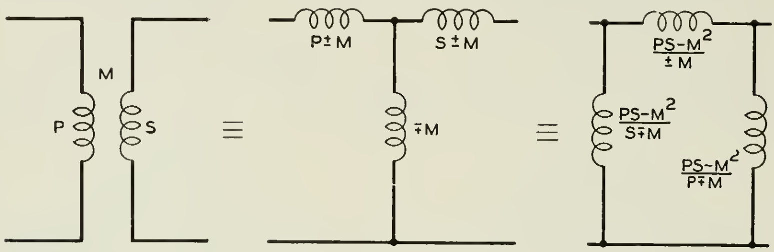


Fig. 10—T and π network equivalences of a transformer.

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.

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

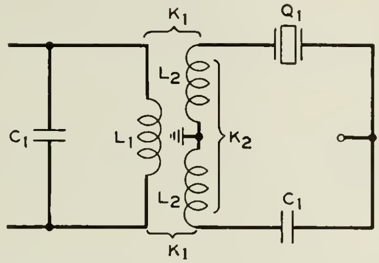


Fig. 11—A three-winding transformer crystal filter.

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

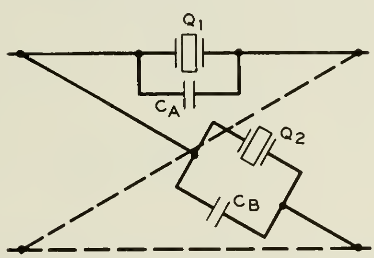


Fig. 12—A narrow-band lattice crystal filter.

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

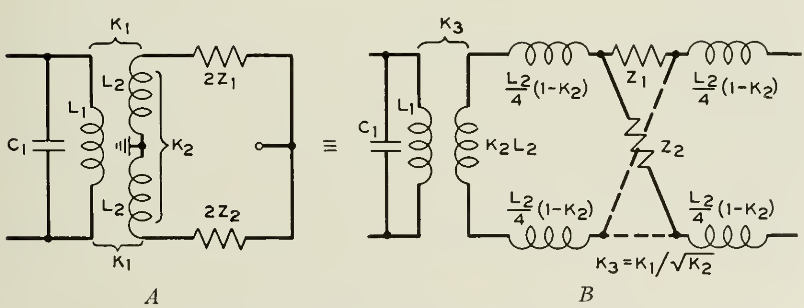


Fig. 13—An equivalent circuit for a three-winding transformer and network.

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.

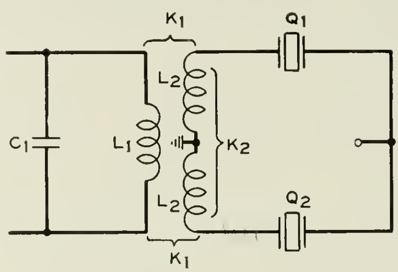


Fig. 14—A three-winding transformer crystal filter with two crystals.

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

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

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 of 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

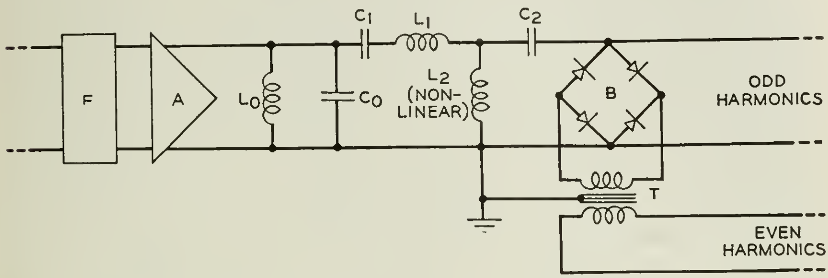


Fig. 1—Circuit diagram of channel harmonic generator.

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.

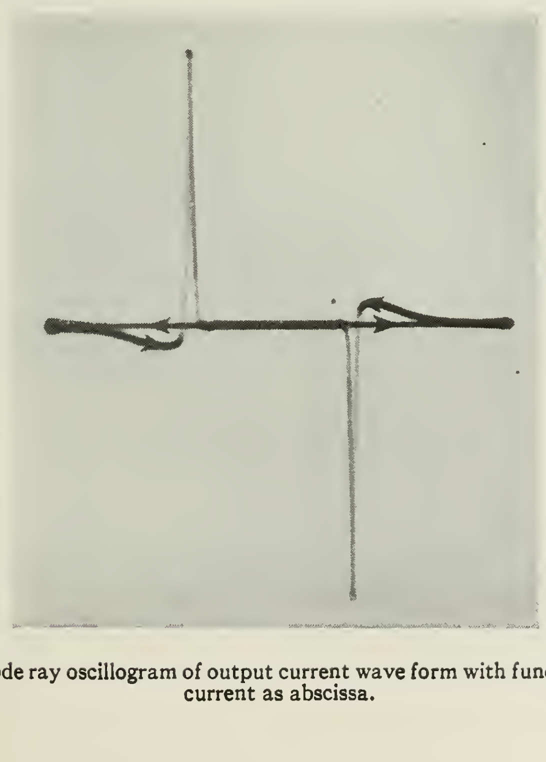


Fig. 2—Cathode ray oscillogram of output current wave form with fundamental input current as abscissa.

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. 4a 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 $\pm$ 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.

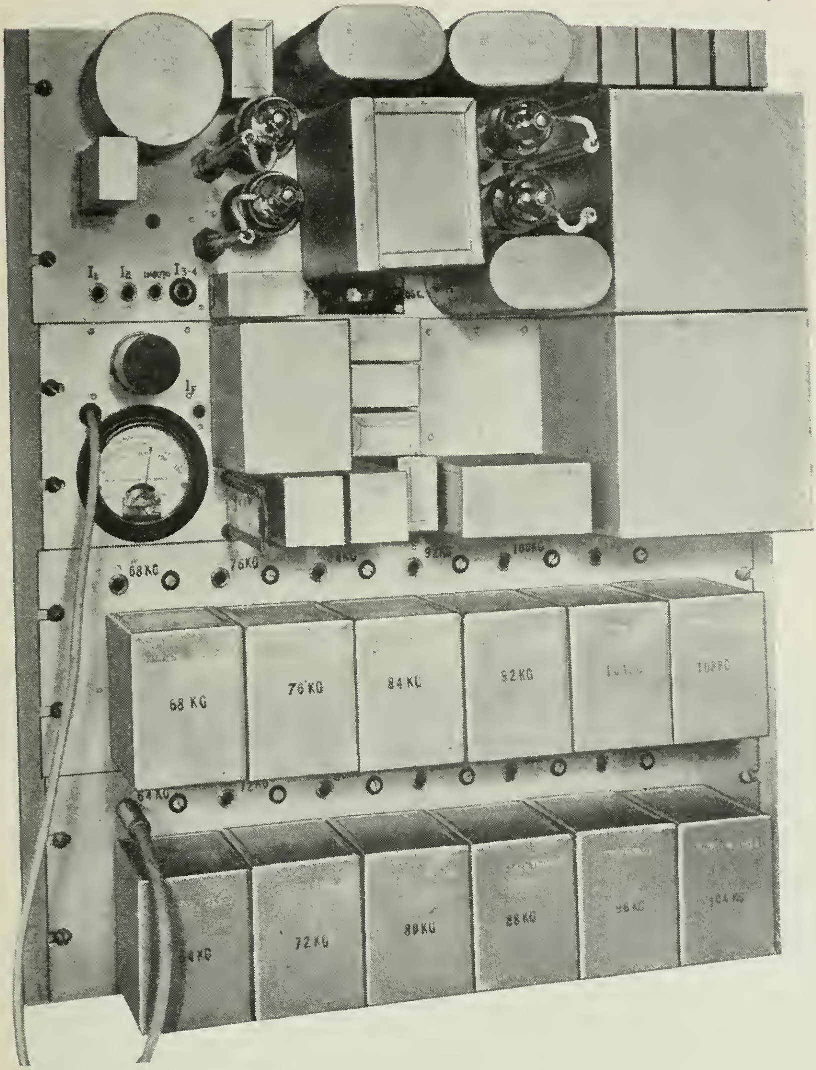


Fig. 3—Carrier supply unit, furnishing twelve harmonics of 4 kc. (experimental model).

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

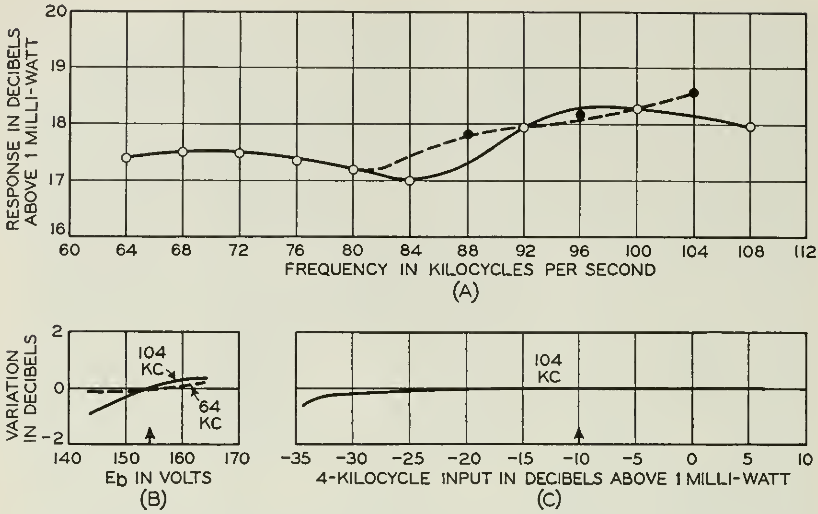


Fig. 4—Performance curves of channel harmonic generator. (*A*) Harmonic outputs; (*B*) Variation of 16th and 26th harmonics with amplifier plate potential; (*C*) Variation of 26th harmonic with fundamental input.

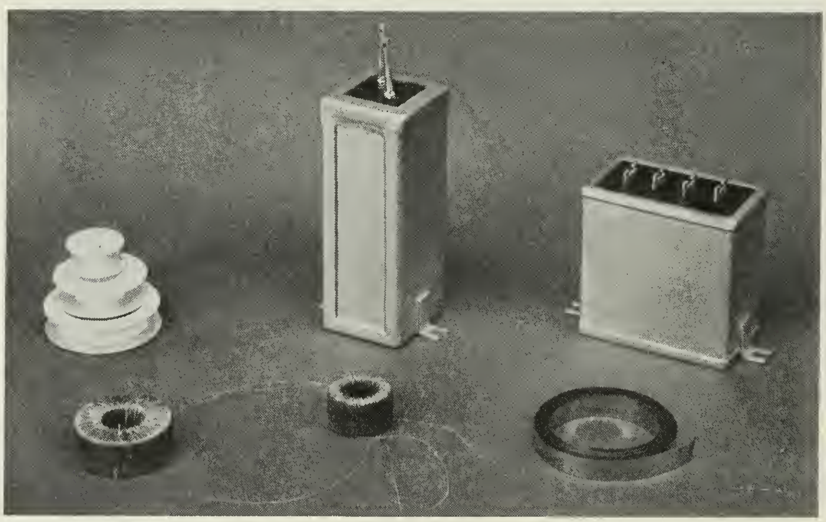


Fig. 5—Construction of experimental non-linear coils used for harmonic generation, showing core forms, magnetic tape, wound coils, and assembled units.

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 $\pm\, I_0$,

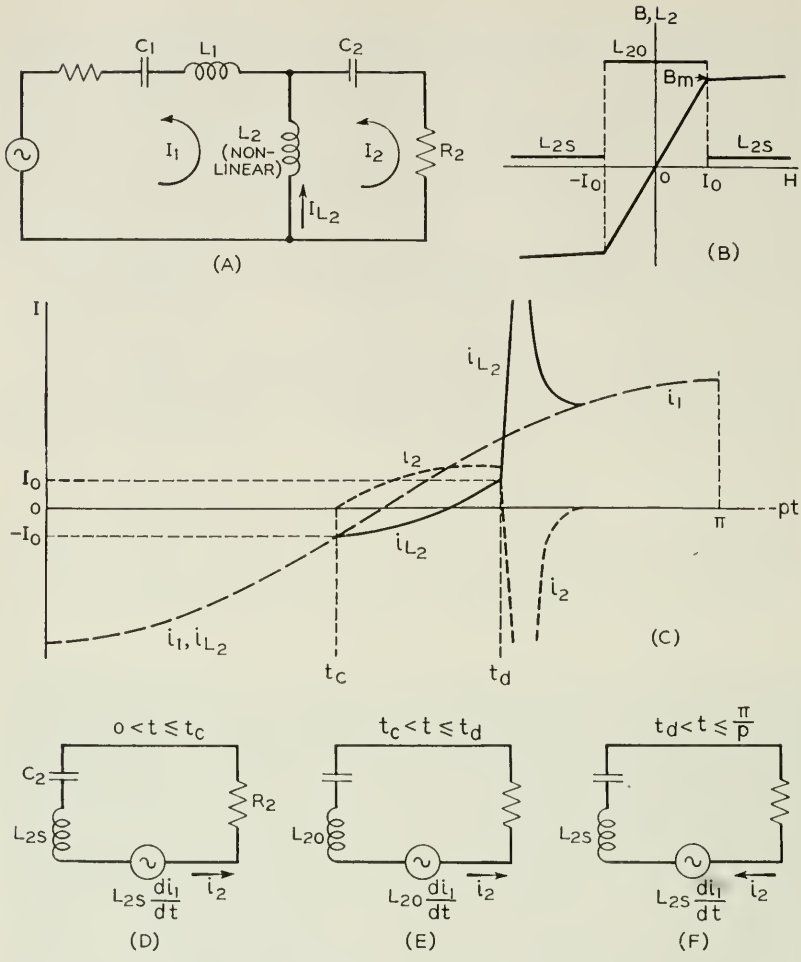


Fig. 6—Diagrams illustrating operation of the harmonic generator. (*A*) Harmonic generator circuit; (*B*) Differential inductance and flux density of assumed non-linear coil as functions of magnetizing force; (*C*) Variation with time of currents in primary and secondary meshes, and in non-linear coil;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

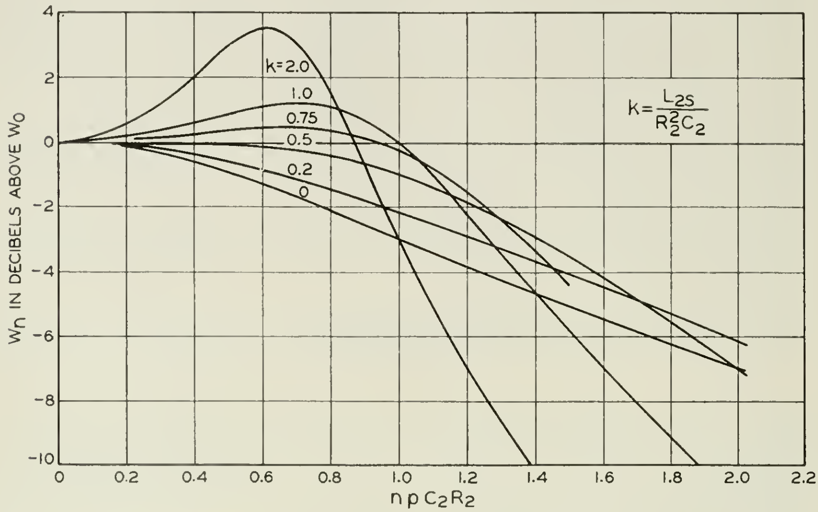


Fig. 7—Harmonic power spectrum plotted from eq. (2) as function of npC_2R_2 with k as parameter.

plitudes 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

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

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

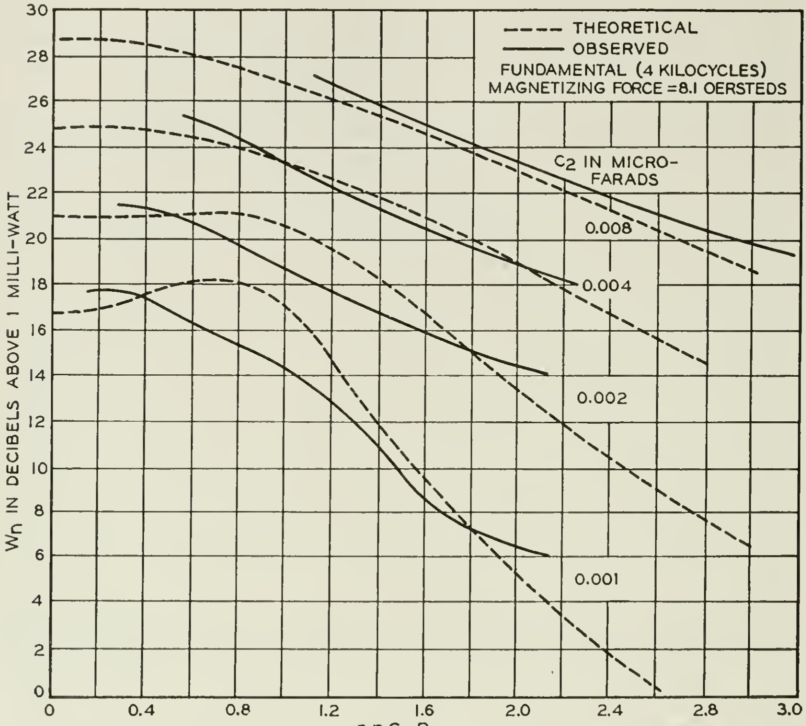


Fig. 8—Comparisons of calculated and measured harmonic distributions, plotted as functions of npC_2R_2, with C_2 as parameter.

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.

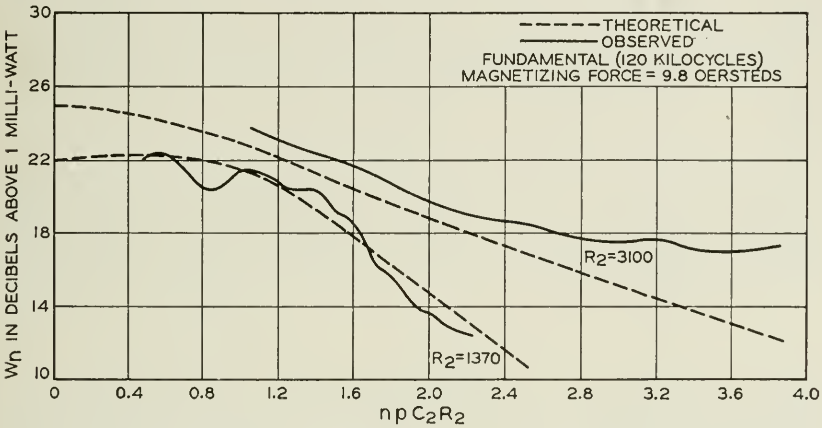


Fig. 9—Comparisons of calculated and measured harmonic distributions plotted as functions of npC_2R_2, with R_2 as parameter.

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 cross-talk, 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

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

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.

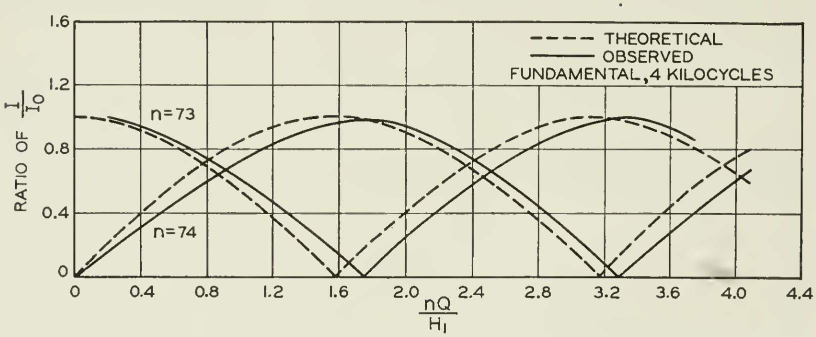


Fig. 10—73rd and 74th harmonic amplitudes as functions of direct current flowing through non-linear coil. Ordinate is ratio of harmonic amplitude with bias indicated, to that of 73rd harmonic with zero bias. Abscissa is harmonic number multiplied by the ratio of bias to fundamental. Dashed lines calculated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

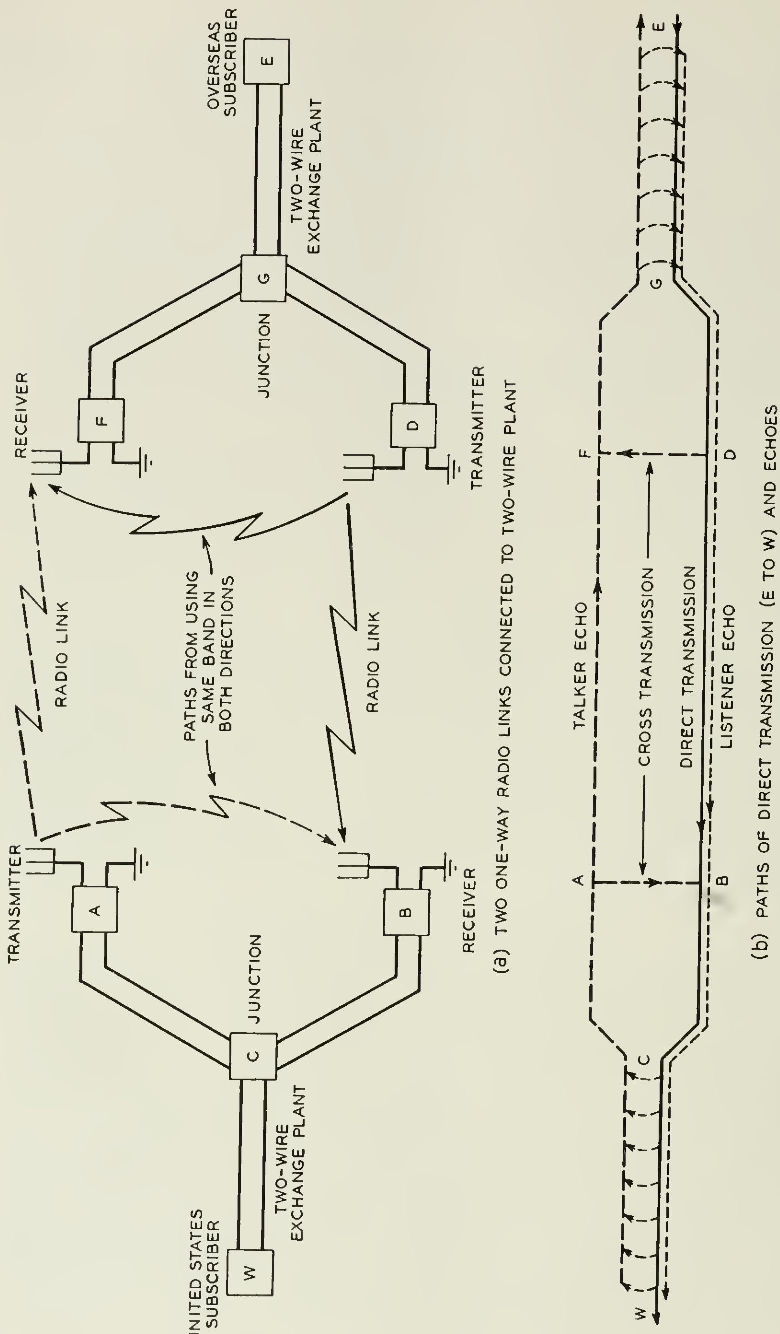


Fig. 1—Echoes in a radio telephone connection.

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.

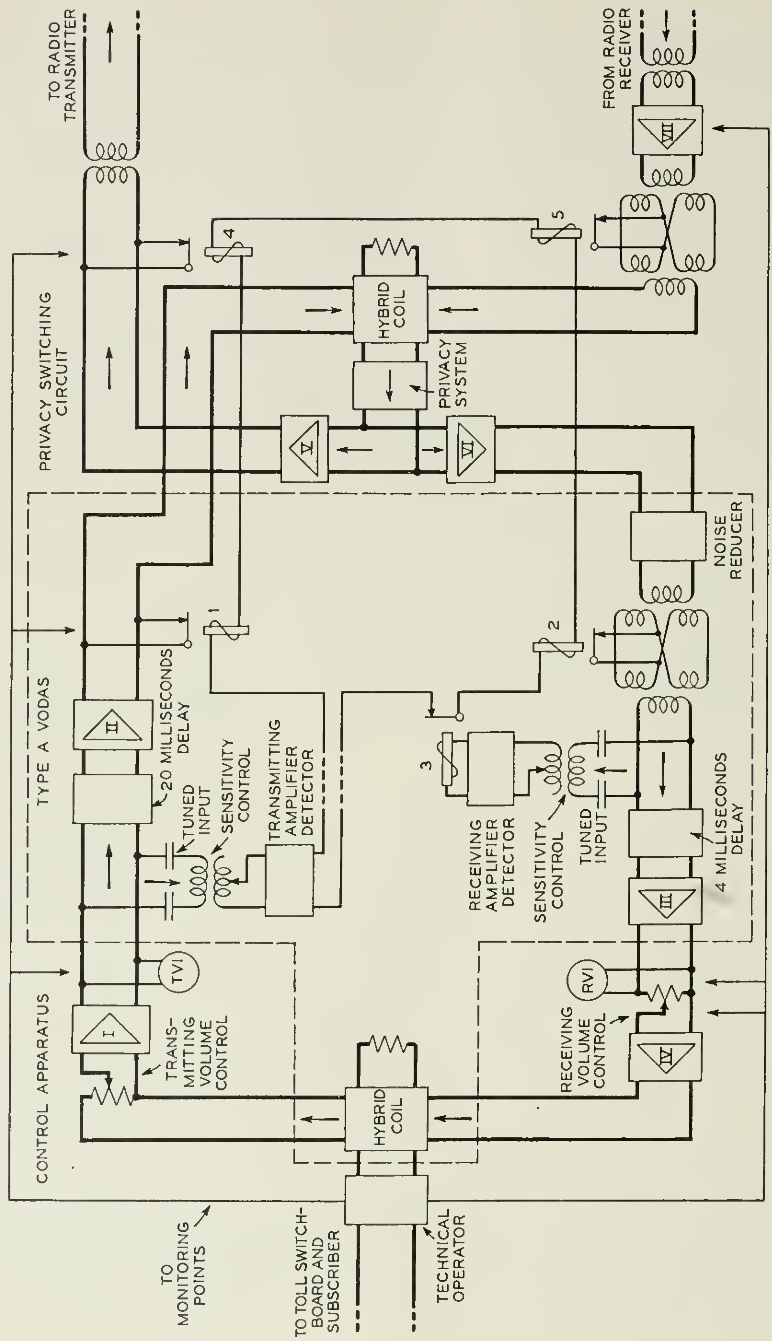


Fig. 2—Schematic of type A control terminal.

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.*

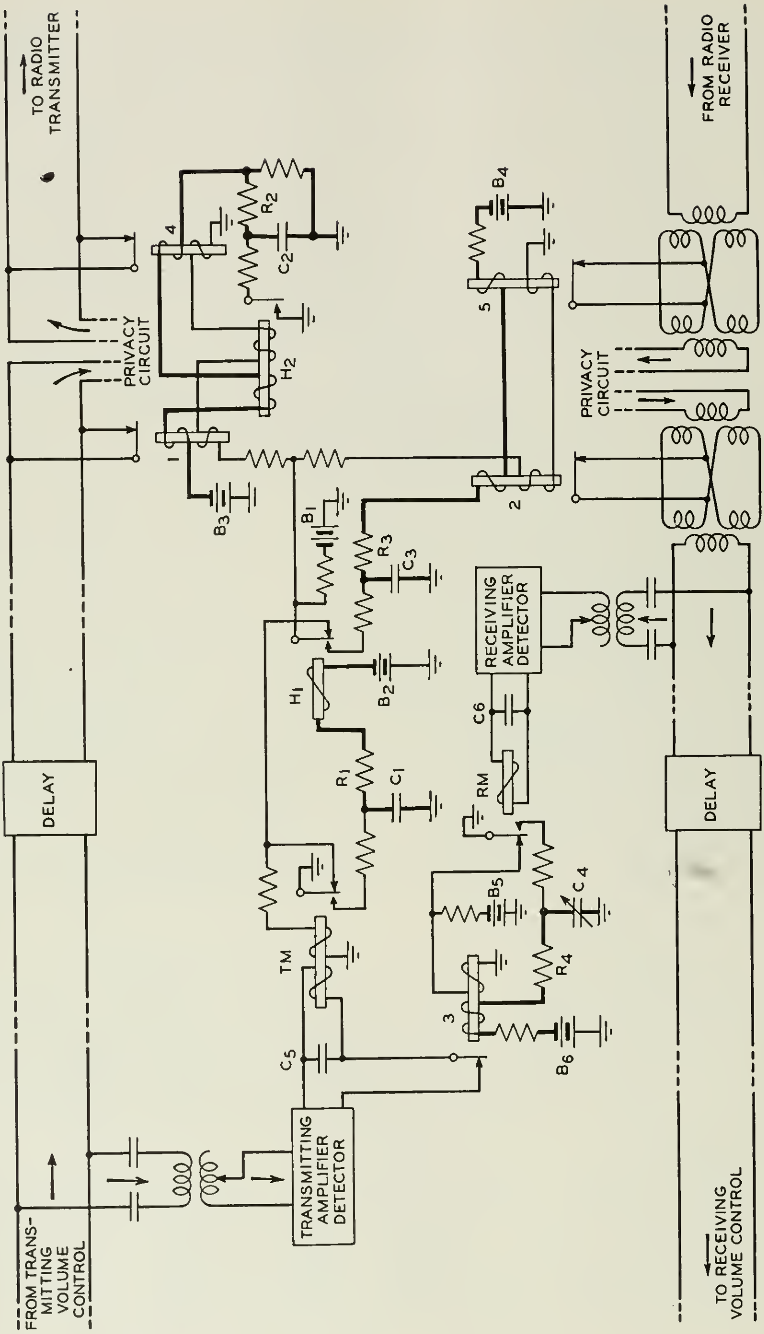


Fig. 3—Vodas relay circuits.

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.

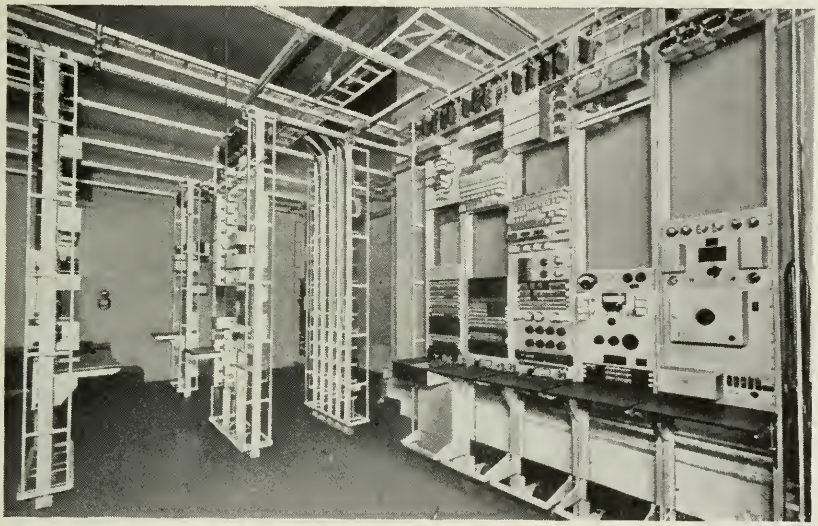


Fig. 4—Type A control terminal at San Francisco.

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 $\frac{1}{2}$ cycle per second. The resulting waves enter the low-pass filter, the output of which contains frequencies between $\frac{1}{2}$ 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

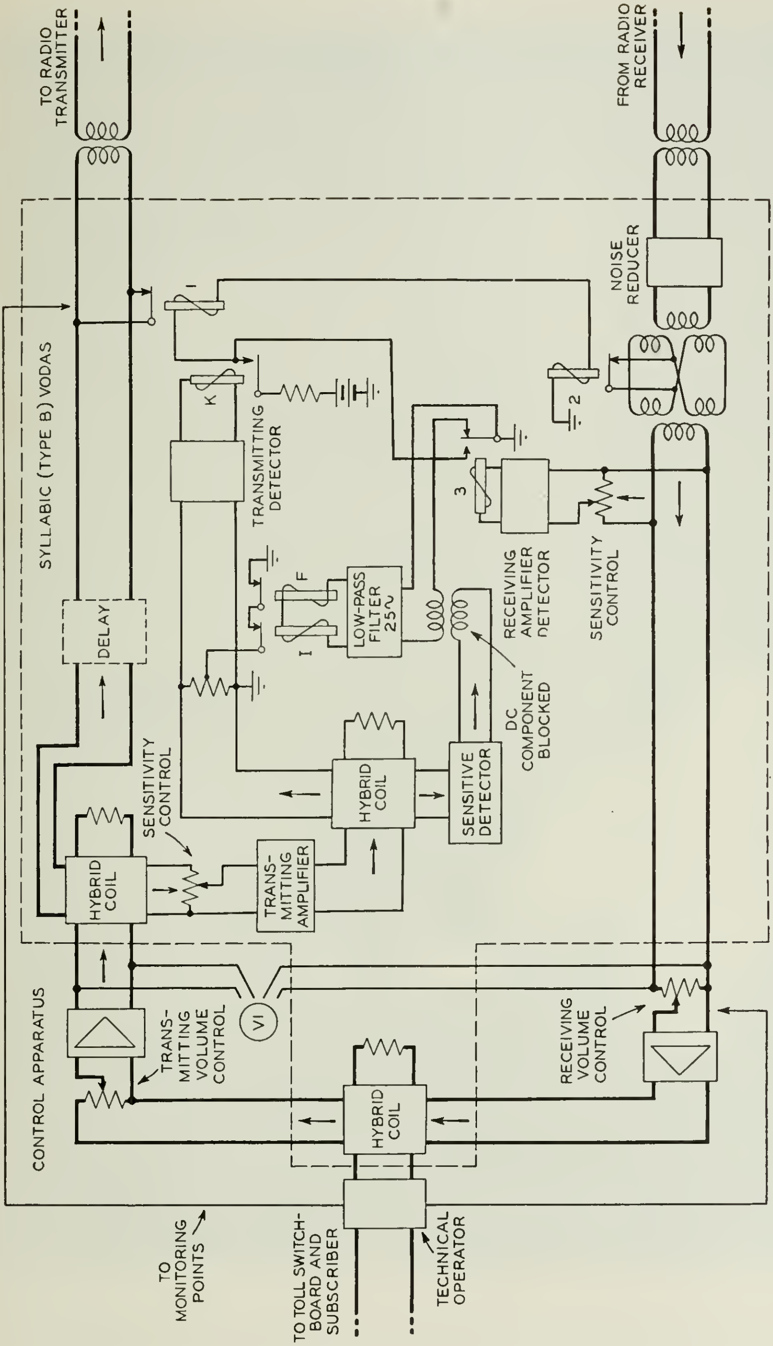


Fig. 5—Schematic of type B control terminal with syllabic vodas.

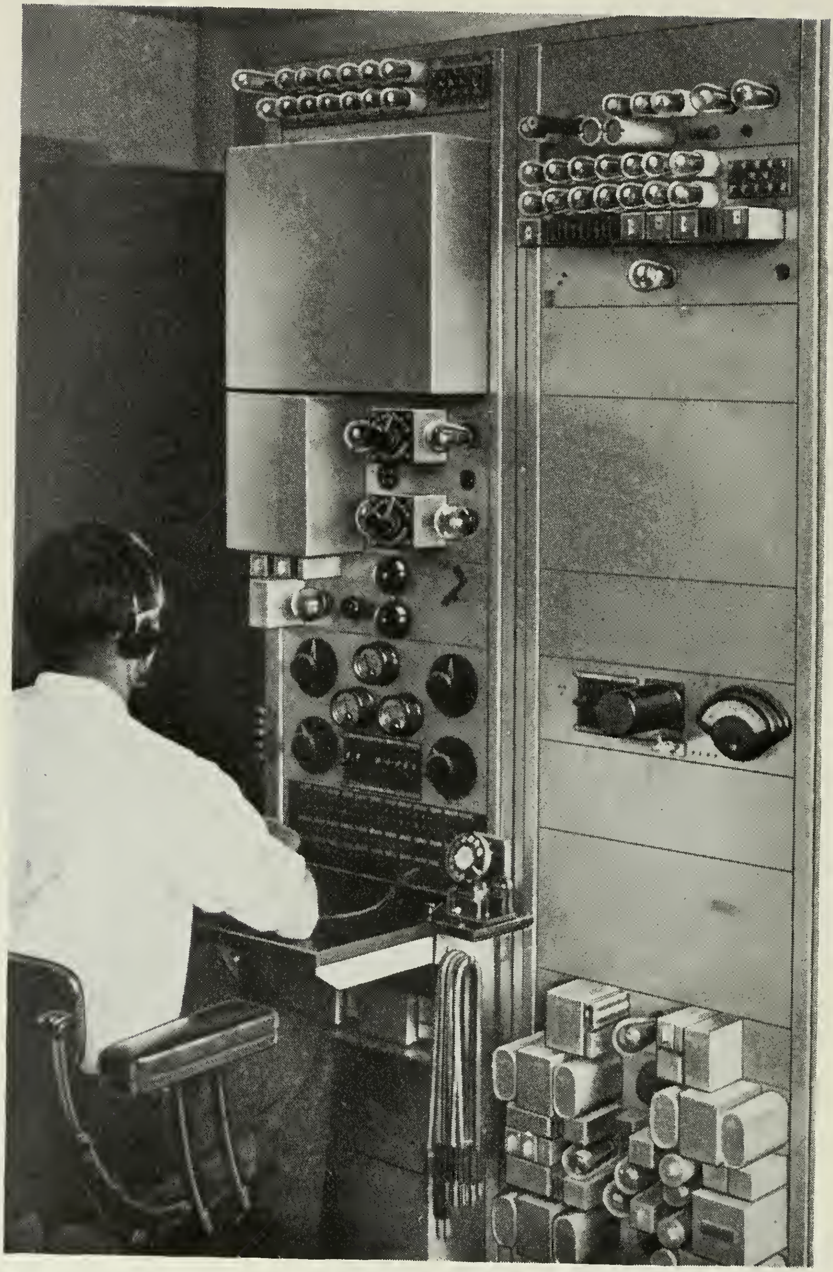


Fig. 6—Technical operator at Forked River, N. J., using a type B control terminal to establish a circuit between a steamship and a shore telephone operator.

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

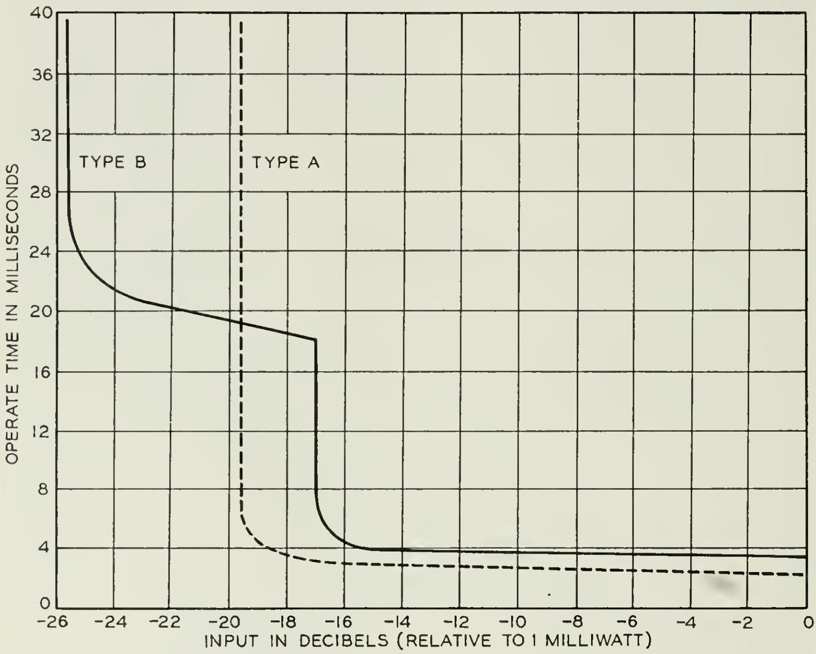


Fig. 7—Vodas operating times with sine waves suddenly applied.

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.

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

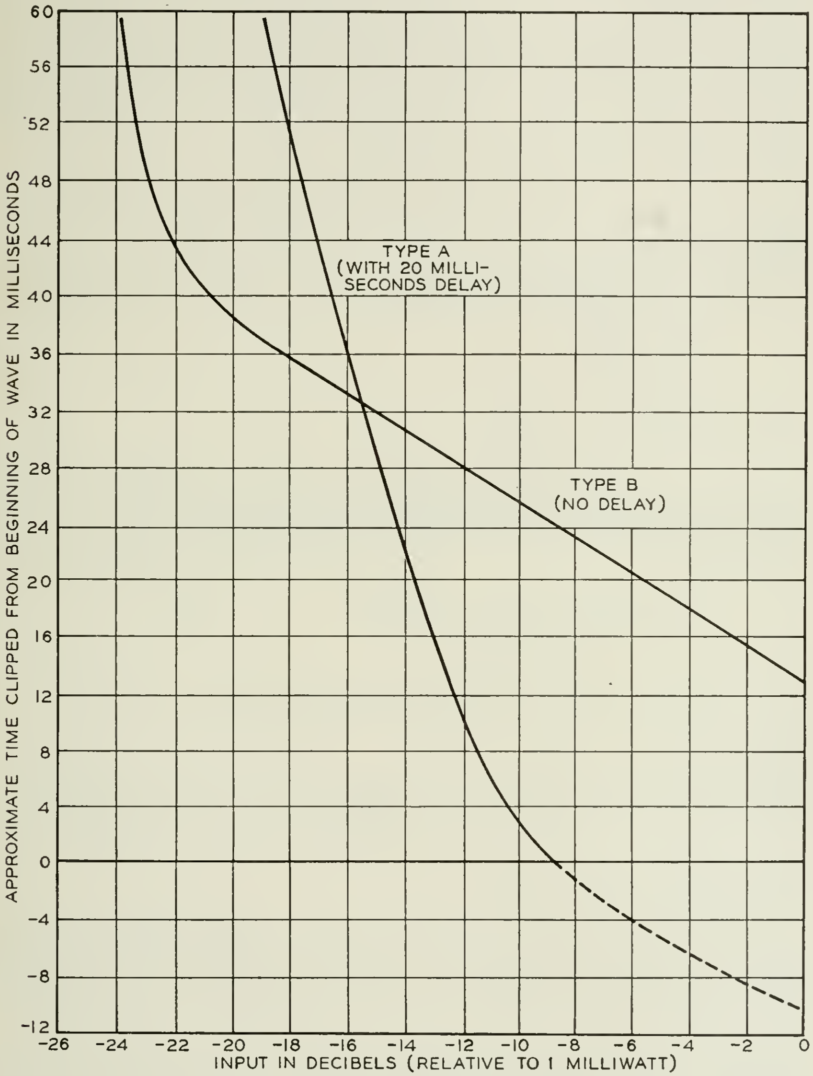


Fig. 8—Operation on a 5-C.P.S. sine wave.

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 delay 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

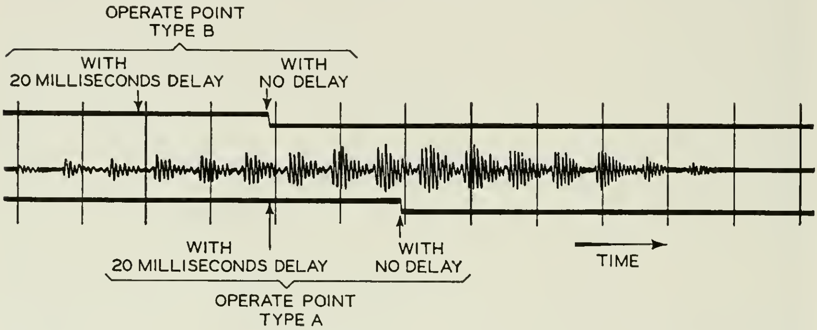


Fig. 9—Oscillogram of the word "SIX," illustrating clipping and its reduction by a delay circuit in the transmission path.

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.

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

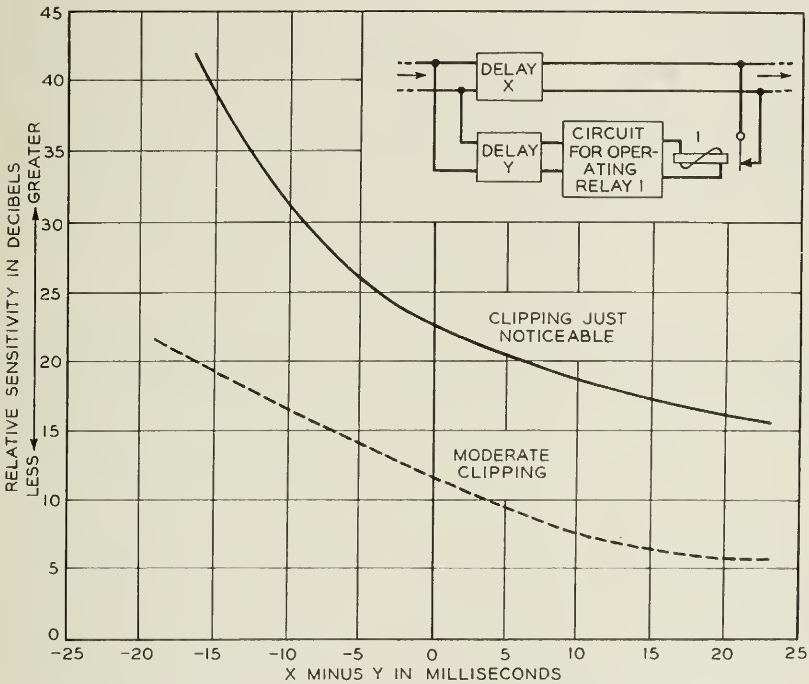


Fig. 10—Typical delay vs. sensitivity for certain clipping effects.

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.

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.

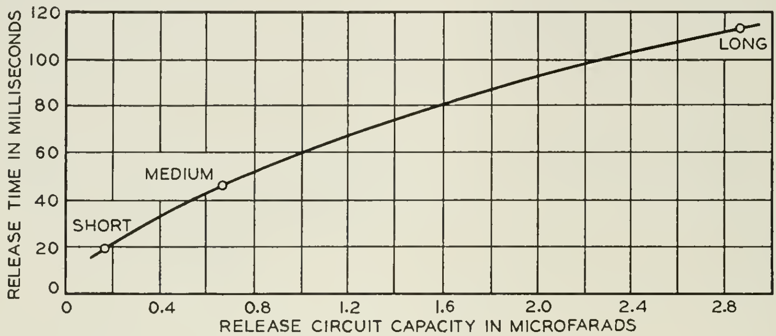


Fig. 11—Release time vs. capacitance.

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.

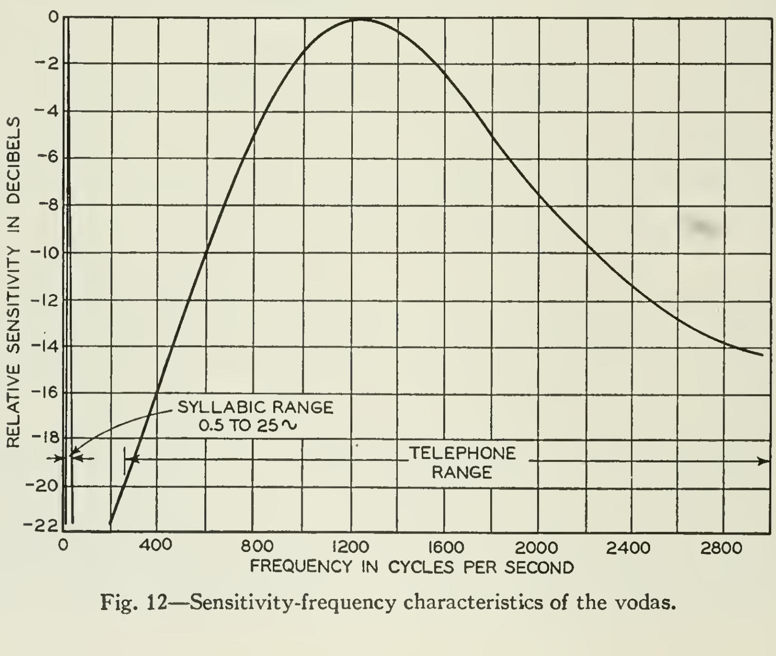


Fig. 12—Sensitivity-frequency characteristics of the vodas.

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 effec-

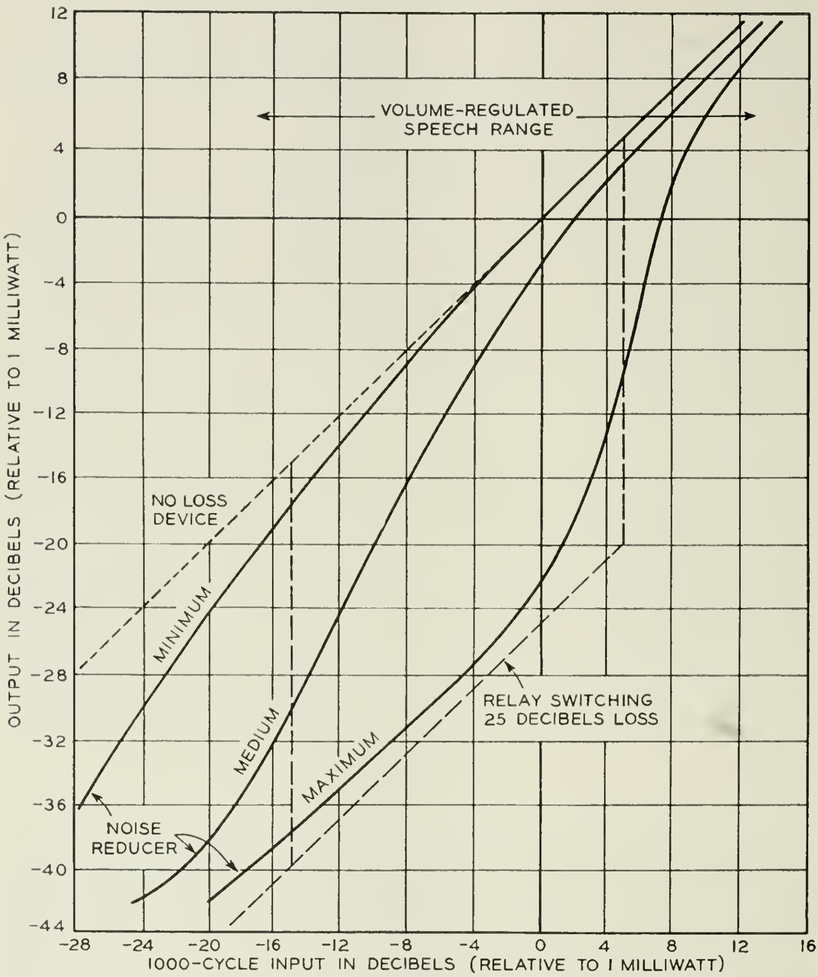


Fig. 1—Input-output comparison of noise reducer and voice-operated relay.

tively control the loss. This contrasts with a very fast limiter acting on high-peak crashes only.

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,

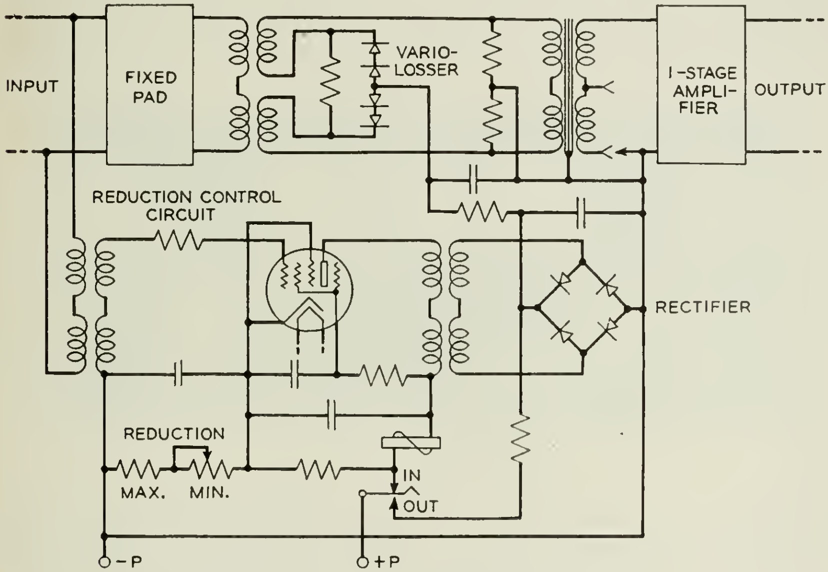


Fig. 2—Simplified schematic of noise reducer.

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.

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

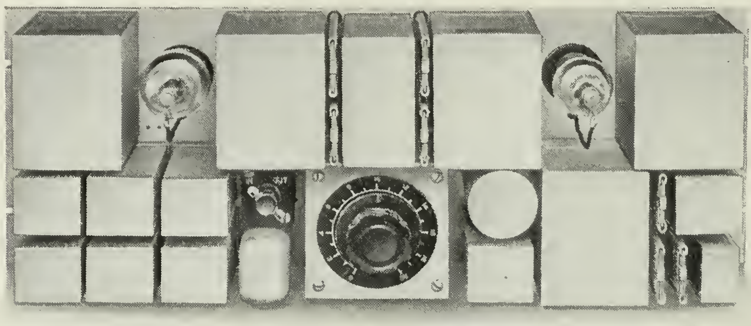


Fig. 3—noise reducer panel.

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.

The noise reducer is contained on a $7\frac{1}{4}$ 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.

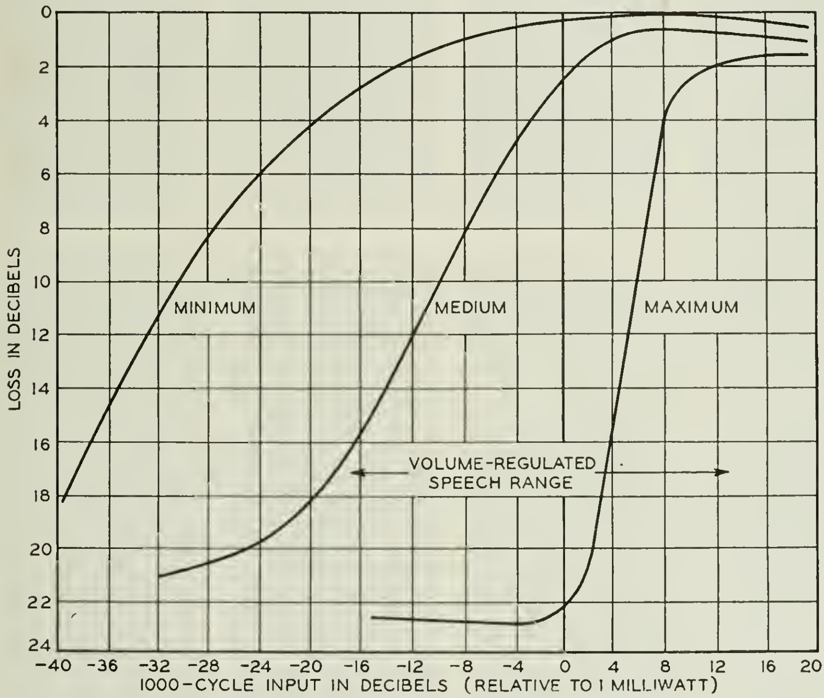


Fig. 4—Loss versus input for several settings of the reduction control.

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.

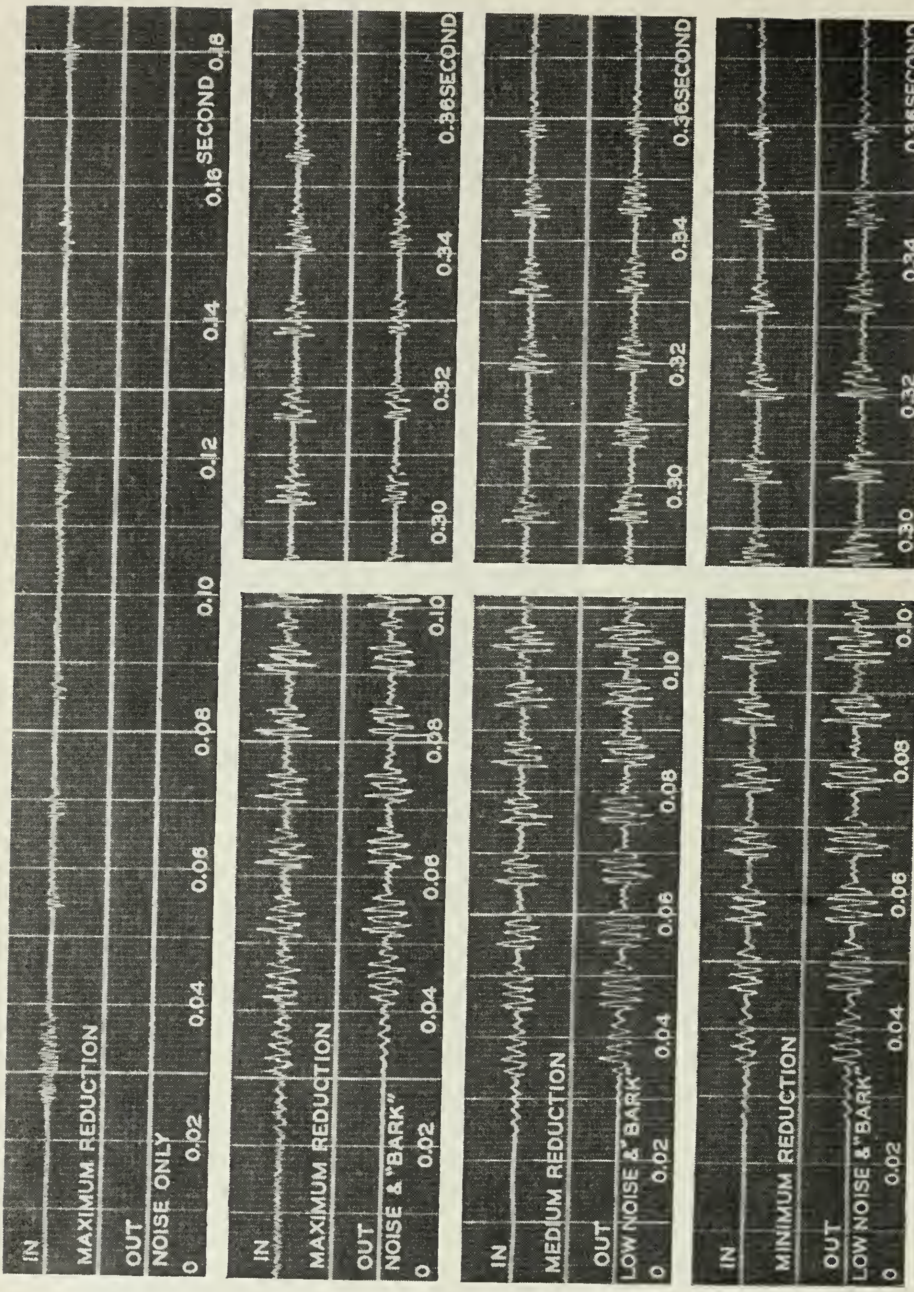


Fig. 5—Input and output for: (1) high noise with maximum reduction; (2) high noise with the beginning and ending of the word "bark"—maximum reduction; (3) low noise, the word "Bark"—medium reduction; (4) low noise, the word "bark"—minimum reduction.

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

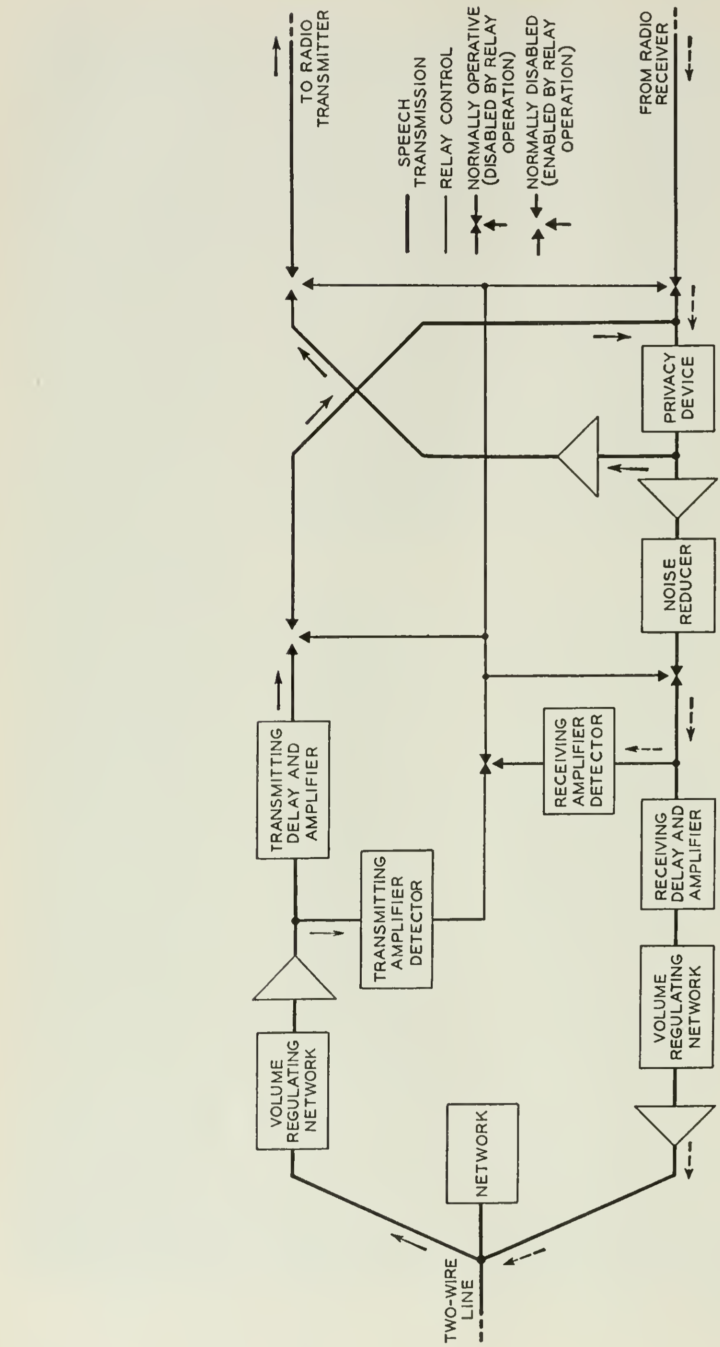


Fig. 6—Application of noise reducer to radio control terminal.

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

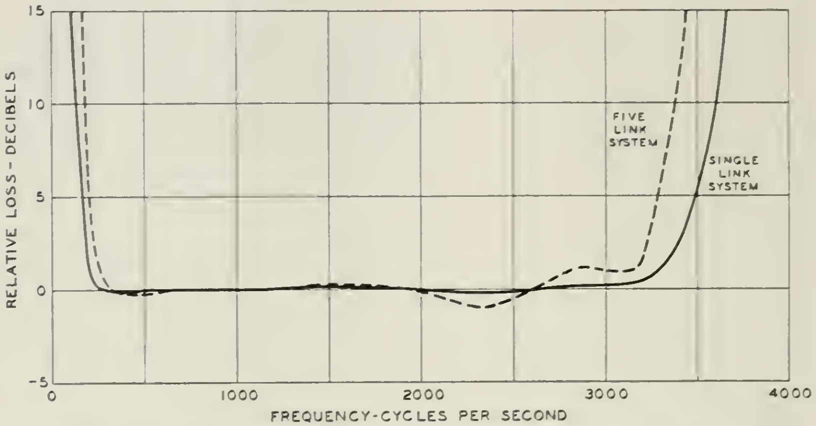


Fig. 1—Transmission frequency characteristics of broad-band systems.

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.

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

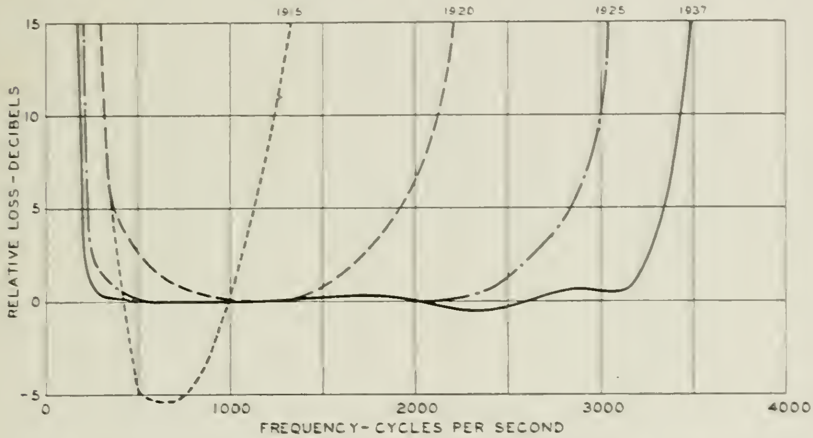


Fig. 2—Representative transmission frequency characteristics of 3000-mile toll circuits.

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.

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

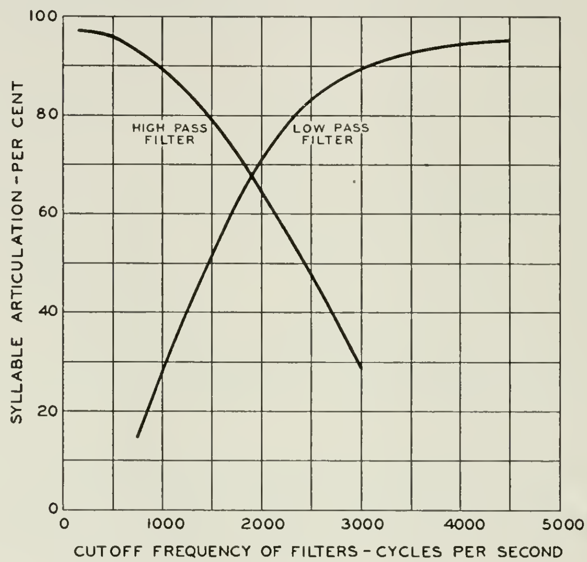


Fig. 3—Effect of cutoff frequency on syllable articulation.

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.

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

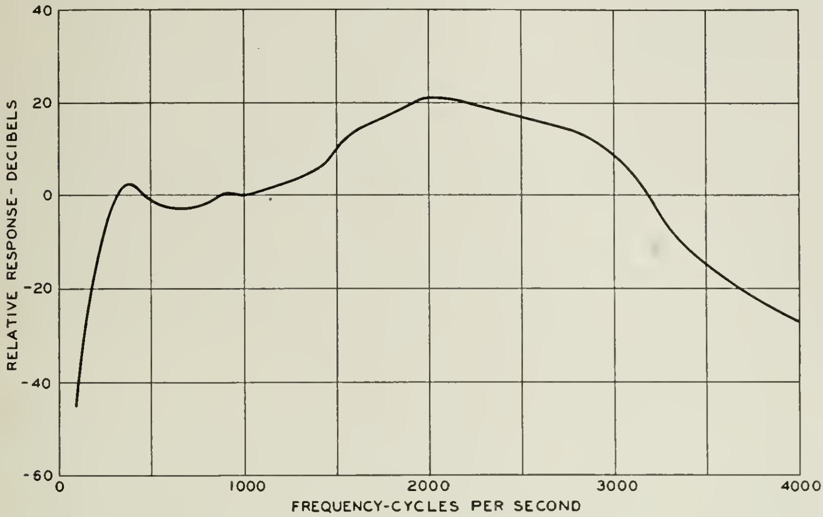


Fig. 4—New station-set characteristics (including two one-mile 24-gauge loops connected by distortionless trunk).

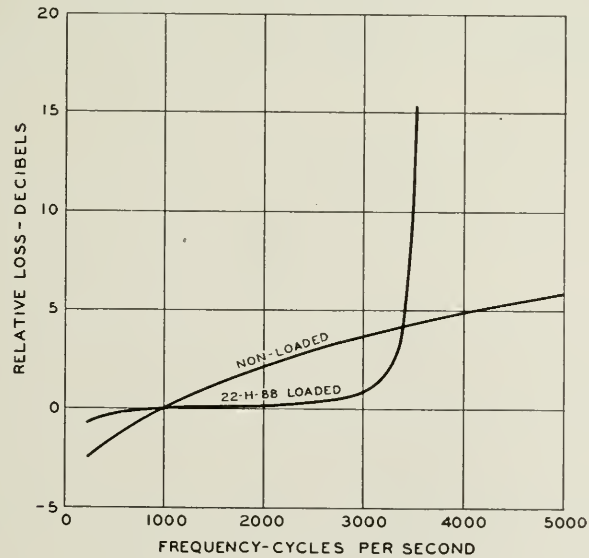


Fig. 5—Toll connecting trunk characteristics.

fraction of a mile to 10 miles or more, with a corresponding effect on the transmission characteristic. It will be noted that the effective 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

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

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

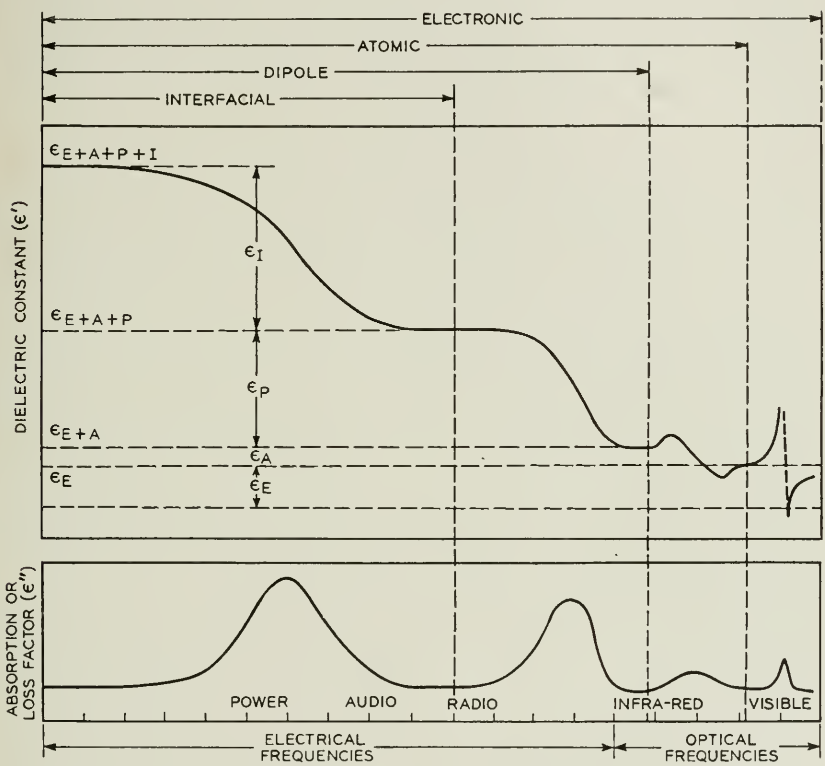


Fig. 1—Schematic diagram of variation of dielectric constant and dielectric absorption with frequency for a material having electronic, atomic, dipole and interfacial polarizations.

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 $f_m = 1/2\pi\tau$.

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

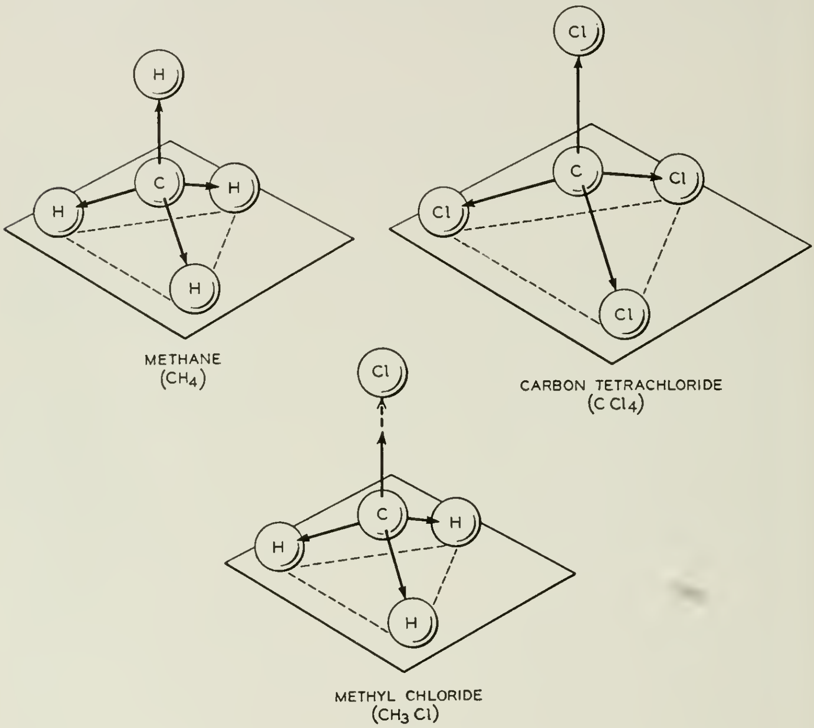


Fig. 2—Methane and carbon tetrachloride are non-polar molecules each having four equal vector moments whose sum is zero. Methyl chloride is polar because the sum of the vector moments is not zero.

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.

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 high-

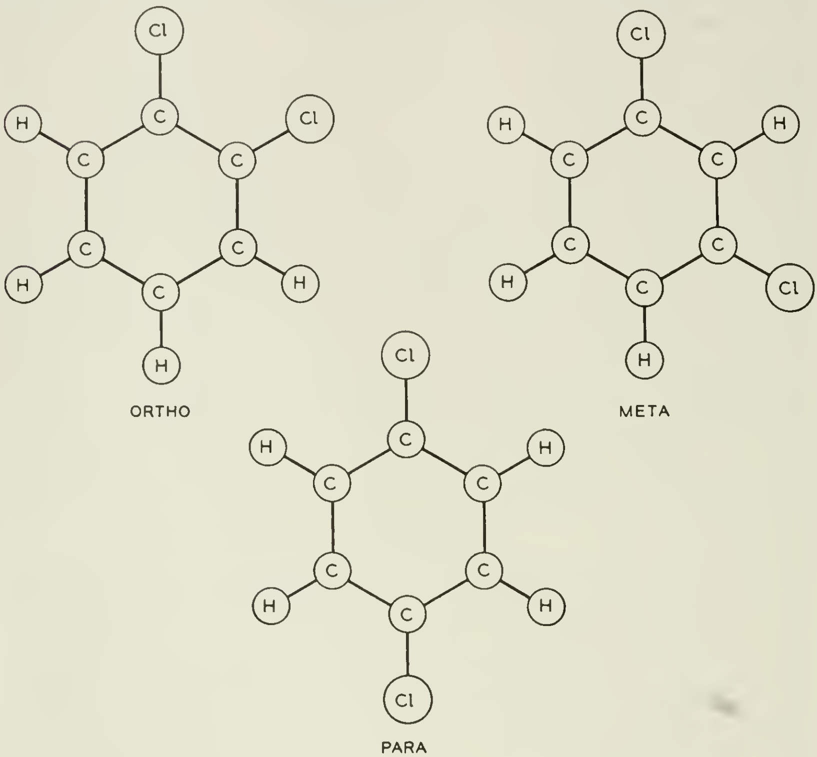
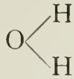


Fig. 3—Ortho dichlorobenzene being the more asymmetrical has a higher electric moment than the meta isomer; the para isomer which is symmetrical has zero electric moment.

est 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.

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 non-mathematical 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.

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

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

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

we have for the current in the network

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

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

$$I = e^{i\int\Omega dt}\cdot\int_0^\infty M(t,\tau)e^{-i\omega\tau}A'(\tau)d\tau + I_T \quad (10)$$
$$= Y(i\omega, t)e^{i\int\Omega dt} + I_T,$$

where

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

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

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

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

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

[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]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$A_n \exp(i\omega_c t + i\omega_n t + i\theta_n). \tag{60}$$

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

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

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

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

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

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

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

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

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

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

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

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

$$\left| \lambda \int_0^T s dt \right| \quad (4a)$$

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

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

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 (2b).

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

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

the noise is representable as the Fourier integral

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

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

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

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

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

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

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

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

Integrating with respect to ω_n, we have

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

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

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

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

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

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

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

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

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

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

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

Consequently, by the Fourier integral energy theorem,

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

and

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

By simple transformations (28*b*) becomes

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

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

We note for reference that

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

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 $\sqrt{\overline{\Delta A_1^2} + \overline{\Delta A_2^2}}$ where A 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 $\bar{K}$), is, approximately

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

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,

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

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

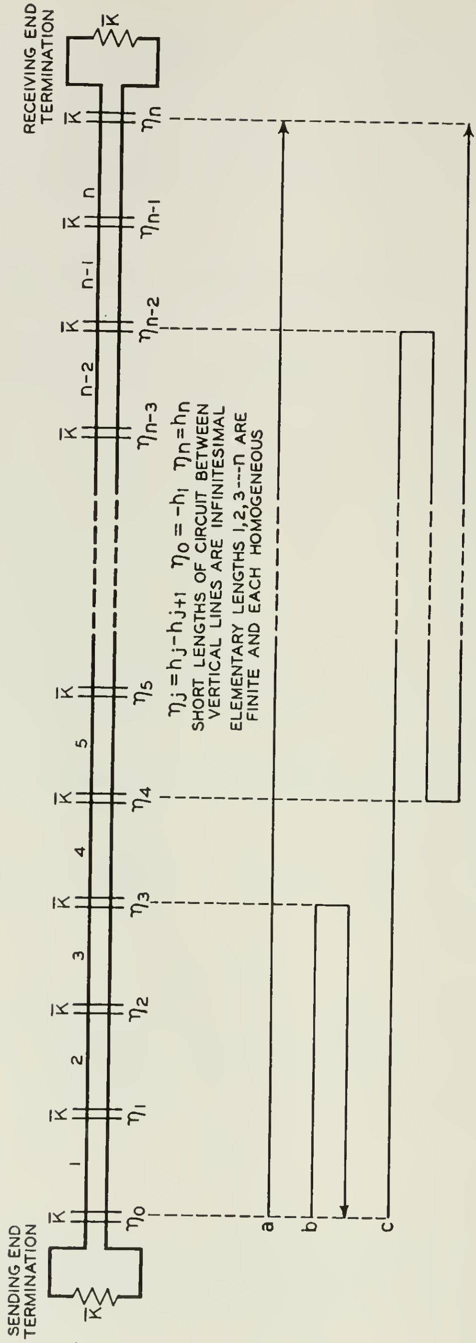


Fig. 1—Inhomogeneous line divided into elementary segments.

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

$$\sqrt{\overline{(\Delta A_1 - \overline{\Delta A_1})^2}} = \frac{\phi^2 \delta^2 \sqrt{n}}{8\sqrt{2}\epsilon^{3/2}} \frac{d\phi}{df} \Delta f. \tag{9}$$

Expressed in terms of the impedance irregularities this amounts to

$$\sqrt{\overline{(\Delta A_1 - \overline{\Delta A_1})^2}} = \left[\frac{\sqrt{\overline{\Delta K_r^2}}}{\overline{K}} \right]^2 \frac{\pi T}{\sqrt{\Lambda}} \Delta f, \tag{10}$$

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 $\pm$ 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:

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

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

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

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

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

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

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

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.

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

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

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

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

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

where F_k is the distribution curve of h_k, and

$$\int_{-\infty}^{\infty} F_k dh_k = 1, \tag{17}$$

$$\int_{-\infty}^{\infty} h_k F_k dh_k = 0. \tag{18}$$

Assuming the h's all have equal distribution curves:

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

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

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

and

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

Likewise

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

except that

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

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

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

$$\overline{H}_0 = \int\int\cdots\int \sum_{j=0}^{n} (h_j^2 - 2h_j h_{j+1} + h^2{}_{j+1}) \times F_1F_2\cdots F_n dh_1 dh_2\cdots dh_n = 2n\overline{h^2}, \quad (25)$$

$$\overline{H}_1 = \int\int\cdots\int \sum_{j=0}^{n-1} (h_j - h_{j+1})(h_{j+1} - h_{j+2}) \times F_1F_2\cdots F_n dh_1 dh_2\cdots dh_n$$

$$= \int\int\cdots\int \sum_{j=0}^{n-1} (-h^2{}_{j+1}) F_1F_2\cdots F_n dh_1 dh_2\cdots dh_n = -n\overline{h^2}, \quad (26)$$

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

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:

$$\overline{E_{cr}^2} = \int\int\cdots\int \left(-\frac{1}{2}H_0 - \sum_{m=1}^{n} H_m B^m \cos m\phi\right)^2$$
$$\times F_1F_2\cdots F_n dh_1 dh_2\cdots dh_n$$
$$= \int\int\cdots\int \frac{1}{4}\sum_{p=0}^{n}\sum_{q=0}^{n}(h_p - h_{p+1})^2(h_q - h_{q+1})^2$$
$$\times F_1F_2\cdots F_n dh_1 dh_2\cdots dh_n$$
$$+ \int\int\cdots\int \sum_{m=1}^{n} B^m(\cos m\phi)\sum_{p=0}^{n}\sum_{q=0}^{n-m}(h_p - h_{p+1})^2$$
$$\times (h_q - h_{q+1})(h_{q+m} - h_{q+m+1})F_1F_2\cdots F_n dh_1 dh_2\cdots dh_n$$
$$+ \int\int\cdots\int \sum_{r=1}^{n}\sum_{s=1}^{n} B^{r+s}(\cos r\phi)(\cos s\phi)$$
$$\times \sum_{p=0}^{n-r}\sum_{q=0}^{n-s}(h_p - h_{p+1})(h_{p+r} - h_{p+r+1})(h_q - h_{q+1})$$
$$\times (h_{q+s} - h_{q+s+1})F_1F_2\cdots F_n dh_1 dh_2\cdots dh_n. \quad (30)$$

Multiplying the factors containing the h's as indicated in (30) gives terms containing $h_a h_b h_c h_d$ 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:

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

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:

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

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

$$E_{cr} - \bar{E}_{cr} = -\frac{1}{2}\left[\sum_{j=0}^{n}(h_j - h_{j+1})^2\right] - \left[\sum_{m=1}^{n} H_m B^m \cos m\phi\right] + [1 - B\cos\phi]n\overline{h^2}, \tag{2}$$

$$\begin{aligned}
\lambda &= -\left[n\overline{h^2}\frac{d(B\cos\phi)}{df} + \sum_{m=1}^{n} H_m \frac{d(B^m \cos m\phi)}{df}\right]\Delta f \\
&= -\left[n\overline{h^2}\left\{B\frac{d\cos\phi}{df} + (\cos\phi)\frac{dB}{df}\right\}\right. \\
&\qquad \left. + \sum_{m=1}^{n} H_m\left\{B^m\frac{d\cos m\phi}{df} + (\cos m\phi)\frac{dB^m}{df}\right\}\right]\Delta f \\
&= \left[n\overline{h^2}(BQ\sin\phi - D\cos\phi) + \sum_{m=1}^{n} mH_m\right. \\
&\qquad \left. \times (B^mQ\sin m\phi - B^{m-1}D\cos m\phi)\right]\Delta f,
\end{aligned} \tag{3}$$

where $Q = d\phi/df$ and $D = dB/df$.

$$\overline{\lambda^2} = \int\int \cdots \int \lambda^2 F_1 F_2 F_3 \cdots F_n dh_1 dh_2 dh_3 \cdots dh_n$$

$$= \int\int \cdots \int n^2 \overline{h^2}^2 (BQ \sin \phi - D \cos \phi)^2 (\Delta f)^2 \times F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n$$

$$+ \int\int \cdots \int 2n\overline{h^2}(BQ \sin \phi - D \cos \phi) \times \left[\sum_{m=1}^{n} mH_m(B^m Q \sin m\phi - B^{m-1} D \cos m\phi) \right] \times (\Delta f)^2 F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n$$

$$+ \int\int \cdots \int \sum_{r=1}^{n} \sum_{s=1}^{n} rs(B^r Q \sin r\phi - B^{r-1} D \cos r\phi) \times (B^s Q \sin s\phi - B^{s-1} D \cos s\phi) \left[\sum_{p=0}^{n-r} \sum_{q=0}^{n-s} (h_p - h_{p+1}) \times (h_{p+r} - h_{p+r+1})(h_q - h_{q+1})(h_{q+s} - h_{q+s+1}) \right] (\Delta f)^2 \times F_1 F_2 \cdots F_n dh_1 dh_2 \cdots dh_n. \quad (4)$$

By methods similar to those employed in Appendix II it follows that

$$\overline{\lambda^2} = \Bigg[(BQ \sin \phi - D \cos \phi)^2 \overline{h^4} + \Bigg[-2(BQ \sin \phi - D \cos \phi)^2 + \frac{(B^2Q^2 + D^2)(3 - 8B \cos \phi + \{6B^2 - 2B^4\} \cos^2 \phi + B^4)}{(1 - B^2)^3} - \left(Q^2 - \frac{D^2}{B^2}\right)\left(1 - \frac{1 + 6B^2 - 3B^4}{(1 + 2B \cos \phi + B^2)^3} - \frac{6B(1 + B^2) \cos \phi + 6B^2(1 + B^2) \cos^2 \phi + 4B^3 \cos^3 \phi}{(1 + 2B \cos \phi + B^2)^3}\right) - 2BQD \frac{\{6(1 + B^2) \cos \phi + 4B \cos^2 \phi + 8B\} \sin \phi}{(1 + 2B \cos \phi + B^2)^3} \Bigg] \times (\overline{h^2})^2 \Bigg] n(\Delta f)^2. \quad (5)$$

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):

$$\begin{aligned} S = \frac{1}{4}\Bigg[& (BQ \sin\phi - D\cos\phi)^2 \\ & + \frac{(B^2Q^2 + D^2)(3 - 8B\cos\phi + \{6B^2 - 2B^4\}\cos^2\phi + B^4)}{(1 - B^2)^3} \\ & - \left(Q^2 - \frac{D^2}{B^2}\right)\Bigg(1 - \frac{1 + 6B^2 - 3B^4}{(1 + 2B\cos\phi + B^2)^3} \\ & - \frac{6B(1 + B^2)\cos\phi + 6B^2(1 + B^2)\cos^2\phi + 4B^3\cos^3\phi}{(1 + 2B\cos\phi + B^2)^3}\Bigg) \\ & - 2BQD\,\frac{\{6(1 + B^2)\cos\phi + 4B\cos^2\phi + 8B\}\sin\phi}{(1 + 2B\cos\phi + B^2)^3}\Bigg]^{\frac{1}{2}}(\Delta f) \end{aligned} \tag{12}$$

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

$$\sqrt{\overline{(\Delta\Lambda - \overline{\Delta\Lambda})^2}} = \frac{Q\phi^2\delta^2\sqrt{n}}{8\epsilon\sqrt{2\epsilon}}(\Delta f). \tag{15}$$

Transocéanic 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

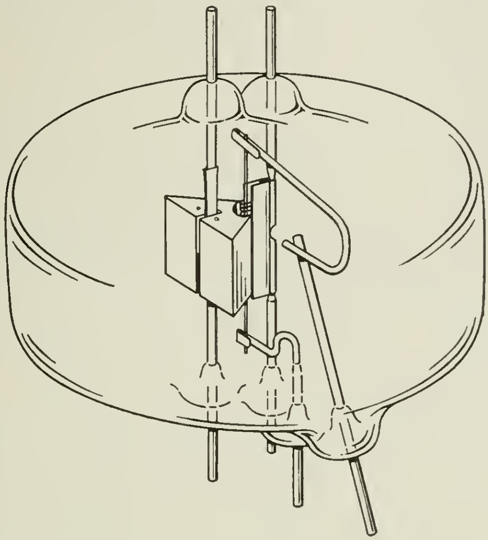


Fig. 2—Section view of one of the double-lead tubes.

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.

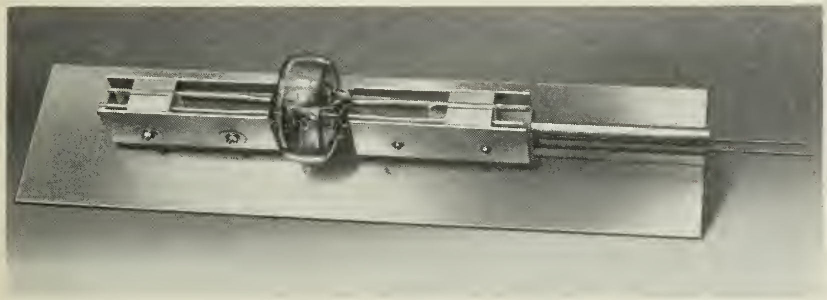


Fig. 3—Typical oscillator circuit.

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

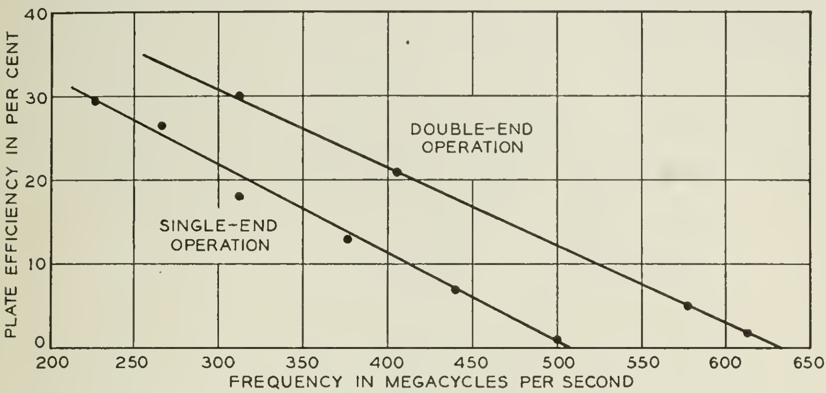


Fig. 4—Comparison plot of output efficiency for the large tube when operated single-ended and double-ended.

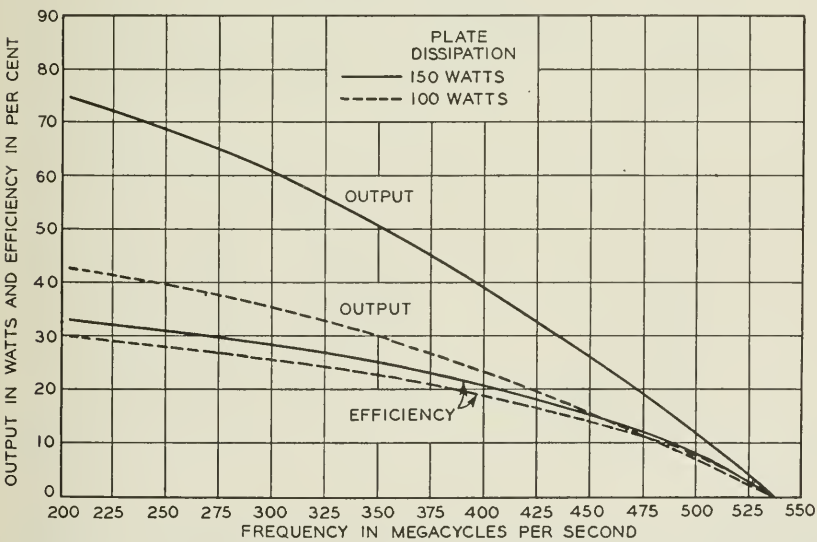


Fig. 5—Output and efficiency as a function of frequency for the large tube.

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 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

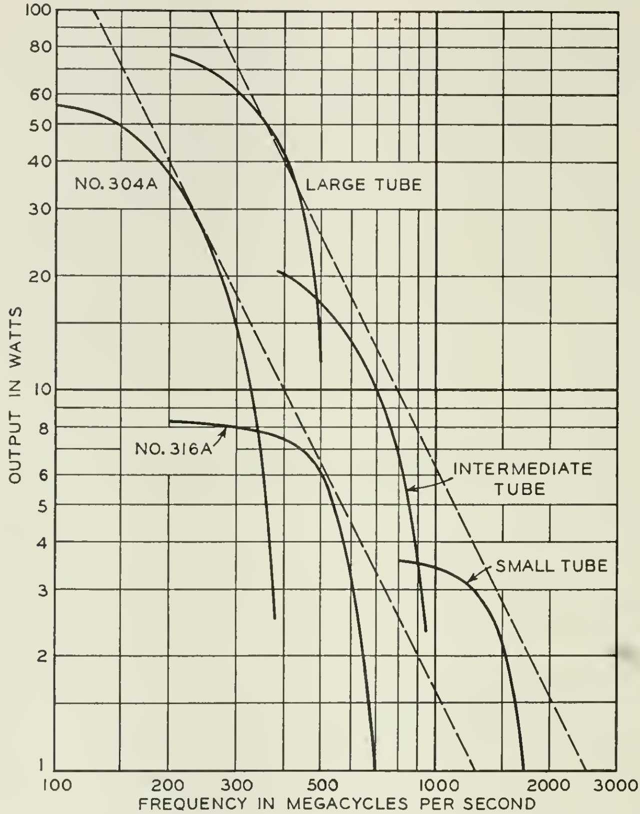


Fig. 6—Comparison plot of the outputs of the double-lead tubes and of commercially available tubes.

outstanding advantages over the more usual "capacitance" schemes. These advantages become even more pronounced with the availability of the double-lead tube.

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

$$\frac{E}{2E} = \frac{W}{1+\tau^2} + \frac{1}{1-\tau^4}\left[\frac{1 - \tau e^{(2\pi id/\lambda)(1-1/\tau)}}{2\pi id/\lambda} + \frac{1 - \tau^2 e^{(2\pi id/\lambda)(1-1/\tau)}}{(2\pi id/\lambda)^2}\right]. \quad (17)$$

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

$$E \to \left[\frac{1}{1-\tau^4}\frac{1+\tau^2}{2\pi\tau^2 id/\lambda}\right]\left[1 - \tau^3 e^{\frac{2\pi id}{\lambda}\left(1-\frac{1}{\tau}\right)}\right]2E_0. \quad (19)$$

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.

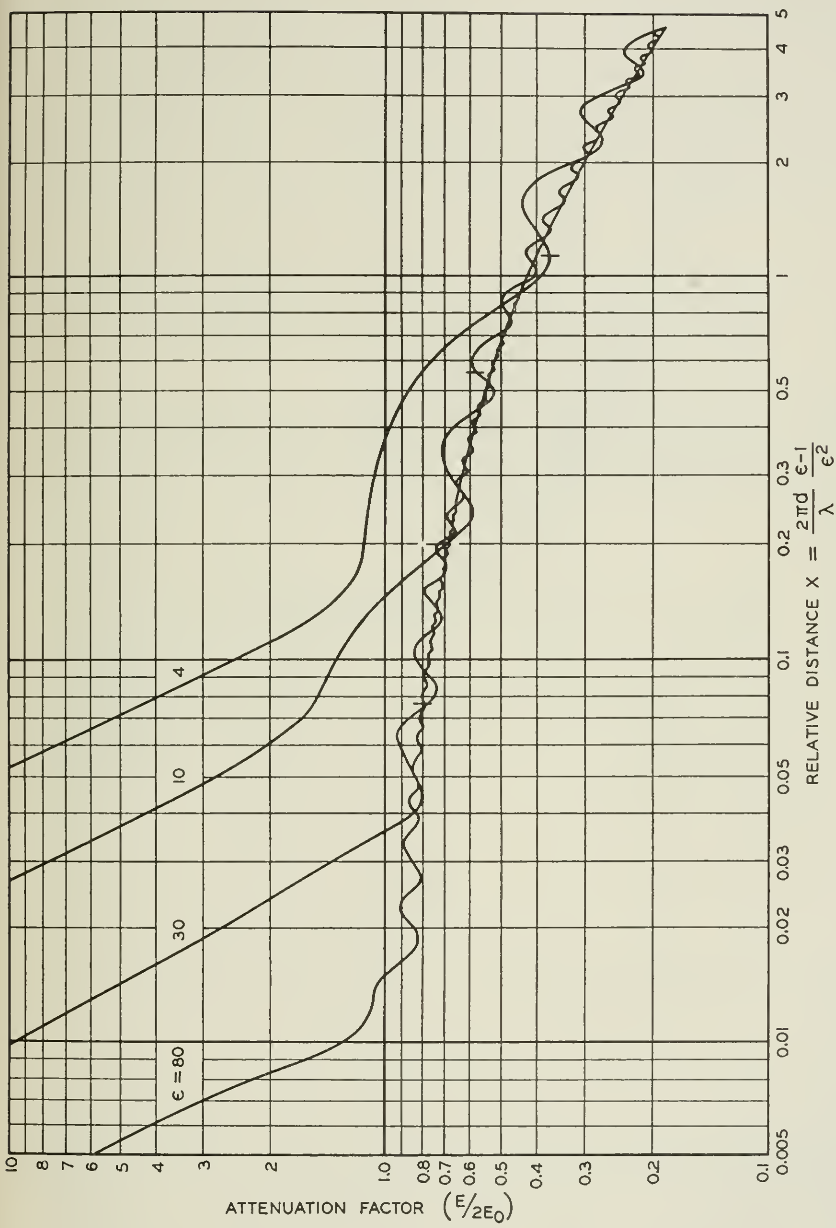


Fig. 3—Attenuation factor for radio propagation over a dielectric plane. The number on each curve gives the value of the dielectric constant to which it applies.

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.

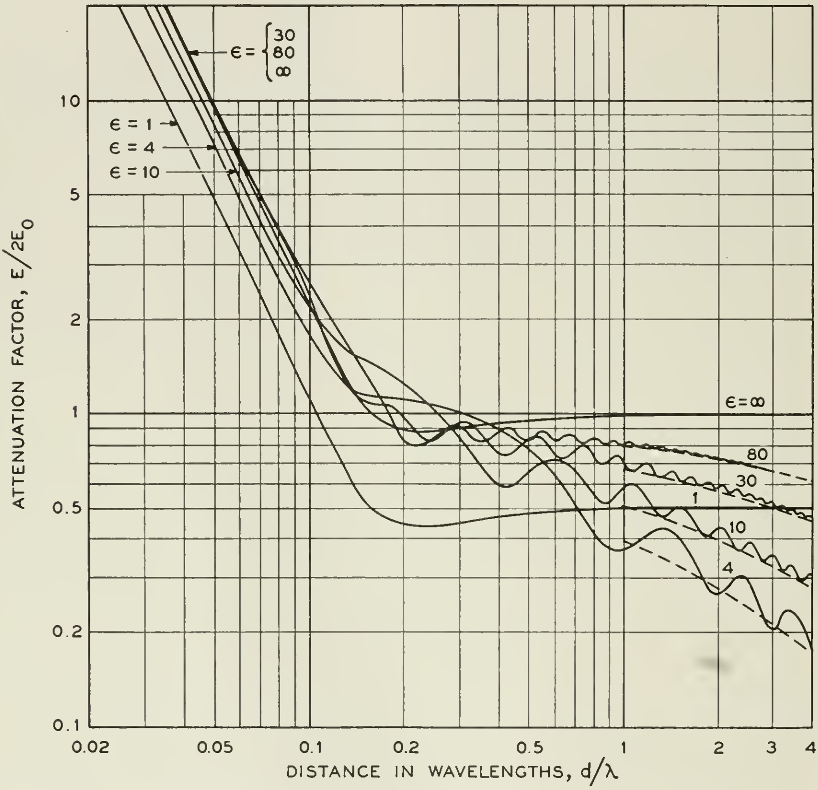


Fig. 3A—Variation of attenuation factor with distance in wave-lengths for transmission over a dielectric plane. For d/λ small,

$$E/2E_0 = \left(\frac{\epsilon}{\epsilon + 1}\right) \Big/ \left(\frac{2\pi d}{\lambda}\right)^2.$$

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 $A\mathbf{\Pi}_0$ becomes

$$-\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial d^2}+\frac{1}{d}\frac{\partial}{\partial d}\right)A\mathbf{\Pi}_0=\left(\frac{A}{1+\tau^2}+\frac{1}{1-\tau^4}\left[\frac{1}{2\pi i d/\lambda}+\frac{1}{(2\pi i d/\lambda)^2}\right]\right)\mathbf{\Pi}_0,$$

when the value of $y = (1 + \tau^2)A\mathbf{\Pi}_0$ 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 $D\mathbf{\Pi}_0 = y/(1 + \tau^2)$ is found to be

$$-\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial d^2}+\frac{1}{d}\frac{\partial}{\partial d}\right)D\mathbf{\Pi}_0=\left(\frac{D}{1+\tau^2}-\frac{1}{1-\tau^4}\left[\frac{\tau}{2\pi i d/\lambda}+\frac{\tau^2}{(2\pi i d/\lambda)^2}\right]\right)\mathbf{\Pi}_0.$$

Adding these two relations gives an expression for $\left(\frac{\partial^2}{\partial d^2}+\frac{1}{d}\frac{\partial}{\partial d}\right)\mathbf{\Pi}$, where

$$\mathbf{\Pi} = 2(A + D)\mathbf{\Pi}_0,$$

which when substituted in the above equation for E and the result divided by $2E_0$, where $E_0 = - 240 i\pi^2\mathbf{\Pi}_0/\lambda$, gives equation (17). Since E_0 is the inverse distance component of the free space field, this relation for 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.