The Bell System Technical Journal

Devoted to the Scientific and Engineering Aspects of Electrical Communication

Published quarterly by the American Telephone and Telegraph Company, through its Information Department, in behalf of the Western Electric Company and the Associated Companies of the Bell System

Address all correspondence to the Editor
Information Department
AMERICAN TELEPHONE AND TELEGRAPH COMPANY
195 BROADWAY, NEW YORK, N. Y.

50c. Per Copy *$1.50 Per Year*

Vol. I • JULY, 1922 No. 1

FOREWORD

MODERN industry is characterized by the extent to which scientific research and technique based on precise study have contributed to its progress. So complete has been the adaptation of and reliance on scientific research in many industries that it is difficult at this time to visualize the state of affairs of two or three decades ago, when substantially all industry on its technical side was dependent for advancement on cut-and-try, rule-of-thumb, methods of development. Today in many industries the management would not think of embarking on a new project without consulting their research engineers.

Many industries have proved the benefits to be derived from the utilization of that organized knowledge provided both in the fields of the physical sciences and in those newer fields which have to do with psychology and economics. There are still greater numbers of industrial organizations where the adoption of scientific methods has been slow. However, the time will undoubtedly come when every industry will recognize the aid it can derive from scientific research in some form as it now recognizes its dependence for motive power on steam or electricity rather than on muscular activity.

Upwards of one hundred years ago there was adopted in earnest by scientific men, principally in university laboratories, the program of searching deeper into the unknown, to discover new principles and new relationships of a kind which had at the time very little apparent practical interest to mankind as a whole.

Out of this work, and in time, have grown entirely new industries. From the fact that these industries sprang directly from the research laboratory, it was inevitable that they should be conspicuous because of the number of their men trained in the methods of scientific research. Equally inevitable was it that these new fields of endeavor, originating as they did and being staffed as they were, should be the ground where industrial research would find its first and largest development. And not the least of the advantages which obtained in these newer industries was the absence of age-long traditions tending to ultra-conservatism as to new undertakings, and more particularly as to the employment of the new types of mind.

The results up to the present indicate clearly that the electrical and chemical fields in industry as we know them today, are the places where the greatest advances have been made in the utilization of research methods and research men. Other, older and more basic industries are rapidly following the general path marked out by the successes already obtained in these fields. Hence, it is expected that shortly all industrial activities will be based on the results obtained by trained investigators, using the tools of modern scientific investigation.

Just as applied electricity is a leading exemplar of the benefits to be obtained by an intelligent use of scientific knowledge, so electrical communication of intelligence is a leading exemplar in the field of applied electricity. This branch of applied electricity is a pioneer among those recognizing the practical value of scientific research. It is interesting to note that electrical communication is credited with having organized a research laboratory prior to the first university course in electrical engineering.

More than ever before, the communication engineer must seek exact solutions of his problems. If his results do not always attain the certainty he desires, the reason is the absence of complete knowledge with regard to one or more essential facts. But true knowledge of what things limit the solution of a problem is frequently more than half the battle of obtaining the missing facts. Sometimes these unknown facts can be obtained by a search through the remoter parts of the vast scientific storehouses which have been built in times past. Frequently, however, the search discloses the entire absence

of the thing sought for, and new researches are begun with definite ends in view. Thus it has come about that the communication engineer has become an original investigator and is extending the boundaries of human knowledge and supplementing the advances of pure science to find solutions for his various and sundry problems.

Hence, while well equipped physical and chemical laboratories are still a necessary part of the communication engineer's equipment, he is equally active in pushing his investigations in many other directions. Questions involved in the making of proper rate schedules and adequate fundamental plans for new construction are originating profound researches in such fields as political science, psychology and mathematics. A casual examination of recent technical literature dealing with electrical communication would show articles which touch upon almost every branch of human activity, which we designate as science.

With this intense and growing interest in the proper application of scientific methods to the solution of the problems of electrical communication, it is natural that a widespread desire should have arisen for a technical journal to collect, print or reprint, and make readily available the more important articles relating to the field of the communication engineer. These articles are now appearing in some fifteen or twenty periodicals scattered throughout the world and in the majority of instances receive their first and last printing in these widely separated mediums. The need already felt for such a journal will grow keener as new developments extend the scope of the art and the specialization of its engineers of necessity increases. It is hoped that the BELL SYSTEM TECHNICAL JOURNAL will fill this need, and as implied above, it is intended that the range of subjects treated in the JOURNAL will be as broad as the science and technique of electrical communication itself.

While many of the articles which will appear in the JOURNAL will be original presentations of some phase of the research or development or other technical work of the Bell System, it is not intended that the JOURNAL should be the sole means by which this work is presented. Just as in the past, original articles and papers will continue to be presented before various societies and in different technical and non-technical magazines. Moreover, the JOURNAL will reprint articles on important research and development work in the communication field generally so that the results of such work may be given greater publicity and become of greater value to communication engineers.

A New Type of High Power Vacuum Tube

By W. WILSON

SYNOPSIS: The type of vacuum tube described in the present article is likely to become one of the most remarkable devices of modern electrical science. Vacuum tubes capable of handling small amounts of power have been extensively used during the past few years as telephone repeaters and as oscillators, modulators, detectors and amplifiers in radio transmission and other fields. Practically all such tubes have depended upon thermal radiation from the plates to dissipate the electrical energy which the device necessarily absorbs during its operation. With present methods of construction, and using glass for the containing bulb, a fairly definite upper limit can be set for the power which a radiation cooled tube can handle; as the author points out, this limit gives a tube capable of delivering about 1 to 2 k. w. when used as an oscillator.

Contrasted with this, one of the water-cooled vacuum tubes described herewith, although scarcely two feet in length and weighing only ten pounds, is capable of delivering 100 k. w. of high frequency energy. Another tube of similar construction, but somewhat smaller in size, and capable of delivering about 10 k. w. is also described. It is expected that these water-cooled tubes will find important applications in radio telephony and telegraphy.

Although the principle of operation of the water-cooled tube described in this article is identical from an electrical point of view with that of the small tubes which are now so very familiar, their practicability has only been made possible by a new and striking development in the art of sealing metal to glass. In the case of the 100 k. w. tube the seal between the cylindrical copper anode and glass portion is 3.5 inches in diameter.

The remarkable character of these copper-in-glass seals is evidenced by the fact that they do not depend upon a substantial equality between the coefficient of expansion of the metal and glass. To Mr. W. G. Houskeeper of the Bell System Research Laboratory at the Western Electric Company, goes the credit for developing the copper-in-glass seals. As the article brings out, Mr. Houskeeper has also invented means for sealing heavy copper wire and strip through glass in such a way that the best vacua can be maintained under wide changes of temperature.—*Editor.*

THE development of wireless telephony and the use of continuous wave transmission in wireless telegraphy have led to the general adoption of the vacuum tube as the generator of high frequency currents in low power installations.

The ordinary form of vacuum tube is, however, ill suited for the handling of large amounts of power, and at the large wireless stations where the plant is rated in hundreds of kilowatts either the arc or the high frequency alternator is used.

The undoubted advantages to be derived from the use of vacuum tubes, especially in the field of wireless telephony where the output power must be modulated to conform to the intricate vibration pattern of the voice, has led to a demand for tubes capable of handling amounts of power comparable with those in use at the largest stations.

That the development of such tubes was of great importance was recognized by the engineers of the Bell Telephone System in the early days of the vacuum tube art. The experiments at Arlington,

Virginia, in which speech was first transmitted across the Atlantic to Paris and across the Pacific to Honolulu, required the use of nearly 300 of the most powerful tubes then available, each capable of handling about 25 watts, and the difficulties encountered in operating so many tubes in parallel gave added impetus to the development of high power units.

It is the object of the present paper to deal with the various steps in the development of high power tubes as carried out in the Bell System research laboratories at the Western Electric Company.

The usual type of vacuum tube consists of an evacuated glass vessel in which are enclosed three elements, the filament, the plate, and the grid. When the tube is in operation an electron current flows between the filament which is heated by an auxiliary source of power and the plate, the magnitude of this current being controlled by the grid.

The passage of the current through a thermionic tube is accompanied by the dissipation in the plate of an amount of power which is comparable to the power delivered to the output circuit and which manifests itself in the form of heat. This causes the temperature of the plate in the usual type of tube to rise until the rate of loss of heat by radiation is equal to the power dissipated. Some of the heat liberated by the plate is absorbed by the walls of the containing vessel which consequently rise in temperature. These factors, together with a consideration of the size of plate that can be conveniently suspended inside a glass bulb and the size of glass bulb that can be conveniently worked, set a limit of about 1 to 2 k. w. for the power that can be dissipated in the plate of a commercial vacuum tube of this type. The plates are generally constructed of molybdenum or some other refractory metal and the containing vessel made of hard glass.

The use of quartz as the containing vessel offers certain advantages which tend to raise the power limit somewhat and this material has been used for power tube purposes in England.

It is apparent then that in the development of vacuum tubes capable of handling large amounts of power means other than radiation must be used for removing the heat dissipated at the plate, and development of tubes along these lines was undertaken by Dr. E. R. Stoekle and Dr. O. E. Buckley.

Dr. Stoekle had already worked for some years on the problem of removing the heat dissipated at the anode of a thermionic tube by making the anode a part of the outside wall of the vessel and thus making it possible to convey the heat directly away from it by means of circulating water. This was clearly the right principle but as is obvious to those who are familiar with these devices, great difficulties

presented themselves in the mechanical construction of large tubes in which vacuum tight joints must be made and maintained between glass and large masses of metal. The importance of the problem, however, was such that Stoekle and Buckley pushed on in the face of difficulties to the construction of tubes which could handle kilowatts where previous tubes could only handle watts.

A step in the direction of overcoming these difficulties was made by Messrs. Schwerin and Weinhart, who were working with Dr. Buckley on the problem, and who suggested that the anode might be made in the form of a tube or thimble of platinum sealed into a glass vessel and kept cool by passing water through it.

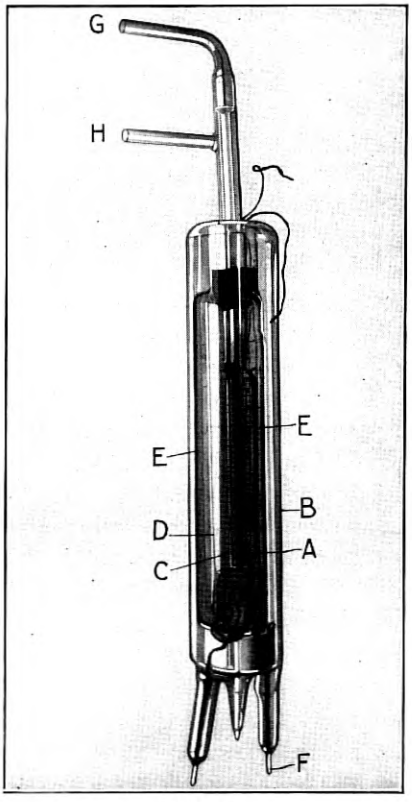


FIG. 1

This suggestion led to the development of a tube which, although not the one finally adopted, is discussed in some detail since it was the first one to be pushed to such a point as to give promise of economical commercial manufacture.

The tube is shown in Fig. 1. The anode consists of a platinum cylinder A, 7″ long and .625″ wide, which is sealed into the center of the glass cylinder B. The end of the platinum cylinder remote from the seal is closed. The anode is surrounded by the grid C and by the filament D, which are supported by the glass arbors E. The current for the filament is led into the tube through the platinum thimbles F.

The anode is kept cool by means of a supply of water passing into the anode through the tube G and leaving by the tube H.

A number of tubes having this general type of construction were made up and it was found possible to dissipate as much as 15 k. w. in the anode.

As soon as the pressure of work more directly connected with the necessities of the war would permit, Mr. W. G. Houskeeper and Dr. M. J. Kelly undertook the further improvement of the water-cooled tube, the former assuming the task of developing the mechanical structure, and the latter that of determining the electrical design and the process of tube exhaust.

Mr. Houskeeper adopted into the construction of the tube a remarkable type of vacuum seal which he had previously developed. These seals are made between glass and metal and can be made in any desired size. They are capable of withstanding repeated heating and cooling over wide ranges of temperature, from that of liquid air to 350° C, without cracking and without impairment of their vacuum holding properties.

It is no exaggeration to say that the invention of these seals has made possible the construction of vacuum tubes, capable of handling in single units, powers of any magnitude which may be called for in wireless telegraph and telephone transmission.

The underlying principle connected with the making of this seal consists in obtaining an intimate connection between the glass and metal, either by chemical combination or by mere wetting, and in so proportioning the glass and metal portions of the seal that the stresses produced when the seal is heated or cooled will not be great enough to rupture either the glass or the junction between the glass and metal.

The three principal types of seals developed by Mr. Houskeeper are known as the ribbon seal, the disc seal and the tube seal.

If a copper ribbon is directly sealed through glass it is found that the glass and copper adhere along the flat faces of the seal but that ruptures occur along the edges as shown in Fig. 2 (a). This is due to the fact that as the seal cools after being made, the glass in contact with metal is capable of resisting the shearing and tensile stresses

that occur along the faces, while the glass wrapping round the edges of the ribbon is called upon to withstand much greater tensile stresses and gives way. If the edges of the ribbon are sharpened as shown in Fig. 2 (b), a tight seal results, the reason being that the forces of

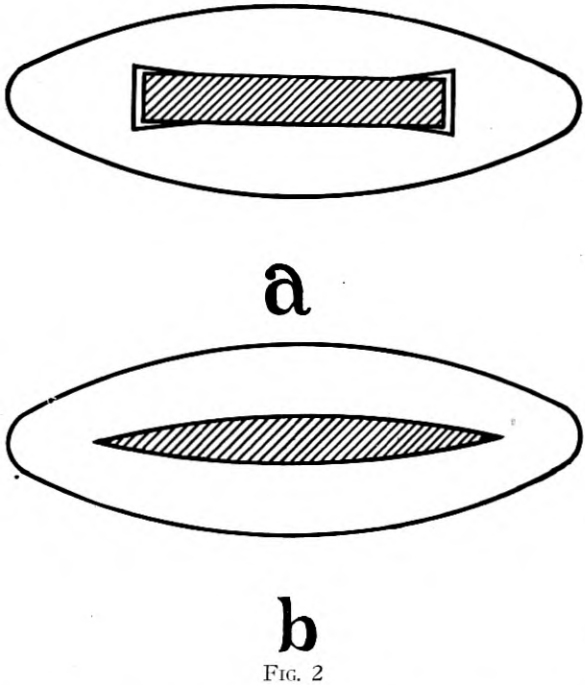


FIG. 2

adhesion between the glass and copper acting along the flat contact faces are sufficient to stretch the thin copper at the edge and prevent its drawing away when cooled. There is a definite relation between the elastic properties of the metal and glass and the angle of edge that can be used for a successful seal.

By proper shaping of the metal ribbon, seals have been successfully made up to very large sizes. Some of these are shown in Fig. 3, the the largest in the photograph being about 1″ in width, and capable of successfully conducting a current of 150 to 200 amperes.

The principles involved in the making of the disc seal are the same as those involved in making the ribbon seal. If a metal disc is sealed wholly into glass the edges must be sharpened or the glass and copper break away from each other as in the case of the ribbon seal.

In the general use to which these seals are put there is no necessity for having the glass surround the circumference of the copper disc

and the necessity for sharpening the edge is obviated by allowing the glass to adhere to the flat portion of the disc only, care being taken to prevent its flowing around the edge. It is necessary to have a ring of glass on both sides of the seal in order to equalize the bending stresses

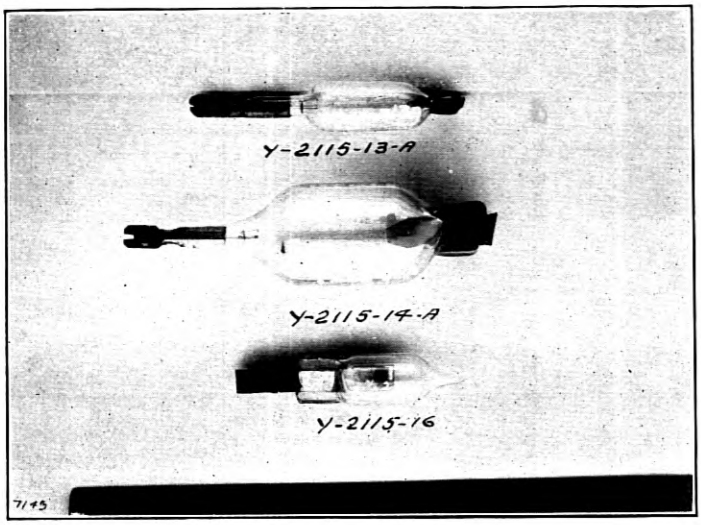


FIG. 3

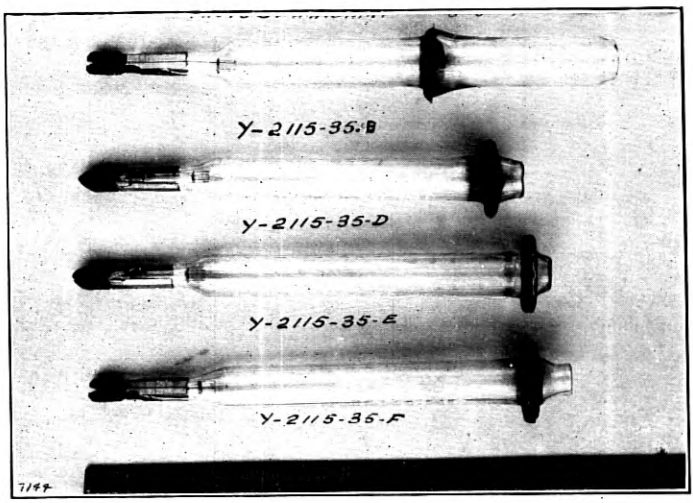


FIG. 4

which would otherwise tend to break the glass and copper away from each other. Successful disc seals have been made with copper up to 1-10″ thick. There is, of course, a certain maximum thickness that can be used for a seal of a given diameter and it is preferable to keep well below this limit.

The seals shown in Fig. 4 close the ends of glass tubes to the other ends of which are sealed pilot lamps for the purpose of testing the vacuum. Tubes sealed in this way have been kept a number of years without any impairment of the vacuum.

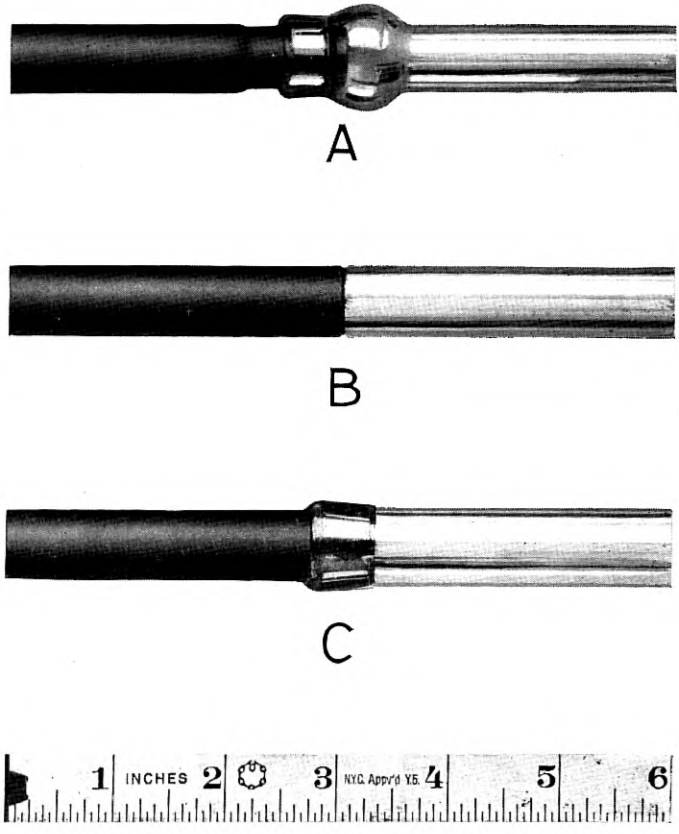


FIG. 5

The third type of seal and the most important in connection with the present problem is the tube seal shown in Fig. 5. This furnishes the means of joining metal and glass tubes end to end and is used in the water-cooled tube to attach the anode to the glass cylinder which

serves to insulate the other tube elements. As in the case of the disc seal, it can be made either with the edge of the metal not in contact with the glass, as shown at A, or with the metal sharpened to a fine edge which is in contact with the glass. The glass may be situated either inside or outside of the metal, see B and C.

The first thermionic tubes in which these seals were embodied were made of copper and were designed to operate at 10,000 volts and to give about 5 k. w. output.

A photograph of one of these tubes is shown in Fig. 6; and the filament grid assembly is shown in Fig. 7.

FIG. 6

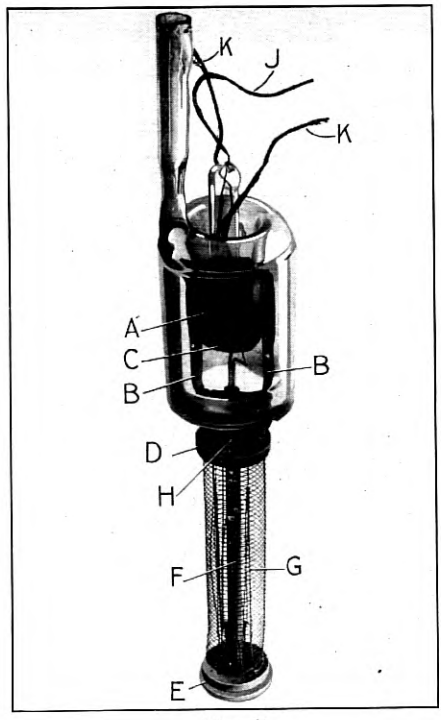


FIG. 7

The anode consists of a copper tube 1.5″ in diameter and 7.5″ long. A copper disc is welded to one end forming a vacuum-tight joint. The other end which is turned down to a knife edge is fused directly to a glass tube.

The filament grid assembly consists of two lavite discs D and E, spaced 5″ apart by a seamless steel tube. The grid F is made in the form of a helix, and is held in position by allowing the ends of the longitudinal wires, to which the turns of the helix are welded, to pass through holes in the lavite blocks D and E. The filament G is mounted between hooks fastened to the lavite blocks and is kept taut by the springs H. The grid lead is shown at J, and the filament leads at K K. In this tube platinum seals are used for the lead wires. The use of the springs H make it necessary to supply the filament with current from the opposite end of the assembly and this is done by passing the current through the steel support tube and returning it through a lead passing through this tube and insulated from it by a quartz tube.

The whole assembly is carried by two supports B B. These supports are welded to a corrugated nickel collar A which grips the glass stem C.

The pumping of these tubes at first presented considerable difficulty, chiefly on account of the large amount of occluded gas contained by the metal parts. This caused the time of pumping of the tube to be very long and a dangerous warping of the internal structure developed owing to the fact that during exhaust the tube elements are maintained at a much higher temperature than they are subjected to during normal operation. The trouble was overcome by heating the various parts of the tube to as high a temperature as possible in a vacuum furnace, prior to the final assembly, and thus getting rid of a large amount of the occluded gases. The anode was preheated before the glass seal was made and the whole filament grid assembly was preheated just before it was mounted on the glass stem. The preheating of the parts brought about an enormous reduction in the time required for pumping and gave a much more uniform product.

Although successful from the standpoint of operation, this tube had several undesirable features that it was thought well to eliminate. In the first place the welding of the end into the tube was not particularly desirable, and in general any troubles that occurred due to leaks in the metal could be traced to this point. Further, in the assembly of the tube there were a very large number of welds to be made which constituted points of weakness at the high temperature necessary for the evacuation of the tubes. It was, therefore, decided to go to a type of tube in which the anode would be drawn in one piece and in which as many welds as possible would be eliminated in the assembly of the internal elements. At the same time it was considered desirable to go to a somewhat larger type of structure in which high

tension insulation could be more easily provided and a larger tube was, therefore, designed capable of delivering 10 k. w. to an antenna at a plate voltage of 10,000 volts.

The final form adopted for this tube is shown in Figs. 8 and 9.

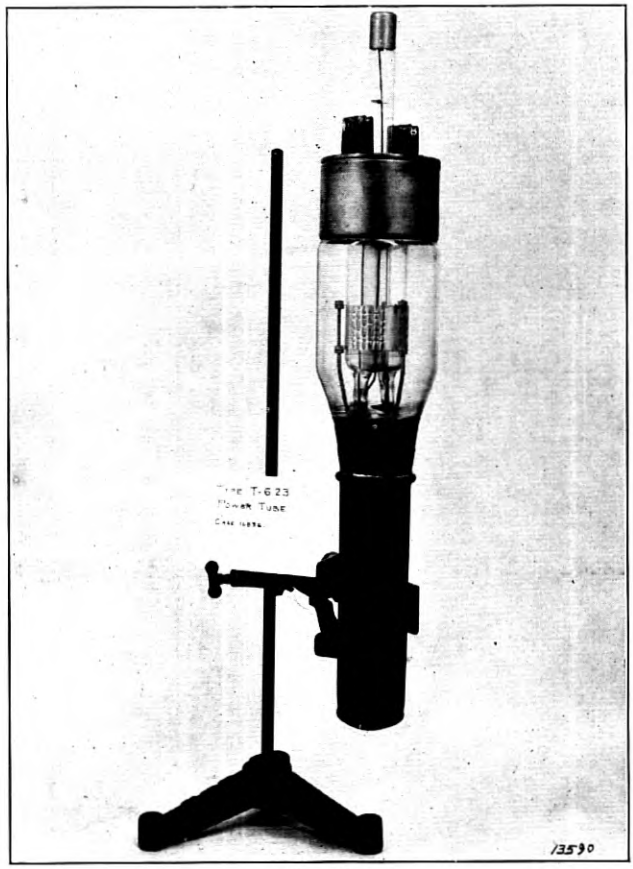


FIG. 8

The anode A is drawn from a piece of sheet copper and is 9″ long and 2″ in diameter. The copper flare B is turned down to a sharp edge and a glass bulb C sealed thereto. The grid and plate assembly is shown at D. The structure is supported by four molybdenum rods, which are threaded and secured by means of nuts to the lavite pieces E and F. The filament is made of 19.5″ of .025 pure tungsten wire purchased from the General Electric Company and is formed and secured to two of the molybdenum rods at G and H. The

power consumed in it during operation is .75 k. w. It is guided by the hooks J. The filament leads are shown at K, K and are led through the glass by the copper disc seals L, L. The grid is a molyb-

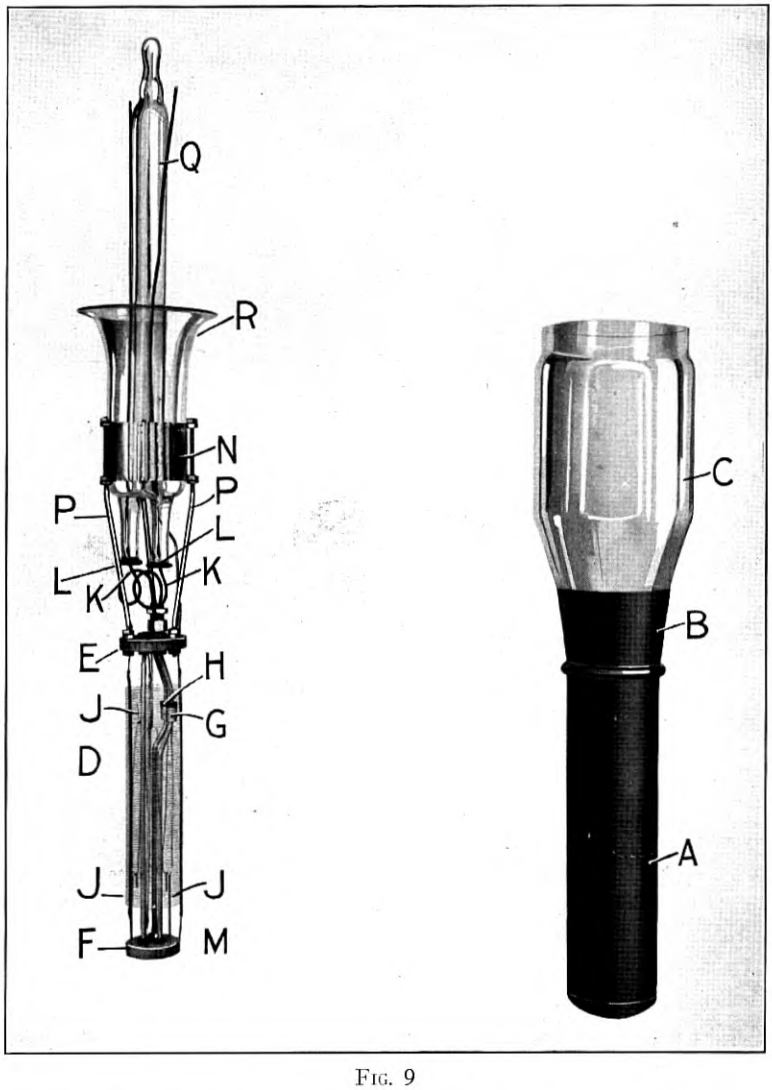


FIG. 9

denum helix and is supported by the molybdenum rods M which are fixed to the lavite block E and slide on the outside of the lavite block F. The whole structure is mounted on the flare R by means of the

nickel collar N and the support rods P. The grid lead is brought out through the tube Q. The tube is completed by sealing together the flare R and the bulb C.

In this tube all welds except those in the collar N are eliminated, the assembly being bolted together. The drawing of the anode does

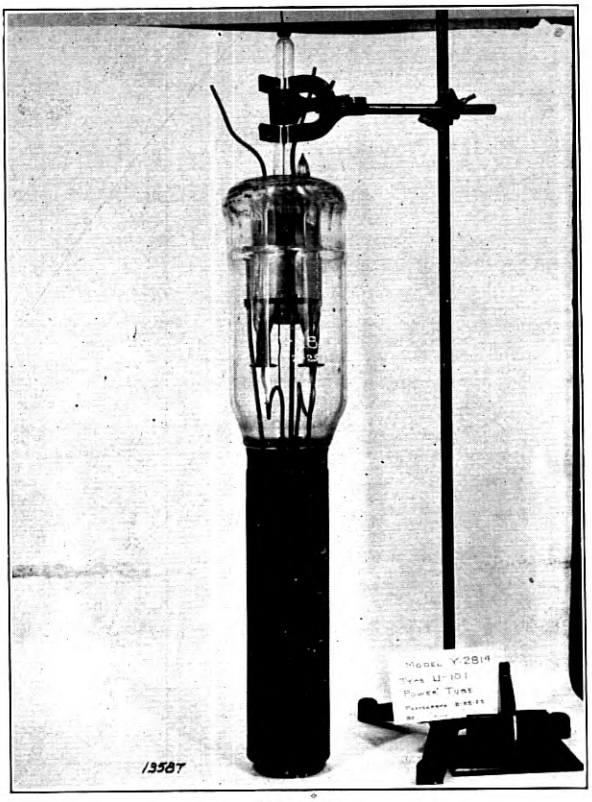


FIG. 10

away with the leaks that were troublesome in the older tubes and the manufacture of the tube can be carried out with certainty.

With this tube as much as 12 k. w. have been obtained in an artificial antenna working at 12,000 volts. This power was obtained at a frequency of 600,000 cycles corresponding to 500 meters wave length. The difficulties of obtaining this amount of power at this frequency using a number of smaller tubes in parallel, are obvious to anyone who is acquainted with the problem. On a D. C. test the anode was found to be capable of dissipating 26 k. w. when cooled with water.

The success which had attended the development of a tube of this high power capacity indicated the possibility of constructing still larger tubes and it was decided to proceed with the development of a tube capable of delivering at least 100 k. w. into an antenna.

The development proceeded with a few minor alterations along the lines of the smaller tube, nominally rated at 10 k. w. and the 100 k. w.

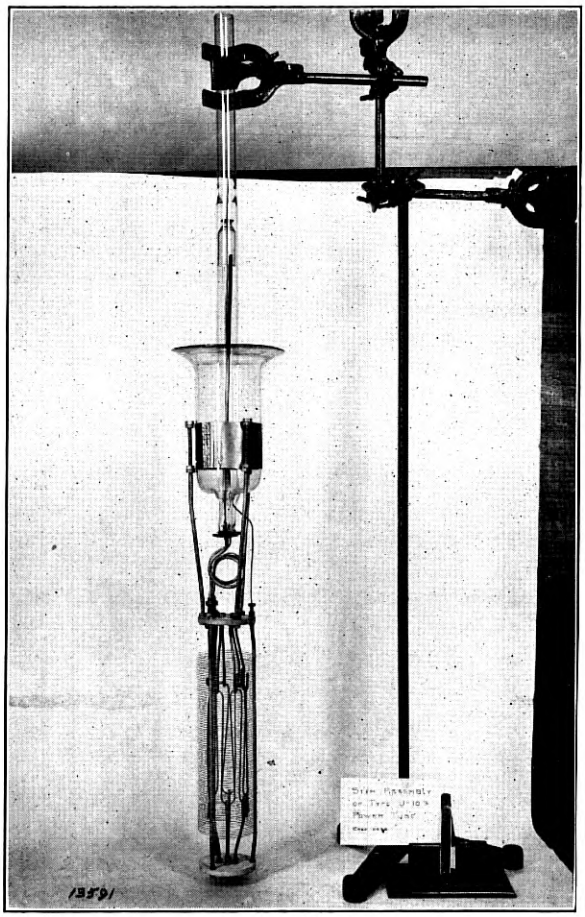


FIG. 11

tube as now developed is shown in Figs. 10 and 11. The anode which is made of a piece of seamless copper tubing closed by a copper disc welded into the end, is 14″ long and 3.5″ in diameter. The filament is of tungsten and is .060″ in diameter and 63.5″ long. The current required to heat it is 91 amperes and the power consumed in it 6 k. w.

The filament leads are of copper rod one eighth of an inch in diameter and are sealed through 1″ copper disc seals. The grid is of molybdenum and is wound around three molybdenum supports.

The handling of the parts of this tube during manufacture presents a task of no mean magnitude and numerous fixtures have been devised to assist in the glass working. It has been found necessary for instance to suspend the anode in gimbals during the making of the tube seal owing to its great weight, and special devices have been made to hold the filament grid assembly in place while it is being sealed in, otherwise the strains produced by its weight cause cracking of the seal.

The significance of this development in the radio art cannot be overestimated. It makes available tubes in units so large that only a very few would be necessary to operate even the largest radio stations now extant, with all the attendant flexibility of operation which accompanies the use of the vacuum tube.

From the standpoint of wireless telephony the development of these high power tubes gives us the possibility of using very much greater amounts of power than have ever been readily available before. The filaments in these tubes have been made so large that the electron emission from them will easily take care of the high peak currents accompanying the transmission of modulated power.

The 100 k. w. tube by no means represents the largest tube made possible by the present development. There is no doubt that if the demand should occur for tubes capable of handling much larger amounts of power they could be constructed along these same lines.

Direct Capacity Measurement[1]

By GEORGE A. CAMPBELL

SYNOPSIS: Direct capacity, direct admittance and direct impedance are defined as the branch constants of the particular direct network which is equivalent to any given electrical system. Typical methods of measuring these direct constants are described with especial reference to direct admittance; the substitution alternating current bridge method, due to Colpitts, is the preferred method, and for this suitable variable capacities and conductances are described, and shielding is recommended. Proposed methods are also described involving the introduction of electron tubes into the measuring set, which will reduce the measurement to a single setting or deflection. This gives an alternating current method which is comparable with Maxwell's single null-setting cyclical charge and discharge method. Special attention is drawn to Maxwell's remarkable method which is entirely ignored by at least most of the modern text-books and handbooks.

THE object of this paper is to emphasize the importance of direct capacity networks; to explain various methods of measuring direct capacities; and to advocate the use of the Colpitts substitution method which has been found preeminently satisfactory under the wide range of conditions arising in the communication field.

About thirty years ago telephone engineers substituted the so-called " mutual capacity " measurement for the established " grounded capacity " measurement; this was a distinct advance, since the transmission efficiency is more closely connected with mutual capacity than with grounded capacity. Mutual capacity, however, can give no information respecting crosstalk, and accordingly, about twenty years ago, I introduced the measurement of " direct capacity " which enabled us to control crosstalk and to determine more completely how telephone circuits will behave under all possible connections.

For making these direct capacity measurements alternating currents of telephone frequencies were introduced so as to determine more exactly the effective value of the capacity in telephonic transmission, and to include the determination of the associated effective direct conductances which immediately assumed great importance upon the introduction of loading.

Telephone cables and other parts of the telephone plant present the problem of measuring capacities which are quite impossible to isolate, but which must be measured, just as they occur, in association with other capacities; and these associated capacities may be much larger than the particular direct capacity which it is neces-

[1] This article is also appearing in the August issue of the *Journal of the Optical Society of America and Review of Scientific Instruments*. An appendix is added here giving proofs of the mathematical results.

sary to accurately measure, and have admittances overwhelmingly larger than the direct conductance, which is often the most important quantity. This is the interesting problem of direct capacity measurement, and distinguishes it from ordinary capacity measurements where isolation of the capacity is secured, or at least assumed.

The substitution alternating current bridge method, suggested to me in 1902 by Mr. E. H. Colpitts as a modification of the potentiometer method, has been in general use by us ever since in all cases where accuracy and ease of manipulation are essential.

After first defining direct capacities and describing various methods for measuring them, this paper will explain how this may all be generalized so as to include both the capacity and conductance components of direct admittances, and the inductance and resistance components of direct impedances.

Definition of Direct Capacity

It is a familiar fact that two condensers of capacities C_1, C_2, when in parallel or in series, are equivalent to a single capacity $(C_1 + C_2)$ or $C_1 C_2/(C_1 + C_2)$, respectively, directly connecting the two terminals. These equivalent capacities it is proposed to call direct capacities. The rules for determining them may be stated in a form having general applicability, as follows:

Rule 1. The direct capacity which is equivalent to capacities in parallel is equal to their sum.

Rule 2. The direct capacity between two terminals, which is equivalent to two capacities connecting these terminals to a concealed branch-point, is equal to the product of the two capacities divided by the total capacity terminating at the concealed branch-point, *i.e.*, its grounded capacity.

These rules may be used to determine the direct capacities of any network of condensers, with any number of accessible terminals and any number of concealed branch-points. Thus, all concealed branch-points may be initially considered to be accessible, and they are then eliminated one after another by applying these two rules; the final result is independent of the order in which the points are taken; all may, in fact, be eliminated simultaneously by means of determinants[2]; a network of capacities, directly connecting the accessible terminals, without concealed branch-points or capacities in parallel, is the final result. Fig. 1 shows the two elementary cases of direct capacities and also, as an illustration of a more complicated system, the bridge

[2] See appendix, section 1, for a discussion of determinant solutions.

circuit, with three corners 1, 2, 3 assumed to be accessible, and the fourth inaccessible, or concealed. Generalizing, we have the following definition:

The direct capacities of an electrical system with n given accessible terminals are defined as the $n(n-1)/2$ capacities which, connected between each pair of terminals, will be the exact equivalent of the system in its external reaction upon any other electrical system with which it is associated only by conductive connections through the accessible terminals.

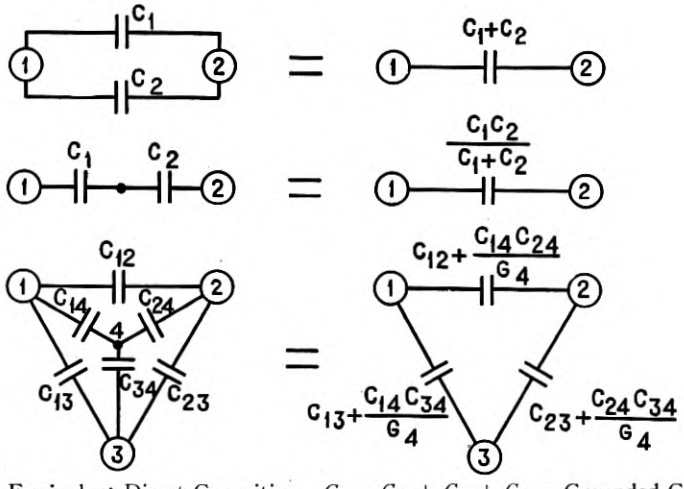


Fig. 1—Equivalent Direct Capacities. $G_4 = C_{14} + C_{24} + C_{34}$ = Grounded Capacity of Branch-Point 4

The total direct capacity between any group of the terminals and all of the remaining accessible terminals, connected together, is called the grounded capacity of the group.

This definition of direct capacity presents the complete set of direct capacities as constituting an exact, symmetrical, realizable physical substitute for the given electrical system for all purposes, including practical applications. Direct capacities are Maxwell's "coefficients of mutual induction," but with the sign reversed, their number being increased so as to include a direct capacity between each pair of terminals.

In considering direct capacities we exclude any direct coupling, either magnetic or electric, from without with the interior of the electrical system, since we have no concern with its internal structure; we are restricted to its accessible, peripheral points or terminals; some care has been taken to emphasize this in the wording of the definition.

Additive Property of Direct Capacities

Connecting a capacity between two terminals adds that capacity to the direct capacity between these terminals, and leaves all other direct capacities unchanged. Connecting the terminals of two distinct electrical systems, in pairs, gives a system in which each direct capacity is the sum of the corresponding two direct capacities in the individual systems. Joining two terminals of a single electrical system to form a single terminal adds together the two direct capacities from the two merged terminals to any third terminal, and leaves all other direct capacities unchanged, with the exception of the direct capacity between the two merged terminals, which becomes a short circuit. Combining the terminals into any number of merged groups leaves the total direct capacity between any pair of groups unchanged, and short-circuits all direct capacities within each group.

These several statements of the additive property of direct capacities show the simple manner in which direct capacities are altered under some of the most important external operations which can be made with an electrical network, and explain, in part, the preeminent convenience of direct capacity networks.

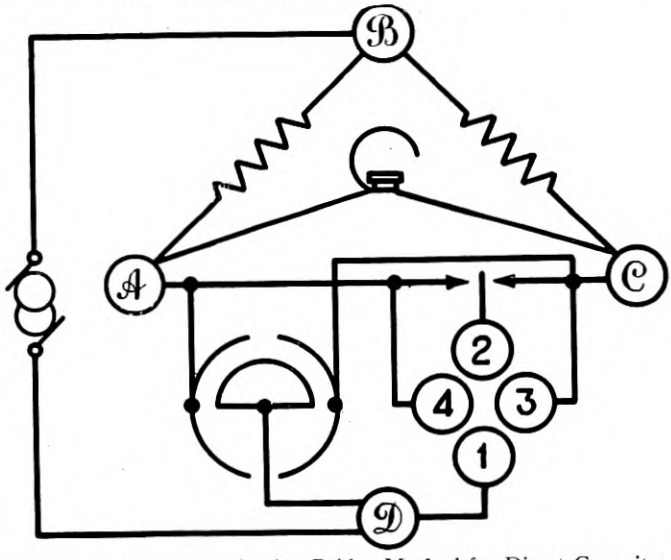


Fig. 2—Colpitts Substitution Bridge Method for Direct Capacity

Since the additive property of direct capacities is sufficient for explaining the different methods of measuring direct capacities we may now, without further general discussion of direct capacities, proceed to the description of the more important methods of measurement.

Colpitts Substitution Bridge Method, Fig. 2

The unknown direct capacity is shifted from one side of the bridge to the other, and the balance is restored by adjusting the capacity standard so as to shift back an equal amount of direct capacity. The method is therefore a substitution method, and the value of the bridge ratio is not involved. Both the standard and the unknown remain in the bridge for both settings, so that the method involves transposition rather than simple, ordinary substitution.

Details of the method as shown by Fig. 2 are as follows: To measure the direct capacity C_{12} between terminals 1 and 2 connect one terminal (1) to corner $\mathcal{D}$ of the bridge, and adjust for a balance with the other terminal (2) on corner $\mathcal{P}$ and then on $\mathcal{C}$, while each and every one of the remaining accessible terminals (3, 4, . . .) of the electrical system is permanently connected during the two adjustments to either corner $\mathcal{P}$ or $\mathcal{C}$. If the direct capacities in the standard condenser between corners $\mathcal{P}$ and $\mathcal{D}$ are C', C'' in the two balances,

$$C_{12} = C'' - C'$$

and if the bridge ratio is unity[3],

$$C_{13} - C_{14} = C' + C'' - 2C^{\circ},$$

where C_o is the standard condenser reading when the bridge alone is balanced.

Two settings are required by this method for an individual direct capacity measurement, but in the systematic measurement of all the direct capacities in a system the total number of settings tends to equal the total number of capacities, when this number becomes large. The number of settings may always be kept equal to the number of capacities by employing an equality bridge ratio, and using the expression for the direct capacity difference given above. The same remarks also hold for the group of direct capacities connecting any one terminal with all the other terminals.

In general, ground is placed upon corner $\mathcal{C}$ of the bridge, but is transferred to corner $\mathcal{D}$, if it is connected to one terminal of the required direct capacity. The arbitrary distribution of the other terminals between corners $\mathcal{P}$ and $\mathcal{C}$ may be used to somewhat control the amount of standard capacity required; or it may be helpful in reducing interference from outside sources, when tests are made upon extended circuits. The grounded capacity of a terminal or group of terminals is measured by connecting the group to $\mathcal{C}$, and all of the remaining terminals together to $\mathcal{D}$.

[3] See appendix, section 2.

The excess of one direct capacity C_{12} over another C_{56} is readily determined by connecting terminals 1 and 5 to corner $\mathcal{D}$, terminals 3, 4, 7, 8, . . . to corner $\mathcal{C}$ or $\mathcal{A}$, and then balance with terminals 2 and 6 on $\mathcal{A}$ and $\mathcal{C}$, respectively, and repeat, with their connections reversed.

Potentiometer Method, Fig. 3

The required direct capacity C_{12} is balanced against one of its associated direct capacities, augmented by a standard direct capacity

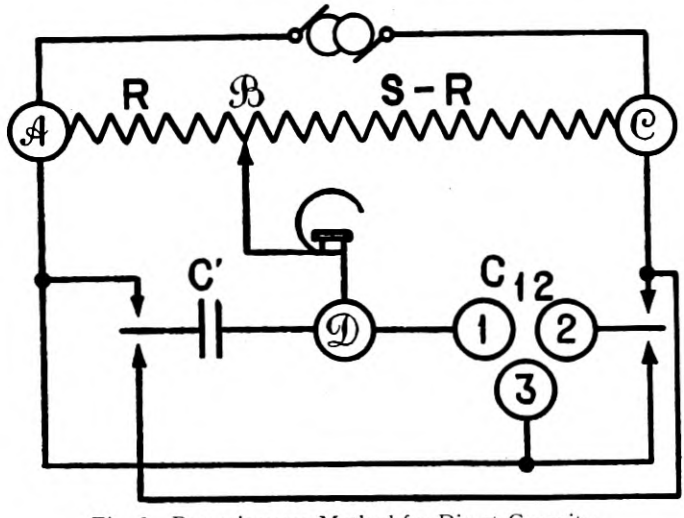


Fig. 3—Potentiometer Method for Direct Capacity

C', and the measurement is repeated with the required direct capacity and standard interchanged. Let R', R'' be the resistances required in arm $\mathcal{A}\mathcal{B}$ of the bridge for the first and second balance, then, S being the total slide wire resistance and G_1 the grounded capacity of terminal 1:[4]

$$C_{12} = \frac{R'}{R''} C',$$

$$G_1 = \frac{S - R''}{R''} C'.$$

This ratio method requires for the bridge a variable or slide wire resistance and a constant condenser, and it may be employed as an improvised bridge, when sufficient variable capacity is not available for the Colpitts method. Not being a substitution method, however,

[4] See appendix, section 3.

greater precautions are necessary for accurate results. There must be no initial direct capacity in arm $\mathcal{CD}$, or a correction will be required. Possibly variable capacity ratio arms would be preferable to resistances.

Null-Impedance Bridge Method for Direct Capacity, Fig. 4

Assuming that the electron tube supplies the means of obtaining an invariable true negative resistance, Fig. 4 shows a method which determines any individual direct capacity from a single bridge setting. The bridge arms are replaced by a Y network made up of two resist-

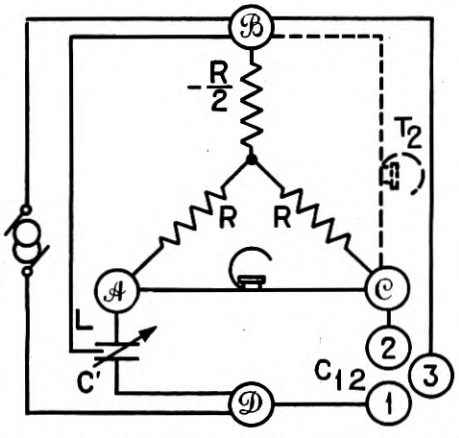


Fig. 4—Null-Impedance Bridge Method for Direct Capacity

ances R, R and a negative resistance $-R/2$; the Y has then a null-impedance between corner $\mathcal{B}$ and corners $\mathcal{A}$, $\mathcal{C}$ connected together[5]. The three terminals 1, 2, 3 of the network to be measured are connected to corners $\mathcal{D}$, $\mathcal{C}$, $\mathcal{B}$ and a balance obtained by adjusting the variable standard condenser C'. Then $C_{12} = C'$ regardless of the direct capacities associated with C_{12} and C', since these capacities either are short-circuited between corners $\mathcal{B}$, $\mathcal{A}$ or $\mathcal{B}$, $\mathcal{C}$ or are between corners $\mathcal{B}$, $\mathcal{D}$ and thus outside of the bridge.

Correct adjustment of the negative resistance may be checked by observing whether there is silence in telephone T_2 after the balance has been obtained. Assuming invariable negative resistance, this test need be made only when the bridge is set up, or there is a change in frequency. The bridge may be given any ratio Z_1/Z_2 by employing a Y made up of impedances Z_1, Z_2, and $-Z_1 Z_2/(Z_1 + Z_2)$.

[5] See appendix, section 4, which also describes a transformer substitute for the Y.

Maxwell Discharge Method[6], Fig. 5

Connect the terminals between which the direct capacity C_{12} is required, to A, B and the remaining accessible terminals of this electrical system to D. The adjustable standard capacity is C' and any associated direct capacities in this standard are shown as C'',

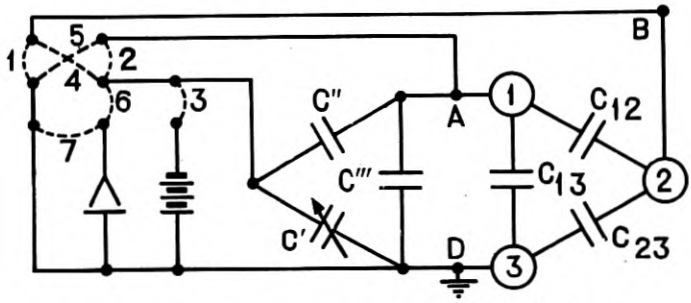


Fig. 5—Maxwell Discharge Method for Direct Capacity

C'''. If C_{12} is a direct capacity to ground, interchange C' and C_{12}. Balancing involves the following repeated cycle of operations:

1. Make connections 1, 2, 3 and 7 for an instant (thus charging C_{12}, C_{13}, C''', C' and discharging the electrometer).
2. Make connections 4, 5 and then 6 (to discharge condensers C_{13}, C''', mix charges of C_{12}, C' with polarities opposed and connect electrometer).
3. Adjust C' to reduce the electrometer deflection when the cycle is again repeated.

When a null deflection is obtained $C_{12} = C'$; the required direct capacity is equal to the standard direct capacity irrespective of the magnitudes of the four associated direct capacities. If all capacities are free of leakage and absorption, this remarkable method accurately compares two direct capacities by means of a single null setting, and it requires the irreducible minimum amount of apparatus.

Balanced-Terminal Capacity Measurement, Fig. 6

This is defined as the direct capacity between two given terminals with all other terminals left floating and ignored, after a hypothetical redistribution of the total direct capacity from the given pair of terminals to every third terminal which balances the two sides of the pair. The balanced-terminal capacity, as thus defined, is equal to the direct capacity between the pair augmented by one-quarter of

[6] Electricity and Magnetism, v. 1, p. 350 (ed. 1892).

the grounded capacity of the pair, neither of which is changed by the assumed method of balancing.

As illustrated in Fig. 6, terminals 1, 2 are the given pair and terminal 3 includes all others, assumed to be connected together. A bridge ratio of unity is employed, and the entire bridge is shielded from ground with the exception of corners $\mathcal{C}$, $\mathcal{D}$ which are initially balanced

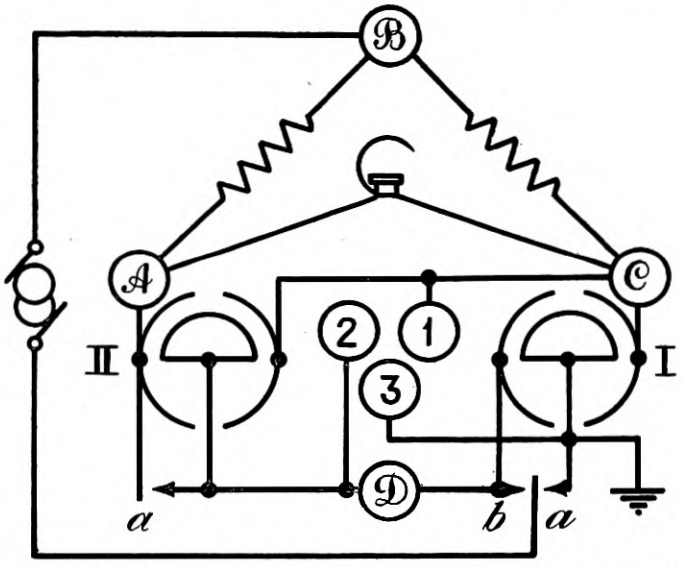


Fig. 6—Bridge for Determining Hypothetical Capacity Between Two Terminals with Other Terminals Balanced and Ignored

to ground within the range of variable condenser I. The following two successive balances are made:

1. With contacts a, a' closed and b open, balance is secured by varying condenser I (the total capacity of which is constant) giving the reading C' for its direct capacity in parallel with terminals 1, 3.
2. With contacts a, a' open and b closed, balance is obtained by varying condenser II, obtaining the reading C'' for its direct capacity in $\mathcal{CD}$.

If C_o', C_o'' are the corresponding readings without the network, the balanced-terminal capacity C_b and the grounded capacity unbalance of the given pair of terminals are:[7]

$$C_b = 2\,(C'' - C_o''),$$
$$G_2 - G_1 = 2\,(C' - C_o').$$

[7] See appendix, section 5.

Any failure to adjust condenser I to perfectly balance the given pair of terminals will decrease the measured capacity C_b. This fact may be utilized to measure the capacity with the second bridge arrangement alone (contacts a, a' open and b closed) by adjusting condenser I so as to make the reading C'' of condenser II a maximum. This procedure presents no difficulty, since the correct setting for condenser I lies midway between its two possible settings for a balance with any given setting of condenser II; furthermore, C'' is not sensitive to small deviations from a true balance in C'.

Balanced-terminal capacity is of practical importance as a measure of the transmission efficiency to be expected from a metallic circuit, if it is subsequently transposed so as to balance it to every other conductor. In practice, when the unbalance of the section of open wire or cable pair, which is being measured, is relatively small, it is sufficient to set condenser I, once for all, to balance the bridge itself and ignore the unbalance of the pair. This favors an unbalanced pair, however, by the amount $(G_2 - G_1)^2/4\ (G_{12} + G_{CD})$ where $G_{12} + G_{CD}$ is the grounded capacity of the pair augmented by that of the bridge.[8] For rapid working, condenser II is graduated to read $2C''$ and by auxiliary adjustment C_o'' is made zero, so that the required capacity is read directly from the balance.

Additional Methods of Measuring Direct Capacity

Measurement of the capacity between the terminals, taken in pairs with all the remaining terminals left insulated or floating, gives $n(n-1)/2$ independent results, from which all the direct capacities may be derived by calculation of certain determinants[9]. Practically, however, we are in general interested in determining individual direct capacities from the smallest possible number of measurements, and the first step is naturally to connect all of the remaining conductors together, so as to reduce the system to two direct capacities in addition to the one the value of which is required. Three measurements are then the maximum number required, and we know that two, or even one, is sufficient if particular devices are employed.

The three measurement method of determining direct capacities from the grounded capacities of the two terminals taken separately G_1, G_2, and together G_{12}, is given by Maxwell.[10] If $G_1 = C'$, $G_{12} = C' + C''$, and $G_2 = C'' + C'''$,

[8] See appendix, section 6.

[9] See appendix, section 7.

[10] *Ibid.*, p. 110.

then

$$C_{12} = \tfrac{1}{2}\,(G_1 + G_2 - G_{12})$$
$$= \tfrac{1}{2}\,C''',$$

which indicates a method by which large grounded capacities can be balanced against three variable capacities, only one of which need be calibrated, and that one need be no larger than the required direct capacity.

Two-setting methods, as illustrated by the Colpitts and potentiometer methods, rest upon the possibility of connecting one of the associated direct capacities between opposite corners of the bridge where it is without influence on the balance, and not altering any associated direct capacity introduced into the working arms of the bridge. Numerous variations of these methods have been considered which may present advantages under special circumstances. Thus, if conductors 1, 2, 3 of Fig. 7 are in commercial operation, and it is not permissible to directly connect two of them together, a double bridge might be employed with a testing frequency differing from

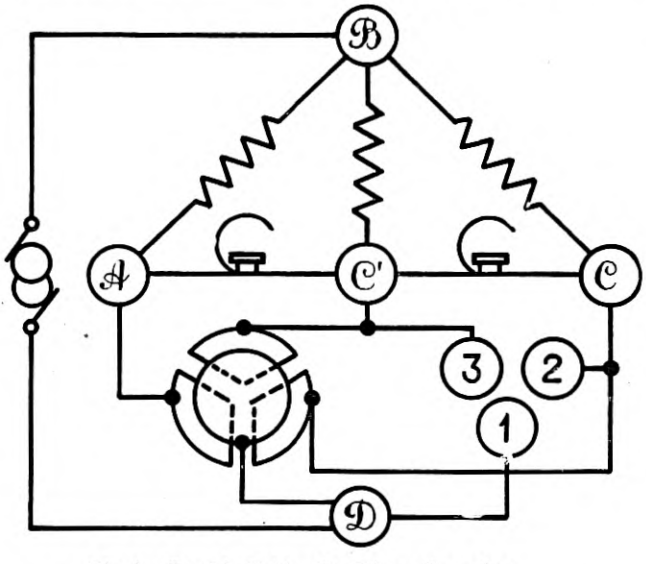


Fig. 7—Double Bridge for Direct Capacity

that of operation. A telephone is shown for each ear, and a constant total direct capacity is divided between the three branches in the proportion required to silence both telephones.

One-setting methods attained ideal simplicity in the Maxwell discharge method, but we found it necessary to use alternating current

methods, and here negative resistances make a one-setting method at least theoretically possible, as explained above. Of possible variations it will be sufficient to refer to the ammeter method Fig. 8. Termi-

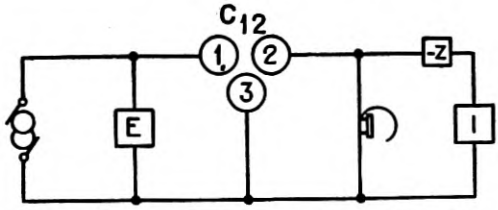


Fig. 8—Ammeter Circuit for Determining Direct Capacity

nals 1 and 2 of the required direct capacity C_{12} are connected to the voltmeter and ammeter terminals, respectively, and all other terminals go to the junction point at 3. Then

$$C_{12} = \frac{I}{2\pi f E},$$

provided the ammeter actually has negligible impedance. The method is well adapted for rapid commercial testing. The ammeter impedance may be reduced to zero by a variable negative impedance device ($-Z$), adjusted to reduce the shunted telephone to silence.

Shielding

In the discussion of the bridge, it has been assumed that the several pieces of apparatus forming the six branches of the bridge have no mutual electrical or magnetic reaction upon each other, except as indicated. In general, however, a balance will be upset by changes in position of the pieces of apparatus, or even by movements of the observer himself, whereas these motions cannot affect any of the mutual reactions which have been explicitly considered. The skillful experimenter, understanding how these variations are produced by the extended electric and magnetic fields, will anticipate this trouble and take the necessary precautions, possibly without slowing down his rate of progress.

Where hundreds of thousands of measurements are to be made, however, substantial savings are effected by arranging the bridge so that reliable measurements can be made by unskilled observers, and here it is necessary to shield the bridge so that any possible movements of the observer and of the apparatus will not affect the results. Magnetic fields of transformers are minimized by using toroidal coils with iron cases. Electrostatic fields are shielded by copper cases;

the principles of shielding were explained in an earlier paper,[11] Fig. 13 of that paper showing the complete shielding of the balance as constructed for the measurement of direct capacity by the Colpitts method. Over five million capacity and conductance measurements have been made with the shielded capacity and conductance bridge and in a forthcoming paper Mr. G. A. Anderegg will give details of actual construction of apparatus and of methods of operation as well as some actual representative results.

Direct Admittance Measurements

For simplicity, the preceding definitions and methods of measurement have been described in terms of capacity, but everything may be generalized, with minor changes only, for the definition and measurement of direct admittances with their capacity and conductance components. The essential apparatus change is the addition, in parallel with the variable capacity standards employed, of a variable conductance standard, which shifts direct conductance from one side of the bridge to the other, without changing the total reactance and conductance in the two sides of the bridge. This may be practically realized in a great variety of ways as regards details, which it will suffice to illustrate by Fig. 9, where C', C'', C''', G', G'', indicate the

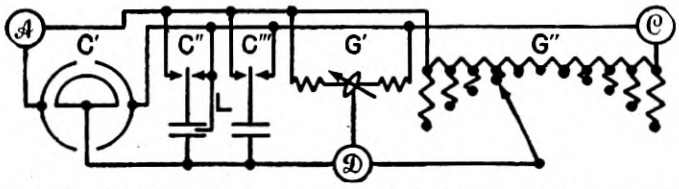


Fig. 9—Variable Direct Conductance and Capacity Standard for Direct Admittance Bridge

continuously variable capacity and conductance standards with enough step-by-step extensions to secure any desired range.

For the continuously variable conductance standard a slide wire is represented, with a slider made up of two hyperbolic arcs so proportioned that, as the slider is moved uniformly in a given oblique direction, conductance is added uniformly on the left and just enough of the wire is short-circuited to produce an equal conductance decrease on the other side. The arcs are portions of the hyperbola $xy = (L^2 - S^2)/4$, where L, S are the total length of the wire and of the portion to be traversed by the slider, and the coordinate axes are

[11] The Shielded Balance, El. W., 43, 1904 (647–649).

the slide wire and the direction of the motion of the slider as oblique asymptotic axes.[12] $L = GS/g = 4\,G/\rho(G^2 - g^2)$, where G is the total conductance and $(G \pm g)/2$ the limiting direct conductance on either side.

If an ordinary slider replaces the hyperbolic arc slider, and the scale reading is made non-uniform so as to give one-half of the difference between the direct conductances $\mathcal{A}$ to $\mathcal{D}$ and $\mathcal{C}$ to $\mathcal{D}$, the conductance standard will still give absolutely correct results with the Colpitts method, provided the bridge ratio is unity. This simplification in connection with the balancing capacity I of Fig. 6 would, however, not be strictly allowable. For improvised testing we have found it sufficient to use two equal resistances (R) with a dial resistance (r) in series with one of them, and take the defect of conductance introduced by the dial resistance as equal to r/R^2 or to $10^{-2}r$, $10^{-1}r$, r, micromho according as R was made 10000, 3162, or 1000 ohms.[13]

For a step-by-step conductance standard, Fig. 9 shows a set of 10 equal resistances, connected in series between corners $\mathcal{A}$, $\mathcal{C}$, to the junction points of which there is connected a parabolic fringe of resistances, the largest of which is 2.5 times each of the ten resistances. With this arrangement the direct conductance in $\mathcal{A}\mathcal{D}$ may be adjusted by ten equal steps, beginning with zero, while the conductance in $\mathcal{C}\mathcal{D}$ is decreased by equal amounts to zero. The total resistance required for this conductance standard is only 21/25 of the resistance required to make a single isolated conductance equal to one of the ten conductance steps; the ratio may be reduced to 1/2 by doubling the number of contacts,[14] and using one fringe resistance for all positions. Resistance may be still further economized by using as high a total conductance as is permissible in the bridge, and securing the required shift in conductance from a small central portion of the parabolic fringe.

Fig. 9 shows the variable capacity standards as well as the variable conductance standards and a few practical points connected with the capacity standards may be mentioned here.

The revolving air condenser standard has two fixed plates connected to $\mathcal{A}$ and $\mathcal{C}$, so that the capacity will increase as rapidly on one side as it decreases on the other side. Since perfect constancy of the total capacity is not to be expected, on account of lack of perfect mechanical uniformity, the revolving condenser should be calibrated to read

[12] See appendix, section 8.

[13] See appendix, section 9.

[14] See appendix, section 10.

one-half of the difference between the capacities on the two sides, as explained above in connection with conductance. The capacity sections employed to extend the range of the revolving condenser include both air condensers C' and mica condensers C'', the latter being calibrated by means of the air condensers and the conductance standard.

A novel feature of our standard air condensers is a third terminal called the leakage terminal, and indicated at L in Figs. 4, 9. Attached to it are plates so arranged that all leakages either over, or through, the dielectric supports from either of the two main terminals, must pass to the leakage terminal. There can be no leakage directly from one of the main terminals to the other. There is thus no phase angle defect in the standard direct capacity due to leakage, and that due to dielectric hysteresis in the insulating material is reduced to a negligible amount by extending the leakage plates beyond the dielectric, so as to intercept practically all lines of induction passing through any support. This leakage terminal is connected to corner C of the bridge; in the revolving condensers, it is one of the fixed plates.

Direct Impedance Measurements

The reciprocal of a direct admittance is naturally termed a direct impedance; substituting impedance for capacity, the definition of direct capacity, given above, becomes the definition of direct impedance. The complete set of direct impedances constitutes an exact, symmetrical, physical substitute for any given electrical system. Direct impedances are often, in whole or in part, the most convenient constants since many electrical networks are made up of, or approximate to, directly connected resistances and inductances. To make direct impedance measurements which will not involve the calculation of reciprocals, we naturally employ inductance and resistance standards in series, the associated direct impedances being eliminated as with direct capacities.

Conclusion

It has been necessary to preface the description of methods of measuring direct capacities by definitions and a brief discussion, since direct capacities receive but scant attention in text-books and handbooks. By presenting direct capacities, direct admittances, and direct impedances as alternative methods of stating the constants of the same direct network, employed as an equivalent substitute for any given electrical system, it is believed the discussion and measure-

ment of networks has been simplified. In another paper the terminology for admittances and impedances will be still further considered, together with their analytical correlation.

APPENDIX

In explaining the different methods of measuring direct capacities it is necessary to start with a clear idea of what direct capacities are, and to make use of the additive property, but it is not necessary to go into any comprehensive discussion of direct capacities. Accordingly, the mathematical treatment of direct capacities has been reserved for another paper, but it seems desirable to append to the present paper proofs of the analytical results given in this paper, since the method of approach giving the simplest proof is not always perfectly obvious.

(1) Reducing the number of terminals which are considered accessible, by ignoring terminals $p, q, r, \ldots$, changes the direct and grounded capacities from (C_{ij}, G_i) to (C'_{ij}, G'_i), the latter being expressed in terms of the former as follows:

$$C'_{ij} = \frac{-\begin{vmatrix} -C_{ij} & -C_{ip} & -C_{iq} & \cdots \\ -C_{jp} & G_p & -C_{pq} & \cdots \\ -C_{jq} & -C_{pq} & G_q & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix}}{\begin{vmatrix} G_p & -C_{pq} & \cdots \\ -C_{pq} & G_q & \cdots \\ \cdots & \cdots & \cdots \end{vmatrix}}$$

$$G'_i = -C'_{ii}$$

where C'_{ii} is given by formula above and $G_i = -C_{ii}$.

To check these formulas note that on substituting $(G_i, -C_{ij})$ for Maxwell's (q_{ii}, q_{ij}) in his equations (18)[15] the coefficients form an array in which the grounded capacity G_i is the ith element in the main diagonal and $-C_{ij}$ is the element at the intersection of row i, column j. The array may be supposed to include every terminal symmetrically by considering the earth's potential as being unknown and writing down the redundant equation for the charge on the earth. Let the charge be zero on terminal j and on all concealed terminals; let there be a charge on terminal i and an equal and opposite charge on all the remaining accessible terminals, connected together to form a single terminal k. Now taking the potential of j as the zero of reference

[15] *Ibid.*, p. 108.

and calculating the potentials of i and k and then allowing the direct capacity between j and k to become infinite, the direct capacity between terminals i and j is $C_{ij} = -Lim\ (C_{jk}\ V_k\ /\ V_i)$. This gives the above formula for C'_{ij}, with $-C'_{ii}$ as a special case. This method is an electrostatic counterpart of the ammeter method shown in Fig. 8 on page 29.

If there is but one ignored terminal the determinant solution takes on a simple form from which Rules 1 and 2 and Fig. 1 may be checked.

If all but two terminals are ignored the equivalent direct network is reduced to a single direct capacity. When, for each pair of terminals, this capacity C'_{ij} is known, from measurements or from calculations, the direct capacities between the terminals may be derived by means of the following formulas

$$C_{ij} = 2\frac{D_{ij}}{D}$$

$$G_i = -C_{ii} = -2\frac{D_{ii}}{D}$$

where D_{ij} is the cofactor of the element in row i column j of the determinant

$$\mathrm{D} = \begin{vmatrix} 0 & S_{12} & S_{13} & \cdot\cdot & 1 \\ S_{12} & 0 & S_{23} & \cdot\cdot & 1 \\ S_{13} & S_{23} & 0 & \cdot\cdot & 1 \\ \cdot & \cdot & \cdot & \cdot\cdot & \\ 1 & 1 & 1 & \cdot\cdot & 0 \end{vmatrix}$$

which has zeros in the main diagonal, a border of ones in the last row and column, while the other elements are $S_{ij} = 1/C'_{ij}$, that is, the reciprocals of the given capacities. The S's form a complete symmetrical system of network constants; Maxwell's coefficients of potential p_{ii} with the two suffixes the same are the same quantities, but he employs only those coefficients of this type which are associated with the earth, his system being completed by adding the coefficients with different suffixes. By starting with Maxwell's results the above formula may be deduced, but more direct proofs, both physical and mathematical, will be given in the theoretical paper referred to at the end of the present paper.

The purpose of this section of the appendix is achieved if the determinant solutions are made so clear as to be available for use in any particular case.

(2) Starting with the bridge alone balanced at reading C° the other two settings involve, in the capacity standard, increases in the direct

capacity on the left of $(C' - C^\circ)$ and $(C'' - C^\circ)$, with equal decreases on the right. Therefore

$$C_{12} + C_{14} + (C' - C^\circ) = -(C' - C^\circ) + C_{13}$$
$$C_{14} + (C'' - C^\circ) = -(C'' - C^\circ) + C_{13} + C_{12}$$

and adding gives the value of $(C_{13} - C_{14})$.

(3) The condition of equal impedance ratios on the two sides, as required for a balance, gives, for both the switches up and down,

$$R'(G_1 - C_{12} + C') = (S - R')\,C_{12},$$
$$R''\,G_1 = (S - R'')\,C',$$

respectively, from which the expressions for C_{12} and G_1 follow.

(4) The Y of Fig. 4 has unusual properties because the total conductance connecting the concealed branch-point of the Y to the three bridge corners $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ is zero. Thus the conductance between any one corner and the remaining two corners joined together is infinite, or in other words, the Y acts as a short circuit under all these three conditions. On the other hand, if corner $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}$ is left floating and ignored the conductance between the other two corners is $2/R$, $1/2R$ or $2/R$, respectively, and the Y is not a short circuit. These statements are verified at once by applying the familiar expressions for resistances in parallel and in series.

On account of the unusual behavior of the Y, even when taken alone, it is not immediately apparent how it will affect the operation of the bridge of Fig. 4 with direct capacities between corners $\mathcal{AB}$ and $\mathcal{BC}$. For this reason it is highly desirable to find an equivalent network the behavior of which is more readily comprehended. It is not feasible to employ the delta network which is equivalent to the Y for this has indeterminate characteristics, being made up of three infinite conductances, only two of which have the same sign. We may, however, make use of the Y which is equivalent to the original Y and direct capacities C_{AB} and C_{BC} taken together. This is found as follows: Any admittance delta may be replaced by a star having admittances equal to the sum of the products of the delta admittances taken in pairs divided by the opposite delta admittance. Applied to the delta of Fig. 1, we find that the star that is equivalent has the capacities

$$C''_{14} = \frac{S}{C_{24}\,C_{34} + G_4\,C_{23}}$$
$$C''_{24} = \frac{S}{C_{34}\,C_{14} + G_4\,C_{13}}$$
$$C''_{34} = \frac{S}{C_{14}\,C_{24} + G_4\,C_{12}}$$

where

$$S = C_{14}C_{24}C_{34} + C_{14}C_{24}(C_{13}+C_{23}) + C_{24}C_{34}(C_{12}+C_{13}) + C_{34}C_{14}(C_{12}+C_{23}) + G_4(C_{12}C_{23} + C_{23}C_{13} + C_{12}C_{13}),$$

which, upon substituting the value of G_4, is the sum of 16 terms, each of which is the product of three capacities, every combination of three capacities being included except the four cases in which the three capacities would form a closed circuit. By allowing the capacities to be complex quantities, any admittances are covered by the formulas.

If $G_4 = 0$

$$\frac{C''_{14}}{C_{14}} = \frac{C''_{24}}{C_{24}} = \frac{C''_{34}}{C_{34}} = \frac{S}{C_{14}C_{24}C_{34}}$$

or the new Y arms present the same ratios as the original Y arms taken alone; that is, the direct capacities C_{AB}, C_{BC} of Fig. 4 have no effect on the bridge ratio. Thus the constancy of the bridge ratio holds for all null-impedance bridges regardless of the ratio Z_1/Z_2 and of the nature of the direct admittances from corners A and C to B.

If $G_4 = 0$ and also $C_{24} = C_{14}$ and $C_{12} = 0$, then

$$C''_{14} = C''_{24} = -\tfrac{1}{2}\,C''_{34} = C_{14} + \tfrac{1}{2}\,(C_{13} + C_{23}).$$

Applying this to Fig. 4, which is possible since the bridge ratio is unity, we find that the three arms of the equivalent Y may be considered as being made up of resistances and capacities in parallel. The resistances are R, R, $-R/2$ and the associated capacities C, C, $-2C$, where R is the original resistance in the Y and C is one-half the sum of the two actual direct capacities from $\mathcal{B}$ to $\mathcal{A}$ and from $\mathcal{B}$ to $\mathcal{C}$. The equivalent bridge thus obtained has ratio arms made up of ordinary resistances and capacities and therefore Fig. 4 used as a bridge can present no unexpected characteristics; the negative resistances and capacities of the equivalent Y merely affect the current supplied to the bridge.

An ideal transformer, if such a device existed, might replace the Y, for it would maintain a constant ratio between the currents in the two windings and act as a short circuit when the bridge is balanced. To determine the error when an actual transformer with impedances Z_p, Z_s, Z_{ps} is employed, take the general expression for the ratio of the capacities derived above which is

$$\frac{C''_{14}}{C''_{24}} = \frac{C_{34}\,C_{14} + G_4\,C_{13}}{C_{24}\,C_{34} + G_4\,C_{23}}.$$

Change to admittances by substituting Y for C and G throughout. Assume the transformer replaced by its equivalent conductance star so that

$$Y_{14} = \frac{1}{Z_p + Z_{ps}},$$

$$Y_{24} = \frac{1}{Z_s + Z_{ps}},$$

$$Y_{34} = -\frac{1}{Z_{ps}}, \text{ and by addition}$$

$$Y_4 = \frac{Z_{ps}^2 - Z_p Z_s}{(Z_p + Z_{ps})(Z_s + Z_{ps}) Z_{ps}}.$$

Substituting these values the expression for the actual ratio of the bridge arms becomes

$$\frac{Y_{14}''}{Y_{24}''} = \frac{Z_s + Z_{ps} + (Z_p Z_s - Z_{ps}^2) Y_{13}}{Z_p + Z_{ps} + (Z_p Z_s - Z_{ps}^2) Y_{23}}.$$

(5) When the bridge alone is balanced at readings C_o' and C_o'', let C_{CD} and G_{CD} be the direct capacity between corners $\mathcal{C}$ and $\mathcal{D}$ and the total direct capacity between these corners and ground. Since G_{CD} is balanced, the effective direct capacity between corners $\mathcal{C}$, $\mathcal{D}$ when earth is ignored, is by Fig. 1, $(C_{CD} + G_{CD}/4)$. Now connect the three terminals 1, 2, 3, as shown with direct capacities C_{12}, $G_1 - C_{12}$, $G_2 - C_{12}$; G_1, G_2 being the grounded capacities of terminals 1 and 2. The first balance with the reading C' requires the equality of the total capacity added on each side, *i.e.*,

$$G_1 - C_{12} + (C' - C_o') = G_2 - C_{12} - (C' - C_o')$$

or

$$G_2 - G_1 = 2(C' - C_o')$$

For the second balance ground may again be considered an ignored terminal, and since terminals 1 and 2 have been balanced to ground, and their total direct capacity to ground is $G_{12} = G_1 + G_2 - 2C_{12}$, the effective direct capacity added to the bridge between corners $\mathcal{C}$ and $\mathcal{D}$ is $C_b = C_{12} + G_{12}/4$. Equating the added capacities on the two sides of the bridge when balanced at the reading C'', we obtain $C_b = 2(C'' - C_o')$.

The direct capacity between $\mathcal{C}$ and $\mathcal{D}$, when ground is considered an accessible terminal, is assumed to be absolutely independent of the setting of the condenser I. To actually meet this condition will require some attention in the design of the variable condenser.

(6) Here the bridge itself is supposed to have equal direct capacities from corners $\mathcal{C}$ and $\mathcal{D}$ to ground, while the added terminals 1 and 2 have different direct capacities to ground, the difference being $(G_1 - G_2)$, while the total direct capacity to ground is $(G_{12} + G_{CD})$. Now two capacities in series may be replaced by their product divided

by their sum, which is equal to one-fourth of the sum minus the square of the difference divided by four times the sum. The correction due to the difference is thus $(G_2 - G_1)^2/4(G_{12} + G_{CD})$, as stated.

(7) These determinants are given at the end of the first paragraph of this appendix. These expressions for the direct capacity are of more special interest in the analytical discussion of networks.

(8) Assume that a wire resistance is to be employed and that a sliding contact is to intercept such an amount of resistance that the equivalent conductance will vary directly with the motion of the slider carrying the contact point. Then if the wire is straight and the intercepted portion is of length x and the slider motion is rectilinear and its extent is y the relation which holds between them is xy = constant, the value of the constant depending upon the units employed.

In the paper it is assumed that the total conductance G, the total shifted conductance g, and the resistance of unit length of the slide wire ρ are given; the total length of wire L and the portion traversed by the slider S are then calculated. The arc employed, for each half of the slider of Fig. 9, extends equally both ways from the vertex to the points where the values of x and y are $(L \pm S)/2$, on the hyperbola $xy = (L^2 - S^2)/4$. Substituting for L and S the values given in the paper, it will be found that this range of x actually gives the range of conductance $(G \pm g)/2$, as required.

(9) The exact defect in conductance is

$$\frac{1}{R} - \frac{1}{R+r} = \frac{r}{R(R+r)} = \frac{r}{R^2}\left(1 - \frac{r}{R} + \cdots\right)$$

(10) At mid-point the total conductance due to the five resistances (R) on each side, taken in parallel, is $2/5R$ and to give this same conductance an end fringe must have the resistance $2.5R$. Assume a parabolic fringe having the resistance $(5 - n)^2 R/10$ at the point connected to $\mathscr{P}$ and $\mathscr{C}$ by resistances nR and $(10 - n)R$. This gives a Y network and by Fig. 1 the equivalent direct conductances are $(10 - n)/25R$, $n/25R$, $(5 - n)^2/250R$ between $\mathscr{DP}$, $\mathscr{DC}$, $\mathscr{PC}$ respectively. The sum of the first two is constant and the first decreases by equal steps, of $1/25R$ each, to zero as the second increases. The parabolic fringe, therefore, gives the required conductances.

The total resistance in the chain of ten resistances is $10R$, in the fringe $11R$, and in the largest single fringe $2.5R$. With the complete fringe the total required is $(10 + 11)\ R = 21R$; with a single fringe, subdivided as required, only $(10 + 2.5)\ R = 12.5R$ is required. Compared with $25R$, which would be required for one of the conductance steps, these resistances are $21/25$ and $1/2$.

The Relation of the Petersen System of Grounding Power Networks to Inductive Effects in Neighboring Communication Circuits

By H. M. TRUEBLOOD

THE purpose of this paper is to present a simple theoretical treatment of those features of the Petersen method of grounding a power network which are of principal interest from the standpoint of inductive effects in neighboring communication circuits. In this method, the neutral of the system is grounded through an inductance which is in resonance, at the fundamental frequency, with the total direct capacity of the system to ground. The theory of the behavior of a power system thus grounded at times of accidental faults to earth has been developed by Petersen in a paper published in 1919,[1] in which the results of field tests and of operating experience with an installation in Germany are also described. The method has also found application in other places in Europe, chiefly in Italy and Switzerland. It does not appear in any of these cases that inductive interference was a factor requiring, or at any rate receiving, consideration. In fact, it does not seem that either Petersen himself, or other engineers in Europe who have made use of his scheme, have considered it except as a method of protecting power systems from the effects of accidental grounds.

The features of the method that are of interest from the viewpoint of inductive interference relate both to normal operating conditions of the power system and to the phenomena which occur when a phase of the system is grounded. With regard to the former, it is principally, though not entirely, the effect of the neutral reactor on the harmonics of frequencies within the voice range that require examination; with respect to the latter, the things of chief, though not exclusive, importance, are the ground currents and unbalanced voltages to ground of fundamental frequency, which are possible sources of disturbance in exposed communication circuits.

These features, particularly those concerned with effects at fundamental frequency, are more or less closely related to questions of primary importance from the standpoint of power system operation. It is impossible that this should not be the case. Accidental disturbances of a character which may interrupt service or endanger apparatus or equipment in a power system may produce inductive

[1] W. Petersen, Elektrotechnische Zeitschrit, 40, pp. 5–7 and 17–19, 1919; Sci. Abs. B, Nov. 29, 1919.

disturbances in neighboring communication circuits, and the question of grounding the neutral, in whatever manner, or of leaving it isolated, exists because of its bearing on the avoidance or limitation of such power system disturbances. Thus a method of grounding the neutral, or any other method of power system operation designed to limit the extent or the severity of accidental disturbances, must necessarily possess importance with respect to inductive effects in exposed communication circuits.

It does not seem necessary, therefore, to apologize for the discussion, in the first section of the paper, of the behavior of the power system at times of faults to earth. In this section, an explanation is given of the principal characteristic effect of the reactor which differs in some respects from that set forth by Petersen in the paper already referred to. The matter of transient over-voltage on a non-grounded phase is also examined in this section, and the bearing of these and the earlier considerations on inductive effects is discussed.

In the second section of the paper, the behavior of the power system with reactor under normal operating conditions is discussed with reference to noise and other inductive effects in neighboring communication circuits.

1. Effects with a Grounded Phase on the Power System

1. Action of Coil in Suppressing Arcs to Ground

Referring to Fig. 1, in which, and in the following discussion it is assumed that the three admittances to ground are equal, the admittance current through the fault from the two sound phases is

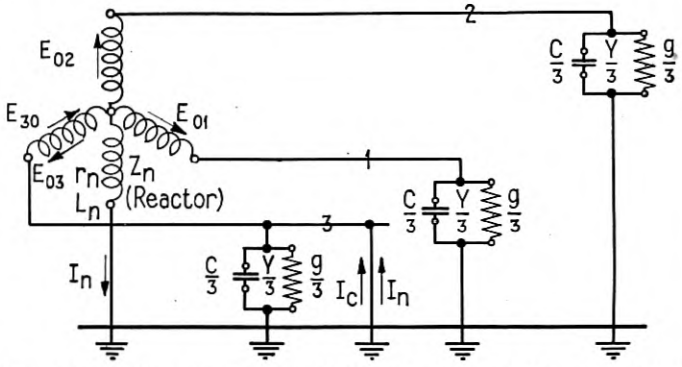


Fig. 1—Single-Phase System with Neutral Grounded Through Reactor. Admittances to Ground Assumed Balanced

$$I_c = \frac{Y}{3}(E_{01} - E_{03}) + \frac{Y}{3}(E_{02} - E_{03}) = \frac{Y}{3}(E_{01} + E_{02} - 2E_{02}),$$
$$= Y E_{30} = (g + jC\omega) E_{30}. \quad (1)$$

The current through the coil and fault is

$$I_n = \frac{E_{30}}{Z_n} = \frac{E_{30}}{r_n^2 + \omega^2 L_n^2}(r_n - j\omega L_n),$$

or, neglecting $\frac{r_n^2}{\omega^2 L_n^2}$ in comparison with unity,

$$I_n = \frac{E_{30}}{\omega^2 L_n^2}(r_n - j\omega L_n).$$

Thus the total fault current is

$$I_c + I_n = I_f = E_{30}\left[g + \frac{r_n}{\omega^2 L_n^2} + j\left(\omega C - \frac{1}{\omega L_n}\right)\right],$$

and, if the coil is adjusted for resonance,

$$I_f = E_{30}\left(g + \frac{r_n}{\omega^2 L_n^2}\right),$$
$$= \frac{E_{30}}{\omega L_n}\left(\frac{r_n}{\omega L_n} + \frac{g}{\omega C}\right). \quad (2)$$

On comparison of this expression with the above equation (1) for the charging current, which constitutes the fault current if the system is isolated, it is seen that, if the losses in the system are small, the effect of the coil is to reduce the magnitude of the current in the fault approximately in the ratio of $\left(\frac{r_n}{\omega^2 L_n^2} + g\right)$ to $|Y|$, *i.e.*, neglecting terms of the second and higher orders, in the ratio $\left(\frac{r_n}{\omega L_n} + \frac{g}{\omega C}\right)$ to unity. Further, as equation (2) shows, the phase of the fault current coincides with that of the voltage impressed between the faulty wire and ground to the degree of approximation here used, *i.e.*, to the second order of small quantities. (The bracket on the right-hand side of (2), if written in full, would include quadrature terms in the square and higher even powers of $r_n/\omega L_n$.) With the system isolated, the phase displacement is nearly 90°.

The action of the coil may be described as a transfer of the charging current from the fault to the coil, leaving nothing but the component of current to supply losses at the fault. With suitable design of the coil, this energy current can be made small.

The coincidence in phase of the fault current and the voltage impressed by the transformer on the faulty wire, together with the small magnitude of the former, are very favorable to the suppression of the

arc. Following its extinction, there is a further action of the resonant system consisting of the coil and the total capacity to ground which acts in such a way as to prevent any over-voltage and to restore the normal potentials to ground *gradually*, thus tending to prevent the arc from restriking. This action is as follows:

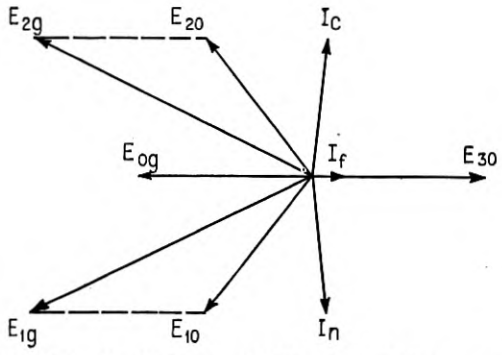


Fig. 2—Vector Diagram Showing Relations of Voltages and Currents at and Following Extinction of Arc

Referring to Fig. 2, the vectors E_{10}, E_{20}, E_{30} represent the emfs. impressed between lines and neutral by the transformer bank. I_c, I_n, I_f are respectively the admittance current to the fault, the current supplied by the reactor to the fault, and the total fault current. The fault is assumed to be on phase 3. The arc will go out when I_f passes through zero. At this instant, E_{30} is also zero and I_n and I_c have nearly their maximum values. Their instantaneous values are, however, exactly equal and opposite in the sense indicated by the arrows in Fig. 1, *i.e.*, regarded as currents fed to the fault by the two parallel circuits (1) coil-fault-faulty wire-E_{30} and (2) admittance of sound phases-fault-faulty wire-transformer bank. These instantaneous currents are exactly equal in magnitude and are in the same direction in the single series circuit consisting of coil, transformers, admittance to ground of the three phases in parallel, and ground. Thus the condition in this series resonant circuit at the instant of extinction is that of an established free oscillation, the energy of the oscillation being at this instant wholly electromagnetic.

The voltage across the reactor due to the current I_n, in the direction of E_{30} (Fig. 1) is represented in Fig. 2 by the vector E_{0g}. This is 180° out of phase with E_{30} and initially of the same amplitude. At the instant the arc goes out, both are practically zero. As the oscillation progresses, E_{0g} dies away, due to damping, and the resultant

voltage of the faulty wire to ground, *viz.*, $E_{2g} = E_{30} + E_{0g}$, passes gradually back to the normal value E_{30}. At the same time the voltages to ground of the two sound phases (E_{1g}, E_{2g},) return to their normal values E_{10}, E_{20}. The ends of the three vectors, E_{3g}, E_{1g}, E_{2g}, may be thought of as sliding at equal rates along the line E_{30} and the dotted lines parallel to it.

The effectiveness of the action just sketched (it has been assumed that the frequency of the series resonant circuit is accurately that of the system fundamental), of course, depends on the accuracy of the tuning and the amount of damping. If the free period of the resonant circuit differs considerably from the fundamental period of the system, the impressing on the faulty wire of a voltage to ground in excess of normal may result, especially if the damping is small. The effect of inexact tuning is discussed by Petersen in the article referred to above. He describes some experiments in which the capacity of the power system to ground was varied some 15 or 20%, each way from the value corresponding to resonance, without apparent effect on the quenching action of the reactor.

2. *Transient Overvoltage on Sound Phase at Time of Grounding*

To simplify the following theoretical discussion a single phase system is treated. This is represented in Fig. 3. Referring to this

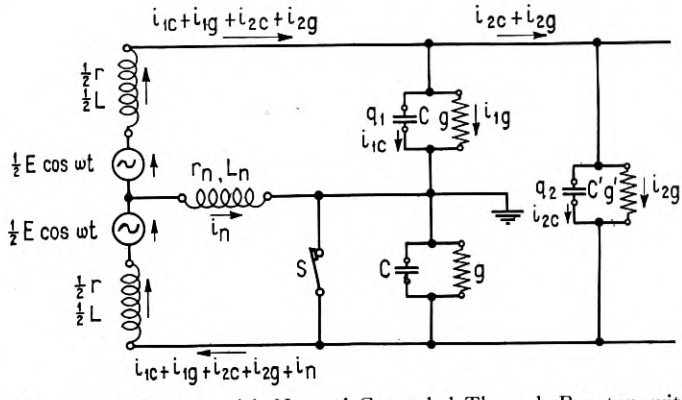


Fig. 3—Three-phase System with Neutral Grounded Through Reactor, with Fault to Ground on One Phase

figure, when one phase is grounded (represented by the closing of the switch S), the following equations, in which D denotes differentiation with respect to time, must be satisfied:

$$\Phi(D)\, i_n = \frac{E}{2}\cos \omega t$$

$$\frac{1}{g}\Phi(D)\, i_{1g} = \left[\frac{r}{2} + 2r_n + \left(\frac{L}{2} + 2L_n\right) D\right]\frac{E}{2}\cos \omega t$$

where

$$\Phi(D) = LC'_0\left(\frac{L}{2}+2L_n\right)D^3+\left[C'_0(rL+2L_nr+2r_nL)+g'_0L\left(\frac{L}{2}+2L_n\right)\right]D^2$$
$$+\left[g'_0\,(rL + 2L_nr + 2r_nL) + C'_0\left(\frac{r^2}{2} + 2rr_n\right) + L + 2L_n\right]D$$
$$+ g'_0\left(\frac{r^2}{2} + 2rr_n\right) + r + 2r_n$$
$$C'_0 = C + C',\ g'_0 = g + g'.$$

To solve the equations, the cubic equation $\Phi(D) = 0$ must be solved. An algebraic solution would be so cumbersome as to be impracticable. The following numerical values of the constants have therefore been inserted, as what is desired is a numerical solution representing the effect in a practical case:

$g = 0.37 \times 10^{-6}$ mho	$g' = 0.18 \times 10^{-6}$ mho
$C = 0.55 \times 10^{-6}$ farad	$C' = 0.30 \times 10^{-6}$ farad
$L_n = 6.4$ henries	$r_n = 200$ ohms
$L = 0.022$ henry	$r = 2.0$ ohms
$E = \sqrt{2} \times 26{,}400 = 37{,}350$ volts	$\omega = 377$

With these assumptions

$$\Phi(D) = 2.4 \times 10^{-7}D^3 + 29.4 \times 10^{-6}D^2 + 12.8\, D + 402,$$

of which the roots are

$$-31.4,\ -45.5 + j7{,}300,\ -45.5 - j7{,}300$$

which may be denoted by $-a'$, $-a + jb$, $-a - jb$ respectively. The resulting equations for i_{1g} and i_n are

$$i_{1g} = Pe^{-a't} + Qe^{-at}\sin(bt + \Theta) + gE\cos \omega t,$$

$$i_n = P'e^{-a't} + Q'e^{-at}\sin(bt + \Theta') + \frac{E}{4820}\sin(\omega t + 4°.7).$$

The relations between the two sets of arbitrary constants may be obtained by inserting these solutions in the following differential equation connecting i_{1g} and i_n:

$$i_{1g}/g - \left[r/2 + 2r_n + (L/2 + 2L_n)\, D\right] i_n = 0,$$

and the three independent arbitrary constants so found are determined by the following conditions when $t = 0$ (it is assumed that breakdown

occurs when the impressed voltage to ground, $E \cos \omega t/2$, is a maximum):

$$i_n = 0,$$

$$i_{1g} + i_{2g} + i_{1c} + i_{2c} = (g'_0 + C_0'D)\frac{i_{1g}}{g} = 0.0173 \text{ ampere},$$

$$q_1 + q_2 = \frac{i_{1g}}{g} C'_0 = 0.0216 \text{ coulomb}.$$

The numerical quantities in the second and third of these equations are respectively the total current supplied to the sound wire and the total charge on this wire at the instant of breakdown. They are obtained by solving the network of Fig. 3, with switch S open and taking instantaneous values when the impressed e. m. f. is a maximum.

The resulting expressing for i_{1g} is

$$i_{1g} = 0.3 \times 10^{-6}e^{-31.4t} - 4.38 \times 10^{-3}e^{-45.5t} \cos 7300t + 0.37 \times 10^{-6} \times 37{,}350 \cos 377t.$$

The non-oscillatory term $0.3 \times 10^{-6}e^{-31.4t}$ is seen to be negligible compared to the others.

The voltage between the sound wire and ground is obtained by dividing the above result by $g = 0.37 \times 10^{-6}$, and is

$$v = 0.8e^{-31.4t} - 11{,}800e^{-45.5t} \cos 7{,}300t + 37{,}350 \cos 377t.$$

This equation is plotted for about 1½ cycles of fundamental frequency in Fig. 4. The non-oscillatory term is negligible. As will be seen, the maximum overvoltage is about 30%.

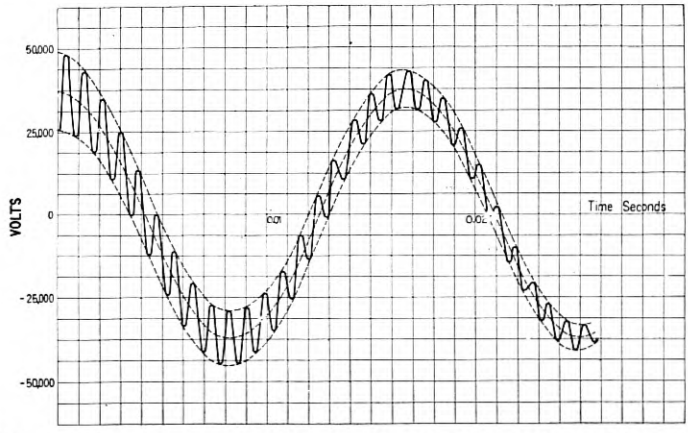


Fig. 4—Voltage from Sound Phase to Ground Following Extinction of Arc.

For comparison purposes, the voltage to ground of the sound phase with the reactor omitted (*i.e.*, with the system isolated) has been calculated, using the same constants as before. The result is

$$v' = -11{,}900e^{-45.8t} \cos 7310t + 37{,}350 \cos 377t .$$

This is practically identical with the earlier result; that is, the voltage to ground of the sound phase is practically the same with the reactor as with the system isolated.

3. *Effects with Respect to Induction in Neighboring Communication Circuits*

An estimate of the value of the Petersen coil must involve a comparison with other methods of grounding the neutral (including grounding through an infinite impedance, *i.e.*, the isolated system) or of otherwise limiting the effects of abnormal occurrences in a power system. As regards the induction of fundamental frequency voltages in exposed communication circuits, the methods of chief importance in such a comparison, at least so far as American practice is concerned, are that in which the neutral is grounded either directly or through a low resistance, and that in which the neutral is isolated.

When accidental grounds occur on a power system with neutral grounded through zero or a low impedance, the resulting heavy short circuit currents to ground may produce severe disturbances in exposed communication circuits. Owing to the fact that these disturbances are produced by electromagnetic induction in a circuit consisting of the communication conductor as one side and the earth as the other, they cannot be avoided by enclosing the communication conductors in lead-sheathed cable, even when this is placed underground.

With the Petersen coil, according to the explanation in the first part of this section, the neutral current of fundamental frequency due to a fault to ground is made equal to the charging current of the system to ground with one phase grounded, and this is generally a small fraction[2] —a few per cent. or less—of the neutral current in an identical system with neutral directly grounded.

The Petersen coil will thus in many cases largely prevent the electromagnetic inductive effects at fundamental frequency which appear when a fault to ground occurs on a system grounded solidly or through

[2] Exceptions to this statement exist in the case of extensive high voltage networks where, with a ground on one phase, the charging current to ground with isolated neutral may be of the same order of magnitude as the short circuit current with dead-grounded neutral if the fault is remote from a point of main power supply.

a low resistance. There will appear, however, electrically[3] induced voltages of fundamental frequency substantially identical with those that would occur with neutral isolated. Where the communication circuits are in underground cable, these voltages are of inappreciable magnitude, and with aerial cables (with metallic sheaths) their effects can in general be controlled without great difficulty. With open wire communication circuits, electrically induced voltages are of much more consequence. They may in some cases equal or exceed the voltages which would be induced electromagnetically with dead-grounded neutral. However, except perhaps in cases of long exposure to high voltage power circuits at close separations, their effects are generally much less severe than the electromagnetic effects, because of the smaller amount of energy transferred to the disturbed circuit. This is in general accordance with experience with open wire circuits exposed to power circuits of moderate voltage. As is explained in the next paragraph, the use of a Petersen reactor to ground the neutral may be expected to lessen the severity of inductive effects which would be experienced from an isolated system, by preventing additional parts of the power system from becoming involved.

As compared to the isolated system, the use of the Petersen coil in the connection from neutral to ground may be expected to have the advantage, according to the theory of the first subdivision of this section, of preventing the formation of an "arcing ground." As experience has shown, an arcing ground in an isolated system is frequently the cause of serious disturbances which may involve portions of a network remote from the location of the original trouble. The advantage of the reactor in this respect is, of course, a fundamental one from the standpoint of power operation. It is in general of proportionate importance from the inductive interference point of view, at least where a power network is involved in parallels with communication circuits at several places, as is not infrequenctly the case near large cities. A breakdown to ground in the power network on a different phase from that originally involved, and in a different locality, may lead to large phase-to-phase currents in the earth, from the second fault to the first, or to a ground intentionally placed on the phase first involved, in order to short circuit the arc. The inductive effects thus become electromagnetic in character, and the interference produced in this manner may be severe.

The possibility that the reactor might tend to produce a greater

[3] "Electrically" is used here and elsewhere in this paper in the sense in which "electrostatically" is perhaps more commonly used. The phenomena involved are not static and the latter word is inappropriate on this account.

overvoltage on a sound phase at the instant of grounding than would be the case in an isolated system is, of course, of importance in this connection. This has been examined from a theoretical standpoint in the second subdivision of this section, with the conclusion that there is no material difference in this respect.

There remains the method of grounding in which a high resistance is employed in the connection from neutral to ground. By "high" here may be meant the "critical"[4] resistance or one of smaller magnitude, but still so large that in the event of a solidly grounded phase, the sound phases are brought to subtsantially full delta voltage above ground. There are probably few cases where electromagnetic indutive effects due to accidental grounds on a power system are a matter of importance, in which a neutral resistance small enough to avoid this rise of voltage on the sound phases would be effective as a measure of relief. This method of grounding would thus not avoid the electrically induced voltages which arise when the reactor is used, although it would presumably be effective in preventing the spread of trouble to other parts of the power system if positive operation of selective relays is secured. Inasmuch as it presents fewer difficulties from this last point of view than the Petersen reactor, grounding through a moderate resistance has a definite advantage over the latter method from the standpoint of inductive effects at fundamental frequency, provided sufficient resistance can be used to limit the electromagnetically induced voltages to tolerable values.

Where this is impracticable from the standpoint of power system operation, the relative merits of the two systems would have to be decided by balancing the effective suppression of the transient electromagnetic inductive effects by means of the reactor, plus the expectation of occasional disturbances continuing over the intervals necessary for the location and disconnection of the faulty line, against the imperfect suppression of the former effects by means of the resistor, plus the limitation to very brief intervals of electrically induced disturbances otherwise the same as with the reactor. It is obvious that the factors controlling such a decision would vary widely in different cases, and that practical experience with both methods would be of great value in estimating their relative importance. It is, of course, possible that future development may remove some of the disadvantage at which the Petersen coil now finds itself in respect to the matter of relay protection for interconnected networks. Such development would presumably be of importance also to the critical

[4] *I.e.*, the resistance for which a discharge to ground passes from the oscillatory to the non-oscillatory type.

resistance which, as a method of grounding the neutral, would generally suffice to prevent interference from ground currents at times of faults to ground, but which apparently presents difficulties from the relay standpoint similar to those involved in the use of the Petersen coil.

II. Effects with Power System in Normal Condition

1. *Fundamental Frequency*

Referring to Fig. 5 (we now take account of inequalities in the admittances to ground),

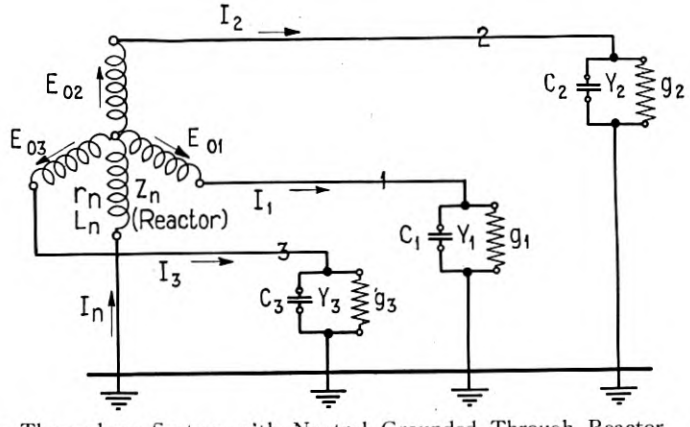


Fig. 5—Three-phase System with Neutral Grounded Through Reactor. Admittances to Ground Not Balanced.

$$E_{01} - \frac{I_1}{Y_1} = E_{02} - \frac{I_2}{Y_2} = E_{03} - \frac{I_3}{Y_3} = Z_n I_n,$$

$$I_1 + I_2 + I_3 = I_n,$$

so that

$$I_n = \frac{Y_1 E_{01} + Y_2 E_{02} + Y_3 E_{03}}{1 + Z_n Y},$$

where $Y = Y_1 + Y_2 + Y_3$ = total direct admittance to ground. If the impressed voltages are balanced

$$I_n = \frac{E_{01}}{1 + Z_n Y}\,(Y_1 + Y_2\, e^{j120°} + Y_3\, e^{j240°}).$$

The parenthesis is the " residual admittance " [5] to ground and, if the three leakances to ground are equal, it can vanish only if

[5] Inductive Interference between Power and Communication Circuits, California Railroad Commission, p. 269.

the three direct capacities to ground are the same. The equation is equivalent to

$$I_n = \frac{E_{rc}}{3} \frac{Y}{1 + Z_n Y},$$

where E_{rc} is the "characteristic residual voltage" [6] of the (isolated) system and is equal in magnitude to 3 E_{01} times the ratio of the residual admittance to ground to the total admittance to ground.

We have

$$YZ_n = r_n g - \omega^2 L_n C + j\omega (C r_n + L_n g)$$

r_n and L_n being the resistance and inductance of the earth coil, and g and C the total leakance and capacity to ground, respectively. Also, for resonance

$$\omega^2 L_n C = 1$$

and hence the above expression for I_n becomes (at fundamental frequency)

$$I_n = \frac{E_{01} (Y_1 + Y_2 e^{j120°} + Y_3 e^{240°})}{\frac{r_n}{\omega L_n} \frac{g}{\omega C} + j \left[\frac{r_n}{\omega L_n} + \frac{g}{\omega C} \right]}$$

if the earth coil is adjusted for accurate resonance.

If in this equation the denominator on the right be denoted by x and the fractional unbalance of the admittance to ground by y $\left(i.e., y = \frac{Y_1 + Y_2 e^{j120°} + Y_3 e^{j240°}}{Y} \right)$, then

$$I_n = \frac{y Y E_{01}}{x} \tag{3}$$

and, V_n being voltage between neutral and ground,

$$V_n = Z_n I_n$$

$$= \frac{y (x - 1)}{x} E_{01}. \tag{4}$$

If the losses in the system (including the earth coil) are small, x is small compared to 1. Thus we get, for the absolute value of V

$$| V_n | = | \frac{y}{x} . E_{01} |, \text{ approximately}$$

and the absolute value of the residual voltage is three times this, approximately.

[6] *Ibid.*, p. 257.

In substance, this means that, at fundamental frequency and with small losses, the fractional admittance unbalance should be kept small compared to the ratio of the resistance to the reactance of the coil, if unduly high voltages to ground are to be avoided. (It is supposed that $g/\omega C$ is of the order of $r_n/\omega L_n$, or smaller—a condition probably satisfied even with the best practiable design of coil, except under very wet line conditions.) The point involved here is, of course, an important one from the standpoint of power system operation. It is also important from the standpoint of electrically induced voltages in exposed communication circuits. The admittance unbalance can be kept within the necessary limits by suitable power circuit transpositions.

The absolute magnitude of the fundamental frequency neutral current is obtained from the expression for $|V_n|$ by dividing by the coil impedance (or directly from equation (3), and is, approximately,

$$|I_n| = \frac{|yE_{01}|}{r_n + \frac{g}{\omega C}\omega L_n}.$$

For a system with dead-grounded neutral the fundamental frequency residual current is $yE_{01}Y$ and is thus smaller than that just found for the case of the reactor in the ratio of $|x|$ to 1, approximately. It is evident, however, that the magnitude of the neutral current with reactor is controllable by means of transpositions, as in the case of the neutral and residual voltages. The inductive effects of this current should be of small consequence with an amount of transposing sufficient to keep the line voltages to ground within limits desirable from the standpoint of power system operation.

2. Harmonics

In the following discussion, we retain the assumption of lumped constants, so that the results are not applicable to extensive networks without modification.

With this restriction, at harmonic frequencies other than the third or one of its odd multiples, the above approximate equation for V_n becomes

$$V_n = \frac{y'(x'-1)}{-1+\frac{1}{m^2}+x'}.\ E_{01}, \text{ approximately,}$$

m being the order and E_{01} the voltage of the harmonic and x and y accented to denote that they are to be taken for the frequency in

question. For small losses x' may be neglected in comparison with unity, as before, giving

$$| V_n | = \frac{| y' E_{01} |}{1 - \frac{1}{m^2}}, \text{ approximately.} \tag{5}$$

For isolated neutral

$$| V_n | = | y' E_{01} |. \tag{6}$$

Thus, even for the fifth harmonic, the right hand side of (5) is only 4 per cent. in excess of the value it would have if the neutral were isolated.

For harmonics whose orders are not divisible by three the residual voltage is three times V_n. Thus from the standpoint of noise interference from voltages, a system grounded through a Petersen coil behaves practically as though the neutral were isolated, so far as these harmonics are concerned. As with the fundamental, power circuit transpositions are available for the reduction of residual voltages of these frequencies.

Residual currents of frequencies belonging to this series of harmonics, which are not present at the ends of the line with isolated neutral, are introduced by grounding through the reactor, but they are of minor importance, as may be judged by comparing the neutral currents with the reactor and with dead grounded neutral. With the reactor, the neutral current of a harmonic of order m not a multiple of 3 is found from (5) to be, in absolute value,

$$| I_n | = \frac{m | y' E_{01} |}{(m^2 - 1) | Z_n |}, \text{ approximately,}$$

Z_n being the coil impedance at fundamental frequency, while with dead-grounded neutral, it would be

$$| I_n | = | E_{01} y' m Y |, \text{ approximately,}$$

in which Y is the total admittance to ground at fundamental frequency. Thus the magnitude of the neutral current with the reactor is approximately $1/(m^2 - 1)$ of its magnitude with dead-grounded neutral. The noise effects of residual currents of these magnitudes will generally be insignificant compared to those arising from other sources, particularly if the power circuit capacities to ground are well balanced.

For the third harmonic, or one of its odd multiples, we get

$$V_n = Z'_n I_n = \frac{E_{01} Z'_n Y'}{1 + Z'_n Y'},$$

in which the symbols for coil impedance and line admittance are accented to denote that they refer to the harmonic frequency in question;

$$V_n = \frac{E_{01}m^2\,(x' - 1)}{1 + m^2\,(x' - 1)};$$

and

$$|\,V_n\,| = \frac{|\,E_{01}\,|}{1 - \frac{1}{m^2}}, \text{ approximately.}$$

For isolated neutral,

$$V_n = E_{01}$$

The neutral is thus subjected to a third harmonic voltage some 12 per cent. greater than if it were isolated, but for the higher harmonics belonging to this series (of the third and its odd multiples), the difference is inappreciable.

The residual voltage for a harmonic of this series is

$$V_r = 3\,(E_{01} - V_n) \tag{7}$$

$$= 3E_{01}\,\frac{1}{1 + m^2\,(x' - 1)};$$

and

$$|\,V_r\,| = \frac{3\,|\,E_{01}\,|}{m^2 - 1}, \text{ approximately.} \tag{8}$$

The corresponding neutral current is

$$I_n = \frac{|\,E_{01}\,Y'\,|}{1 + Z'_n Y'},$$

and

$$|\,I_n\,| = \frac{|\,E_{01}Y'\,|}{m^2 - 1}, \text{ approximately.} \tag{9}$$

From the standpoint of noise interference in telephone circuits, residuals of the series consisting of the third harmonic and its odd multiples are frequently troublesome where the neutral of a three-phase system is grounded directly or through a low resistance. These residuals, of course, are not affected by power circuit transpositions, either as to their magnitudes or as to their inductive effects upon exposed telephone circuits. It is therefore of interest to examine the expressions just obtained for the case in which the neutral is grounded through the coil. While the neutral current will not be the same as

the residual current except when only one line is supplied from the transformer bank, the effect upon the former should in general be at least approximately proportional to the effect upon the residual current in any line supplied from the bank.

In writing equation (7), any difference between the induced voltage E_{01} and the voltage appearing between line and neutral has been ignored. To the extent that this is justifiable, the expressions for the case of the solidly grounded neutral may be obtained by making the denominators of the right hand sides of (8) and (9) each unity. If the transformer bank is provided with a delta winding of low impedance, in particular if it is connected delta on one side, this procedure gives a fair approximation to the correct expressions, since the impedance through which the voltage E_{01} regulates is in this case merely the transformer leakage impedance. The resulting conclusions with respect to the advantage of the reactor—for example, that the third harmonic residual voltage or neutral current is 1/8 as large, the ninth 1/80 as large, etc., with the reactor as with solidly grounded neutral—should not, in any event, be unfavorable to the latter method of grounding unless the electrical length of the line approaches the point at which its reactance to ground becomes positive.

If the transformers are so connected as to provide no path for triple harmonic magnetizing currents other than through line admittance to ground and the impedance between neutral and ground, the induced voltage E_{01} is not the same for the two methods of grounding under consideration, because the impedance to the triple harmonic magnetizing currents is appreciably different in two cases. A convenient method of taking this effect into account is to regard the induced voltage as due to a fictitious impedanceless generator of determinate voltage regulating through the mutual impedance of the transformer windings for the frequency in question.[7] If Z'_m is one-third of this mutual impedance and V_{01}[8] is the voltage of the fictitious generator, the expression for the neutral current with ground connection through the reactor becomes

$$I_n = \frac{V_{01} Y'}{1 - m^2 + Y' Z'_m}, \text{ approximately,} \qquad (10)$$

[7] H. S. Osborne, Trans. A. I. E. E. 34, p. 2175, 1915.

[8] The voltage thus assumed is, of course, not identical with the induced voltage for which the symbol E_{01} has hitherto been used, and for this reason the new symbol V_{01} is used for it. The corresponding E_{01} would be V_{01} diminished by the drop through the mutual impedance.

and the residual voltage is

$$V_r = 3\,(V_{01} - I_n Z'_m - V_n)$$
$$= \frac{3V_{01}}{1 - m^2 + Y' Z'_m} \text{ approximately.} \qquad (11)$$

The corresponding expressions for solidly grounded neutral are obtained by omitting m^2 in the denominator for each of the equations just derived. Thus the advantage of grounding through the reactor relative to grounding directly depends on the magnitude of $Y'Z'_m$ as compared to the square of the order of the harmonic. Z'_m depends upon the voltage and the kva. capacity of the transformers and is mostly inductive reactance. For high voltage transformer banks of small capacity feeding very extensive networks, the gain indicated by equations (10) and (11) from the use of the reactor would probably not be large. It would be important, however, where the aggregate capacity of the supply transformers is moderate or large and the connected network is of moderate extent and voltage. For instance, using the data of the example considered in an earlier part of this paper and taking $Z'_m = jL'_m\omega' = \frac{1}{3} \cdot j\,9{,}000$ as an appropriate value for a total transformer capacity of 7,000 to 8,000 kva., with line voltage from 20,000 to 30,000, we should have $L'_m C\omega'^2$ equal to about 4 at 180 cycles/sec. In other words, in this case, the employment of the reactor would reduce the residual voltage and the neutral current of the third harmonic frequency due to a star-star solidly grounded transformer bank by about 75 per cent., and residuals of other frequencies belonging to the same series probably by larger amounts.

In the earlier discussion relating to harmonics not belonging to the triple series, comparison was made between a system grounded through a Petersen reactor and the isolated system. In a similar comparison with respect to the triple harmonic series, the isolated system has the advantage, since residuals of this series theoretically do not appear in such a system, as the voltages are not impressed between wires. As a practical matter, an isolated system would probably not be entirely free of triple harmonic residuals, owing to dissimilarities in transformers or elsewhere. Such accidental effects can hardly be taken into account in a theoretical discussion. However, in setting up a comparison between the isolated system and that grounded through the reactor, an idea of the relative importance of the triple harmonic residual voltages existing in the latter case can perhaps be obtained by comparing their theoretical magnitudes with the theoreti-

cal magnitudes of residual voltages in the isolated system of non-triple frequencies.

The residual voltage due to one of these non-triple frequencies, which is three times the neutral voltage, is $3y' E'_{01}$, according to equation (6). Here y' is the fractional residual admittance and E'_{01} may be taken as the induced voltage in the transformer for the frequency in question. For a harmonic belonging to the triple series, with neutral grounded through the reactor, the absolute value of the residual voltage is $\frac{3 \mid E''_{01} \mid}{m^2 - 1}$ (equation (8)) m being the order of the harmonic and E''_{01} the induced voltage, if we assume the transformer bank provided with a low impedance path for triple harmonics, and therefore neglect the difference between the induced and the terminal voltages. The ratio of the triple-series residual voltage to the other is thus, in absolute value, $\frac{\mid E''_{01} \mid}{(m^2 - 1) \mid y' E'_{01} \mid}$.

If we take the ninth as the harmonic of the triple series and assume equal values of the induced voltages E''_{01} and E'_{01} it will be seen that $\mid y' \mid$ must be of the order of 0.01 if the residual voltage of the triple harmonic series is to be as large as the other. This amount of unbalance is somewhat larger than has been found at this frequency (540 cycles/sec.) in an actual transposed line.[9] If we consider the higher harmonics of the triple series, $\mid y' \mid$ would have to be made progressively smaller in order that the ratio might remain unity. Thus, for the 21st harmonic, $\mid y' \mid$ would have to be of the order of 0.002. While, of course, $\mid y' \mid$ may be made as small as desired by sufficiently close power circuit transpositions, it appears that in practical cases where transformer banks have delta windings, one may expect the residual voltages of the triple series, introduced by changing from an isolated system to one grounded through the reactor, to be relatively unimportant except in the case of the third harmonic and perhaps in that of the ninth. This statement would not be true if, as with star-star transformers under some circumstances, no low impedance path is provided for magnetizing currents of the triple harmonic series. Such cases are not common in operating practice.

The method of estimating comparative effects here applied to the case of triple harmonic residual voltages is not available for residual currents. To take account of the latter in comparing the isolated

[9] Inductive Interference between Power and Communication Circuits, California Railroad Commission, Technical Report No. 51.

system and the system with neutral grounded through the reactor, recourse may be had to the indirect method of reference to the solidly grounded neutral system, as in the discussion of residual currents of the non-triple series on page 52. Such a procedure, of course, involves a reference to general experience also. It has been shown in the earlier discussion that for a triple series harmonic of order m the neutral current with the reactor is approximately $1/(m^2 - 1)$ as large as with the dead-grounded neutral if a low impedance path for triple frequency magnetizing currents is provided, as by a delta winding. The establishment of this system of neutral currents, even though they are small, when a previously isolated system is grounded through a Petersen reactor, constitutes an addition to the residuals which produce induction in neighboring circuits. However, it is not to be expected that the added inductive effects would be important. Where no low impedance path for the triple series magnetizing currents exists, the reactor is relatively less effective in suppressing residual currents of this series. The triple harmonic neutral currents of a power system connected in this manner and grounded through a Petersen coil might in some cases lead to inductive effects of some significance.

In general, for harmonics of orders not divisible by three, grounding through a moderate resistance (large, however, compared to other impedances involved in a short circuit to ground) will be more advantageous as regards residual voltages, and less advantageous as regards residual currents, than grounding through the reactor. Grounding through zero impedance would, of course, generally lead to the smallest residual voltages and the largest residual currents of these frequencies. For frequencies belonging to the triple series, grounding through the reactor will be considerably more advantageous than grounding through a moderate resistance as regards both residual voltages and residual currents. It may be expected that with moderate neutral resistance, residual currents and voltages of the triple series will both be nearer in magnitude to those obtaining with zero neutral impedance than to those obtaining with the Petersen coil. The moderate neutral resistance is relatively more effective at the higher frequencies in reducing residual currents of all harmonics and residual voltages of the triple series; for harmonic residual voltages not belonging to the triple series, it is relatively more effective at the lower frequencies.

I wish to express my gratitude for helpful suggestions and criticism received in the preparation of this paper from Messrs. L. P. Ferris and R. G. McCurdy, and also from Mr. R. K. Honaman.

SUMMARY

1. At times of a fault to ground on a power system with neutral grounded through a Petersen reactor, the action of the latter tends to extinguish the arc and to prevent its restriking. Theoretical considerations, applied to a practical case, indicate that the transient over-voltage on a sound phase at the instant of occurrence of the fault is substantially the same as in a system with isolated neutral.

2. Grounding the neutral through a Petersen earth coil instead of directly or through a low resistance would largely prevent the electromagnetic inductive effects to which exposed communication circuits are liable at times of faults to ground in systems grounded in the latter manner. (Extensive high voltage networks are perhaps an exception to this statement. But even here, the electromagnetic inductive effects would in general not be greater with the reactor than with isolated neutral.) However, effects due to electric induction similar to those from an isolated system may be expected to appear. Except for long, close parallels involving open-wire communication circuits these effects should in general be much less severe than the electromagnetic inductive effects from a system with dead-grounded neutral. The extent and severity of the inductive effects experienced from the system grounded through the reactor would further tend to be smaller than with the isolated system, because of the effect of the reactor in preventing arcing grounds.

3. Grounding the neutral through a resistance large compared to other impedances involved in a short circuit to ground should have an advantage over grounding through the Petersen reactor, in that the former method presents fewer difficulties in respect to power system protective relays, so that it would reduce the possibility of the continuance of inductive disturbances over considerable periods of time, which might be involved in grounding through the reactor, under present relay practice. From an inductive interference standpoint, a choice between the two methods would depend upon the circumstances of particular cases. Advances in the art of relay protection would improve the position of the reactor in such considerations.

4. Under normal power system operating conditions, the use of the reactor may lead to excessive residual voltages of fundamental frequency if the admittances from phases to ground are unbalanced. Such unbalance may be reduced to the extent necessary from this point of view by power circuit transpositions.

5. Under normal operating conditions, it is to be expected that the residual voltages and currents of the triple harmonic series occurring

with neutral grounded through zero or a low impedance would be largely reduced by grounding through the Petersen reactor instead. Residuals of other harmonic frequencies should be substantially the same as with isolated neutral, and are controllable by means of power circuit transpositions. The method of grounding the neutral through a resistance of moderate value is favorable to the reduction of residual voltages of the harmonics whose orders are not multiples of three, but is relatively unfavorable to the suppression of residual currents of these frequencies. It is also considerably less effective than the reactor in preventing residuals, either voltages or currents, belonging to the triple harmonic series, which are not amenable to treatment by transpositions.

Philadelphia-Pittsburgh Section of the New York-Chicago Cable[1]

By JAMES J. PILLIOD

SYNOPSIS: Engineering and construction features involved in a complete telephone cable system over 300 miles in length and connecting Philadelphia and Pittsburgh, Pa., are described in the following paper. This cable is designed to operate as an extension of the Boston-Washington underground cable system with which it connects at Philadelphia. It is also designed for operation in connection with the Pittsburgh-Chicago cable now under construction, and other cable projects included in a comprehensive fundamental plan.

Beginning with the fundamental factor of public requirements for communication service between cities separated by various distances, there are next considered the methods available to provide this service. Small-gage, quadded, aerial cable, which was decided upon for use in this section after careful economic studies, is described in a general way and the important advantages of the application of loading and telephone repeaters are outlined. The use, in connection with this cable, of the recently developed metallic telegraph system for cables is referred to and some facts are given regarding power plants, test boards and buildings. A few of the many possible combinations of cable and equipment facilities into complete telephone circuits, which will furnish the service required as economically as now possible, are illustrated.

The necessity of complete coordination of the many factors involved in a project of this kind is emphasized.

INTRODUCTION

THE placing in service in the latter part of 1921 of the final section of a continuous telephone cable over 300 miles in length between Philadelphia and Pittsburgh marked a new point of achievement in the steady development and construction of facilities designed to render to the public the best possible long-distance telephone service. Furthermore, this cable forms an important part of a comprehensive plan of long-distance cable construction throughout that section of the United States lying in general east of the Mississippi River and north of the Ohio and Potomac Rivers.

In the discussion of a project of this kind which involves many new practices and the expenditure of several millions of dollars and which, with related work already completed, forms the groundwork for large expenditures in the future, it is usual to inquire first into the underlying reasons for carrying out the project and then into the methods adopted. In the following discussion an endeavor will therefore be made to furnish some information on these two items in their relation to the Philadelphia-Pittsburgh cable, although, as will be obvious, the many different points can be covered in only

[1] Presented at a meeting of the Philadelphia Section of the A. I. E. E., January 9, 1922, presented at the Annual Convention of the A. I. E. E., Niagara Falls, Ont., June 26–30, 1922, and appearing in the Journal of the A. I. E. E. for August, 1922.

the most general way in the space available. However, before going ahead with the discussion, I would like to point out that this project is not unlike many others in that, as a whole and in the component parts, there have been required, first, the careful consideration and decisions of the executives, then the underlying work of many scientists, inventors and engineers, then the skilled work of the manufacturers and construction forces, and finally the maintenance and operation by trained people who are responsible for the continuous service so vitally necessary to the industrial and social structure of the country. The point to be emphasized here is that the coordination of all of these factors and the close cooperation of all of the many hundreds of people concerned are the important things.

General Cable Plans and Routes

Fig. 1 is an outline map of a section of the United States and shows the routes of existing and proposed long telephone cables of the Bell system. It will be noted that the present and proposed routes follow in a general way the routes of trunk-line railroads. This general section contains more than 50 per cent of the entire population of the United States but less than 15 per cent of the area, and the industrial and telephone development is, of course, very great. Furthermore, the nearby surrounding states, supplying as they do large quantities of food products and raw materials, are commercially related to this section in a very peculiar way and this fact greatly influences the long-distance telephone development along the particular cable routes indicated. The routes through the State of Pennsylvania and the offices at Philadelphia and Pittsburgh, which are the terminals of the cable that is more particularly the subject of this discussion, occupy strategic positions in this system.

Circuits of the American Telephone and Telegraph Company and the Bell Telephone Company of Pennsylvania are carried over these routes and this cable was jointly planned and installed by these companies.

Fig. 2 is an outline map of the State of Pennsylvania and shows the situation in this section a little more in detail. On this map are shown some of the larger cities and routes of the longer and more important toll and long-distance telephone lines. As indicated, these lines are mainly of the familiar aerial wire type which has been generally used in the past for this purpose and which is today the most efficient and economical type of construction for many cases. In the general section between Philadelphia and Pittsburgh the

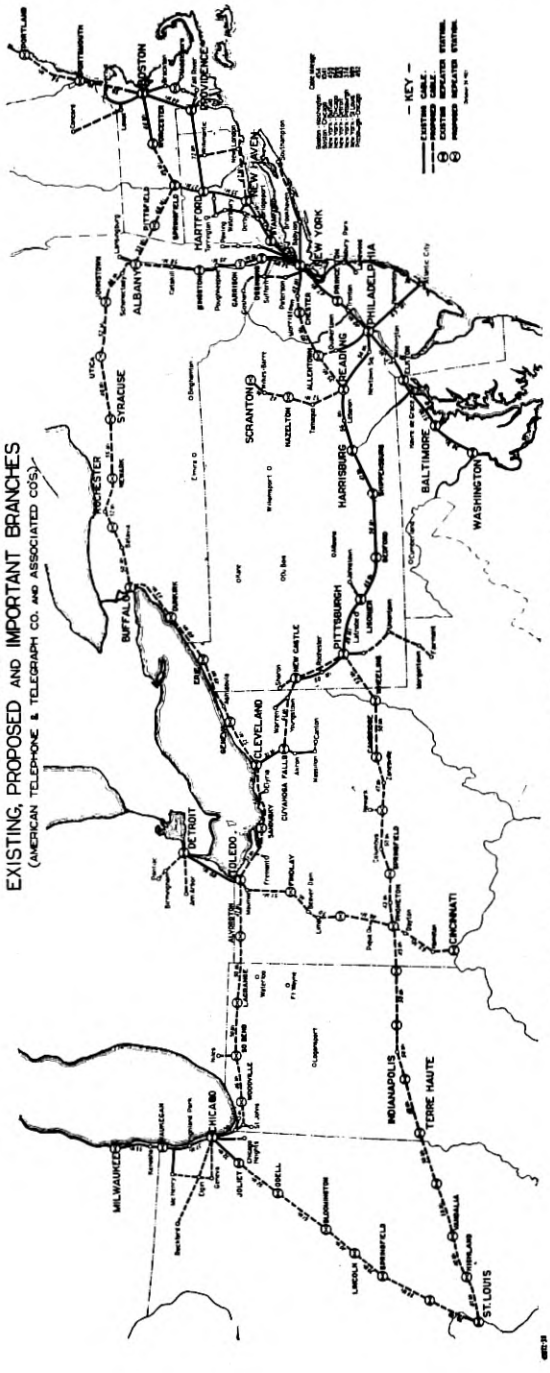


Fig. 1—Routes of Existing and Proposed Long Telephone Cables

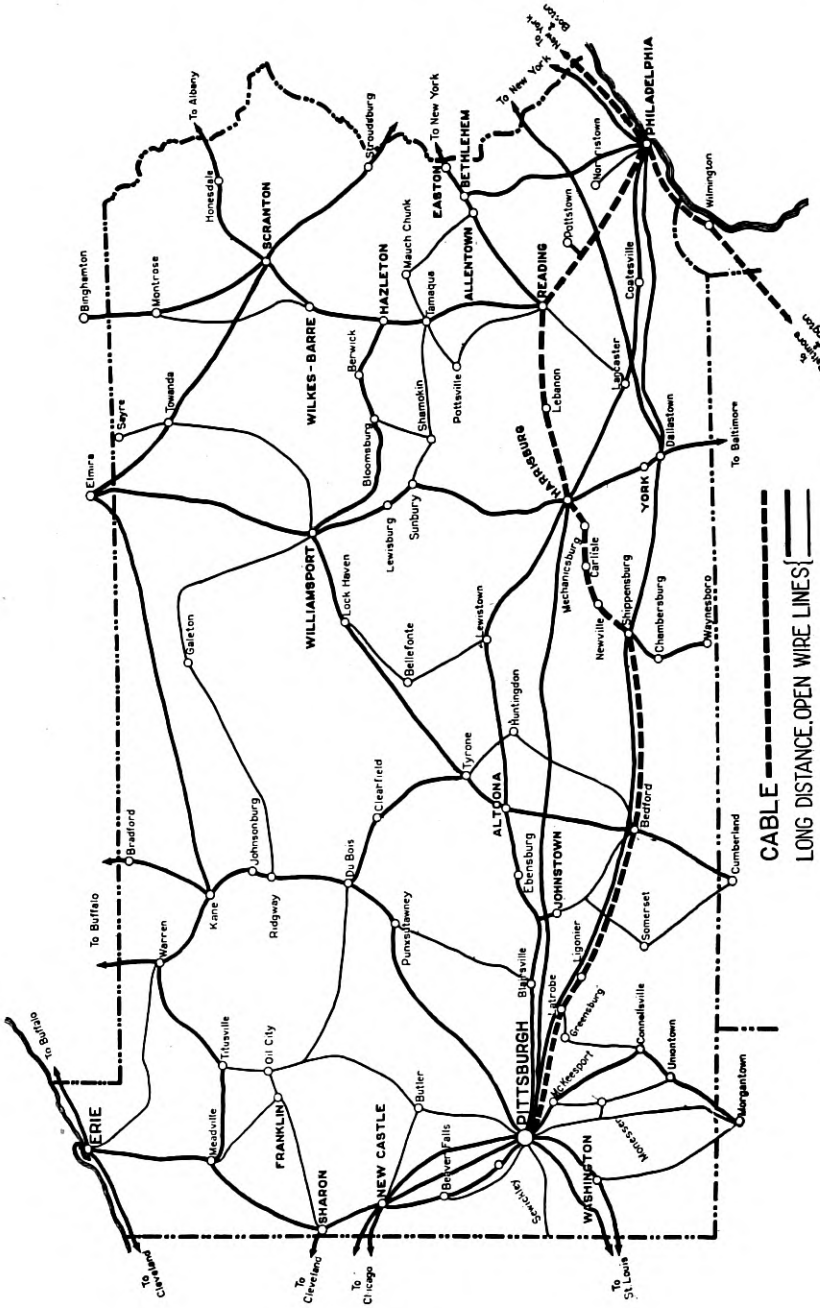


Fig. 2—Outline Map of Pennsylvania, Showing Aerial Line and Cable Routes

requirements for circuits are very heavy and in addition, as is well-known, the topography of the country is such that the through routes which can economically be used for pole lines are limited. At present, these few routes are fully occupied by the pole lines of the various utilities and included in these lines are three fully loaded telephone trunk lines. Another item of importance in the consideration of aerial wire construction is the very severe damage frequently experienced in many sections of the country on heavy aerial wire lines from ice and wind storms. Even lines built with exceptional strength fail in these storms and the interruptions to service are serious matters to the users as well as to the telephone companies. The restora-

Fig. 3—Damage to Section of New York-Boston Main Line Near Worcester, Mass. Storm of November 28, 1921

tion costs under the conditions that naturally exist at such times are abnormally high.

Figs. 3 and 4 show the effects at one point of the ice and wind storm in New England on November 28, 1921, and are proof that this problem is real. This particular spot is near Worcester, Mass., and the line is a section of one of the principal aerial wire routes between New York and Boston. In this storm, many thousands of poles were broken and even where a few poles remained standing due to specially strong construction, the load of ice combined with the wind was too great for the wires to withstand. There is therefore a practical limit to the number of wires that can be safely and economically carried on a pole line.

Where the practicable routes for pole lines are limited, where the

existing pole lines are fully loaded, and where estimated future circuit requirements are of considerable magnitude, it is obvious that different methods of providing facilities, if available, must sooner or later be given serious consideration. The conditions between Philadelphia and Pittsburgh and in general along all of the cable routes shown on Fig. 1 are now, or are expected within a few years to be, such as to make the use of some type of construction other than aerial wire desirable for most of the circuits.

After careful studies of the circuit requirements for future periods

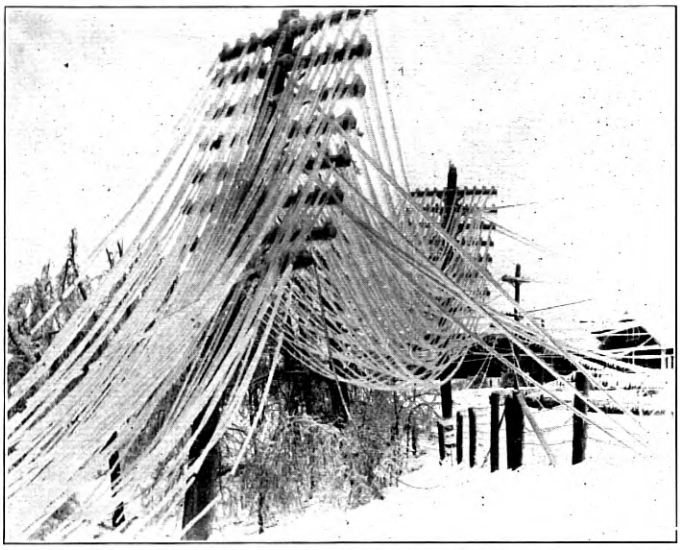


Fig. 4—Section of New York-Boston Main Line Showing Wires Heavily Loaded with Ice. November 28, 1921

and of the methods available for providing long-distance telephone facilities, which in general are aerial wire and cable, it has been decided that for relief in these sections the cable method will give the best and most economical results. Long underground cables, as is well-known, have been in operation for many years between Boston, New York, Philadelphia, Baltimore and Washington, Chicago and Milwaukee and in other sections. However, the type of cable and associated apparatus which is now being used in the development of the more comprehensive plan is quite different from that originally used between Boston and Washington and in the other sections, particularly in the use of copper conductors of a smaller gage combined with improved loading coils, the vacuum tube telephone repeater

and other methods and apparatus which are the result of recent developments. Lead-covered aerial cable supported on wooden pole lines is to be used in general on all of the routes except in the two sections just mentioned and through cities or where special conditions exist for short distances. The possibility of now using conductors of No. 16 and No. 19 A. W. G. instead of conductors up to

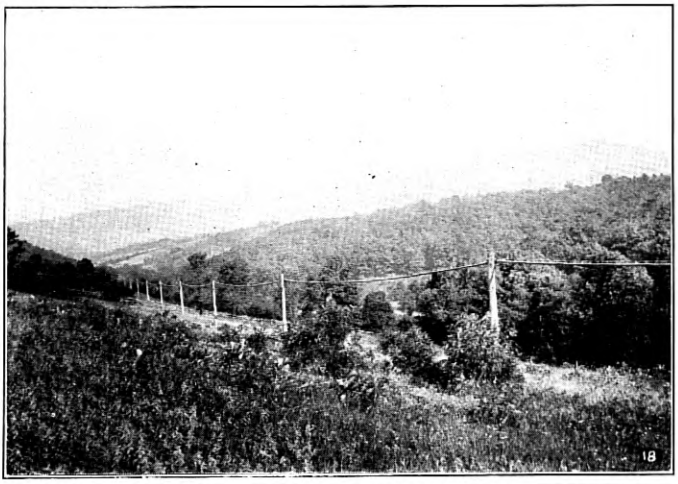


Fig. 5—General View of Pole Line Carrying Aerial Cable

No. 10 A. W. G. as in the older cables, has contributed to make aerial construction rather than underground conduit the more economical in many sections, as one cable will provide for a much greater number of circuits and consequently fewer cables will be required.

Line Construction

The general type of aerial construction which was used for over 250 miles of the total distance of 302 miles from Philadelphia to Pittsburgh may be seen from Figs. 5 and 6 which illustrate the poles, steel suspension strand, metal supporting rings and the cable. The poles are 25-foot untreated chestnut spaced 100 feet apart and designed to carry additional cables in the future. While the poles are new and carry only one cable they have a factor of safety of about 9 under the most severe storm conditions expected, but this will, of course, be reduced as other cables are placed and will gradually be decreased on account of decay at the ground line until it becomes necessary to start replacing the poles. Many of these poles were grown near the locations where they now stand. In other sections, it is planned

to use butt-treated chestnut or cedar poles, or creosoted pine poles where these prove to be the more economical.

The galvanized steel suspension strand has a breaking strength of about 16,000 pounds and the actual tension under normal conditions is about 7,000 pounds. In placing the strand, it is necessary to pull it to just the right tension in order that when the cable is hung it will have the proper sag. The correct tension is readily determined by what is known as the "oscillation" method. The metal rings are spaced 16 inches apart and the cable weighs about 7½ pounds per foot.

The size and make-up of the cable vary somewhat with the number of circuits of the various types that are to be provided in the different sections, but in general it is full size, that is, its over-all diameter is

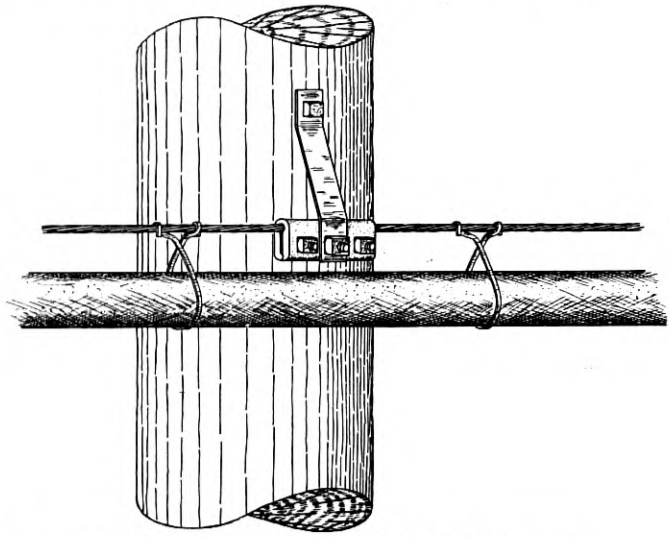


Fig. 6—Method of Supporting Aerial Cable and Messenger

2⅝ in. which is about the maximum size of telephone cable. The sheath is of lead-antimony alloy, one-eighth of an inch thick, and under normal conditions it is, of course, air-tight to keep moisture from entering. The cable for the aerial section was received from the factory in 500-foot lengths, this being largely determined by the arrangement necessary to permit the proper installation tests.

Route

We might next consider the route selected and for this purpose Fig. 2 will again be helpful. It will be noted that starting at Phila-

delphia, the cable is routed to Reading touching Pottstown, Phoenixville and other points. From Reading to Harrisburg the cable follows closely the William Penn Highway, although in sections it was necessary to obtain private right-of-way or to use longer routes removed from this highway on account of the lines of various kinds already in operation there. It is very desirable for economic reasons to keep the length of these cables as short as possible and in some cases this is absolutely necessary to obtain proper operating conditions.

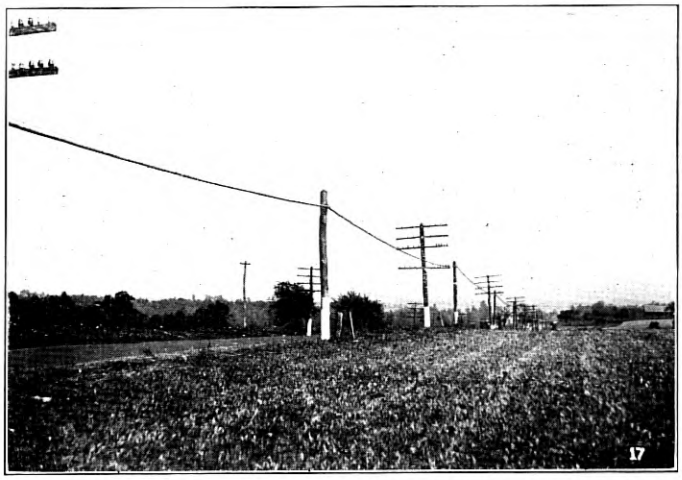


Fig. 7—Cable Line on Seven-Mile Stretch of Lincoln Highway. Aerial wire line to be dismantled later.

However, the most direct routes cannot always be used, for many obvious reasons, and this problem required careful consideration in all sections of the cable.

Between Harrisburg and Pittsburgh the Allegheny Mountains had to be crossed and for this crossing only two general routes were found practicable, the first following an existing pole line which is the New York-Chicago telephone line through Lewiston, Altoona, etc., and which we may call the northern route, and second a southern route through Shippensburg, Bedford and Ligonier for the most part along the Philadelphia-Chicago line and also the Lincoln Highway. A middle route which is now used for the Harrisburg-Pittsburgh line was not seriously considered as the country was too rough for economical construction and maintenance and no important advantages were to be obtained. After careful surveys and cost studies, taking into account all existing and anticipated conditions, such as circuit

requirements and towns to be reached, length of practicable routes, maintenance conditions, freedom from probable physical and electrical interference, etc., it was decided to build on the southern route.

This route, while of nearly the same length as the northern one and offering some important advantages, was not free from difficulties as it crosses the Allegheny Mountains within a few miles of the highest

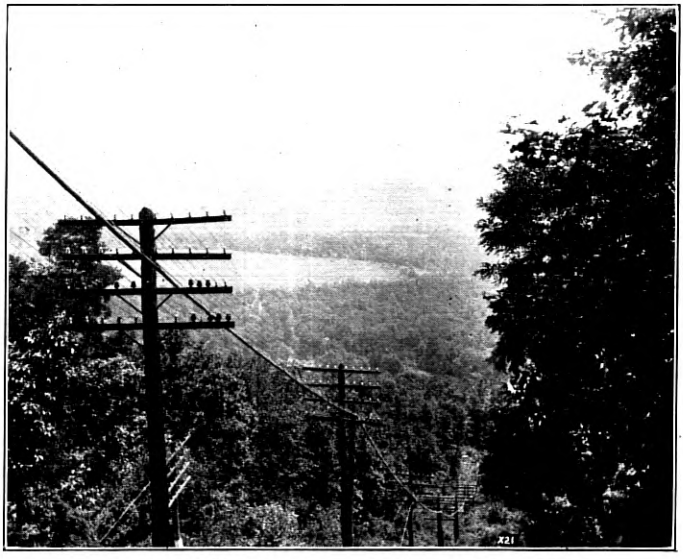


Fig. 8—Cable Line Across Valley at Grand View

point. Fig. 7 shows the cable line on what is known as the seven-mile stretch of the Lincoln Highway east of Ligonier, and here the going was fairly good. The Philadelphia-Chicago aerial wire line is also shown and two of the crossarms carrying 10 wires each are to be removed in the near future and the circuits operated in the cable. It is planned to remove the remaining two crossarms later on. Fig. 8 shows the cable across a valley and is taken from the point on the Lincoln Highway called Grand View. Fig. 9 shows the crossing of the Juniata River east of Bedford where special construction was used. Fig. 10 shows just one example of the conditions encountered in crossing the many mountains and a photograph does not do the scenery or the construction difficulties justice. On account of the steep slopes, clamps are used at many points to fasten the cable to the strand.

Narrow-gage timber railroads were used in the mountains where possible to get material to the job and Fig. 11 shows one of the regular

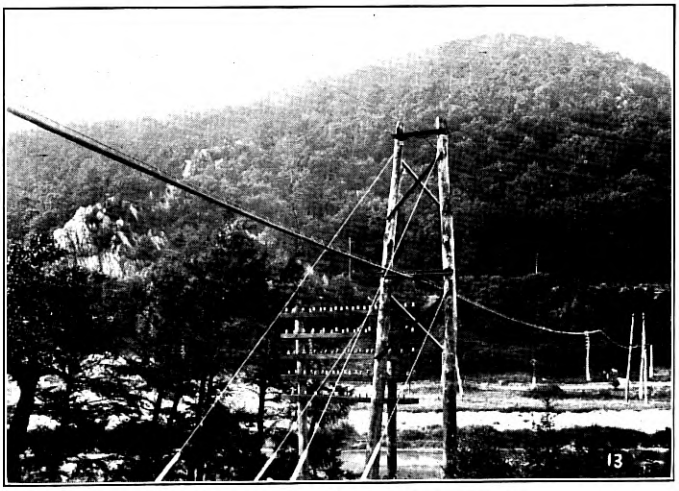


Fig. 9—Cable Crossing at Juniata River

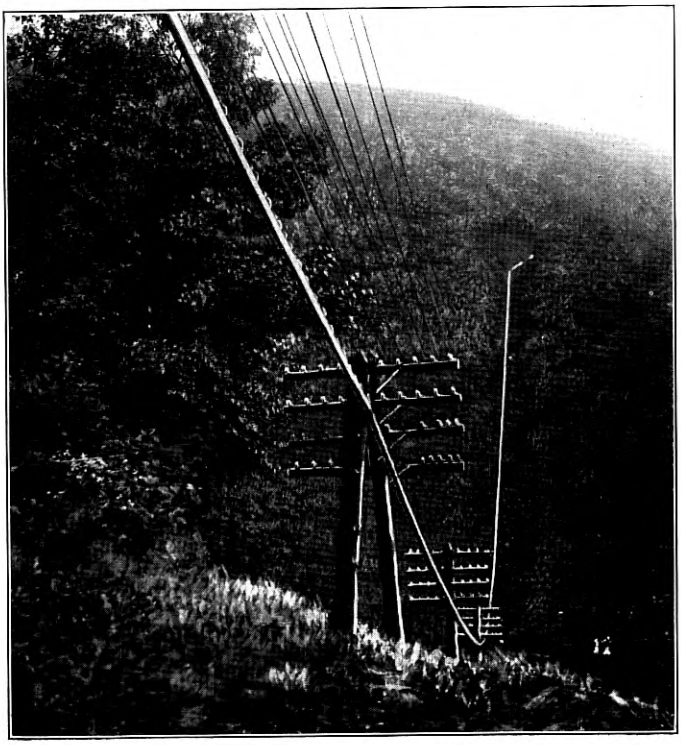


Fig. 10—Cable Line on Steep Slopes

flat cars adapted for our purpose. Fig. 12 shows two 5-ton tractors in action on top of one of the mountains. As many sections of the country are very rough and highways several miles distant it seemed that no other method of transporting the cable reels, which weigh

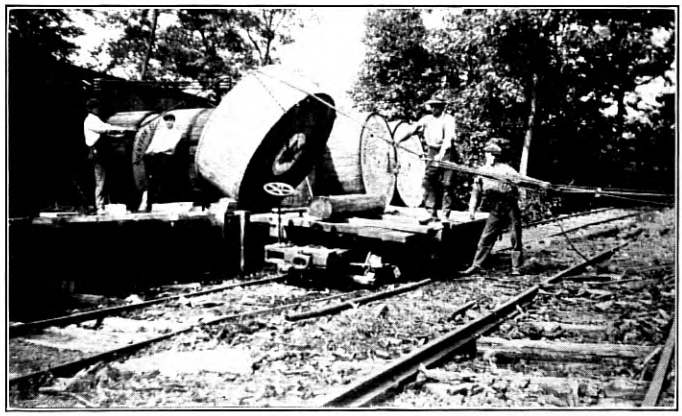


Fig. 11—Narrow-Gage Mountain Railroad and Flat Car

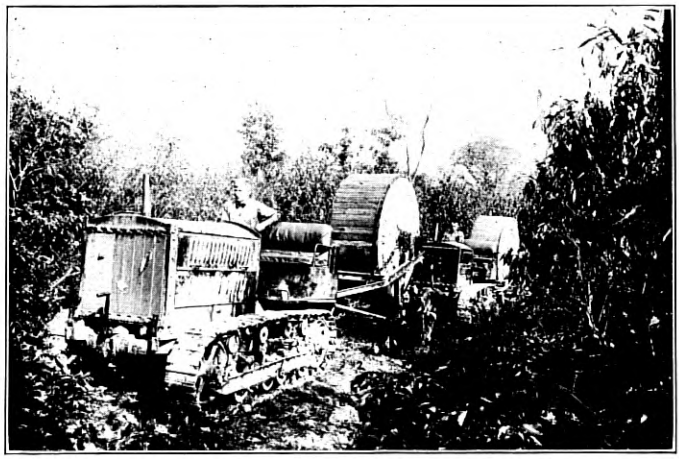


Fig. 12—Tractors Placing Cable Reels in Rough Country

nearly 5,000 pounds, could possibly be used, and certainly no other means would have been as satisfactory. Even with these methods the cable reels could not always be delivered where desired and in some cases it was necessary to pull the sections of cable through the rings for a distance of nearly a mile to get them in place.

Cable Make-Up

As stated before, the make-up of the cable varies somewhat with the circuit requirements in the different sections but the wires and arrangement in a typical section of cable are roughly illustrated in Fig. 13.

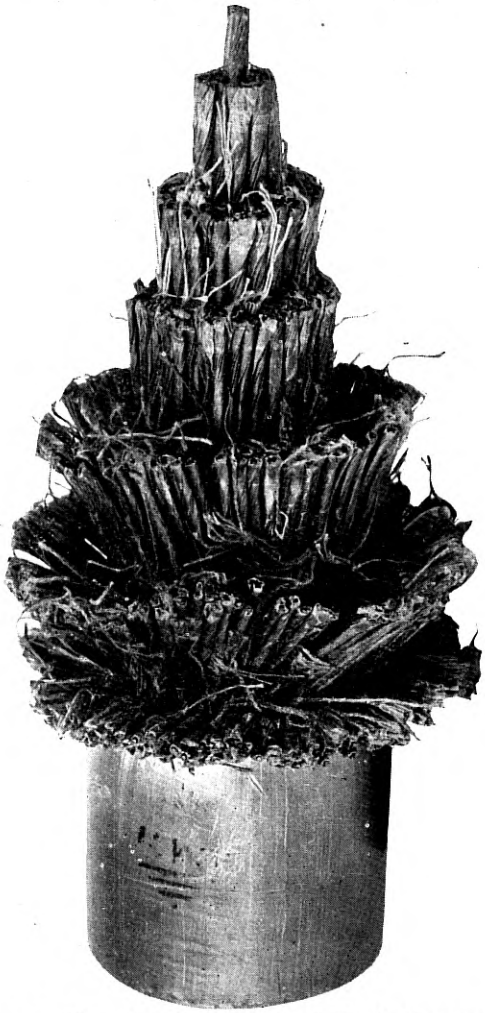


Fig. 13 Piece of Cable with Sheath Partly Removed

The cable is of quadded construction, that is, the wires are first wrapped with dry paper for insulation and twisted into pairs and then two pairs are twisted into what is called a quad. These quads are

arranged in concentric layers as shown and great care and skill are required in the design and manufacture or there is certain to be serious cross-talk between the several hundred circuits when used for long-distance service. Even after the application of the best present manufacturing methods, tests are made on all circuits at three points in each loading section of 6000 feet while the cable is being spliced. These tests are made in order to determine the best possible arrange-

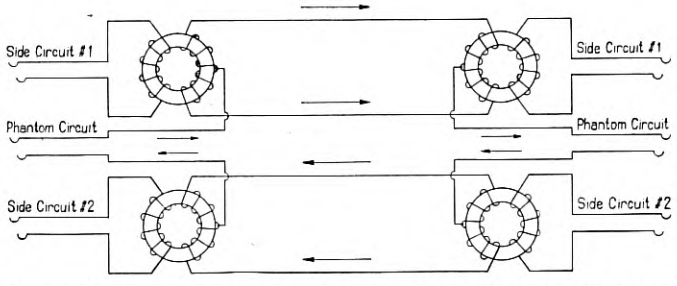


Fig. 14—General Phantom Circuit Arrangement. Four wires providing three circuits

ment of conductors for still further reducing cross-talk between circuits, and the splicing is done accordingly.

There are 19 quads of No. 16 A. W. G. and 120 quads of No. 19 A. W. G. pure copper conductors in one of the principal sections, and the arrangement of the four wires in each quad is such that two physical circuits and one phantom circuit are made available. The method of obtaining three telephone circuits from two pairs of wires is old and extensively used. It is illustrated in Fig. 14. The method results in a 50 per cent increase in the number of available circuits and its application to this project is therefore of very great economic importance. Now the total of 139 quads multiplied by 3 gives 417 circuits or as many as could be carried on about 14 heavily loaded pole lines if aerial wire were used, but as will be described later, we will have to use two of these circuits to make one telephone circuit in some cases where the distances are comparatively great, so it is expected that only about 300 telephone circuits will be obtained for regular service. This is as many as could be carried on 10 heavily loaded pole lines if aerial wire were used. It is now thought that in some sections this number of circuits will take care of future demands for about 10 years after allowing for the dismantling of some existing aerial wire.

As these cables can be obtained in any size desired up to the maximum, the period for which they should be engineered can be determined from studies of circuit requirements and costs. These studies are of very great importance and the cost considerations include, of course,

annual costs of the various plans over proper periods as well as first costs.

Loading

Loading coils are now connected to many of the circuits and all of the circuits in this cable are intended to be equipped with coils located at 6000-foot intervals. The theory and practice of loading are described in papers previously presented before the Institute[1] and for our purpose it will be sufficient to state that these loading coils very materially reduce the attenuation losses and improve the quality of transmission as compared to cable circuits not so equipped. The improvement in so far as the attenuation losses are concerned, varies with the type of circuit and loading coils, but with one of the No. 19

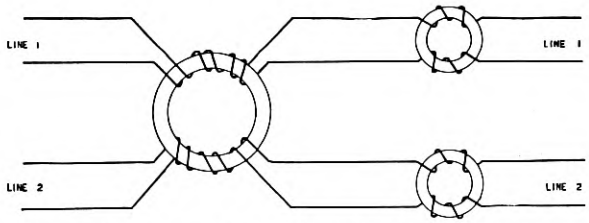


Fig. 15—Loading Coils Connected to a Group of Four Wires and Arranged for Phantom Operation

A. W. G. circuits in this cable loaded with coils having an inductance of 0.175 henry located at 6000-ft. intervals, the losses are only about one-third as great as in a similar circuit without the coils. The connections and arrangements of the coils are shown in Fig. 15 and it will be noted that coils have been connected to both the physical and phantom circuits. The arrangement is such that there is no appreciable interference between circuits due to magnetic action in the iron cores of the different coils or to the necessarily close electrical relation in the windings.

The loading coils are potted and sealed in iron pots, two of which are shown in Fig. 16, and in the country these are mounted on pole fixtures. Each pot contains 36 groups of 3 coils each. The pots are nearly 30 inches in diameter at the flange, 52 inches high and weigh about 2700 pounds. The pots can be obtained in different sizes depending upon the number of coils which it is desired to install at one time. When the cable was installed, extra lead sleeves were

[1] Papers by M. I. Pupin, Transactions of A. I. E. E., XVII, May 1900 and XV, March 1899.

Paper by Bancroft Gherardi, Transactions of A. I. E. E., XXX, June 1911.

placed at the loading points and a little slack left in the wire to facilitate the connection of four additional loading pots to the cable at

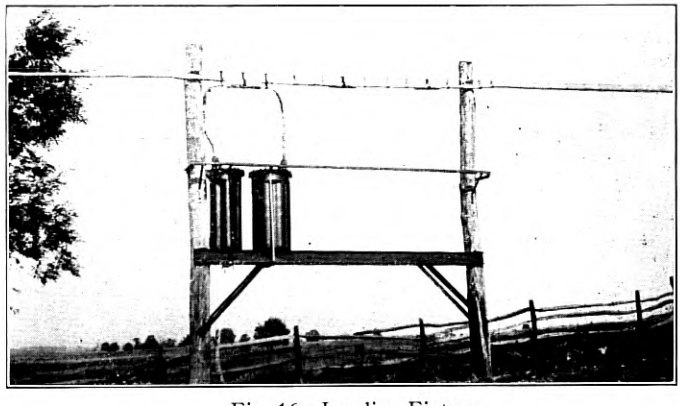


Fig. 16—Loading Fixture

some later date when the circuits are needed. The loading points must be uniformly spaced in order to obtain the proper impedance characteristics in the circuits as will be referred to later. Fig. 17 shows the iron core of a loading coil and Fig. 18 shows this core

Fig. 17—Loading Coil Core

Fig. 18—Loading Coil with Winding Completed

wound with insulated wire and then wrapped with cloth and the terminals brought out nearly ready for potting. Fig. 19 shows several

coils arranged on one of the spindles which will be placed in the iron pot also shown. This particular pot will hold 7 spindles and when they are in place, the pot will be filled with compound and thoroughly sealed.

Telephone Repeaters

Even with the improvement in the quality of transmission and reduced attenuation losses effected by the use of loading coils, loaded

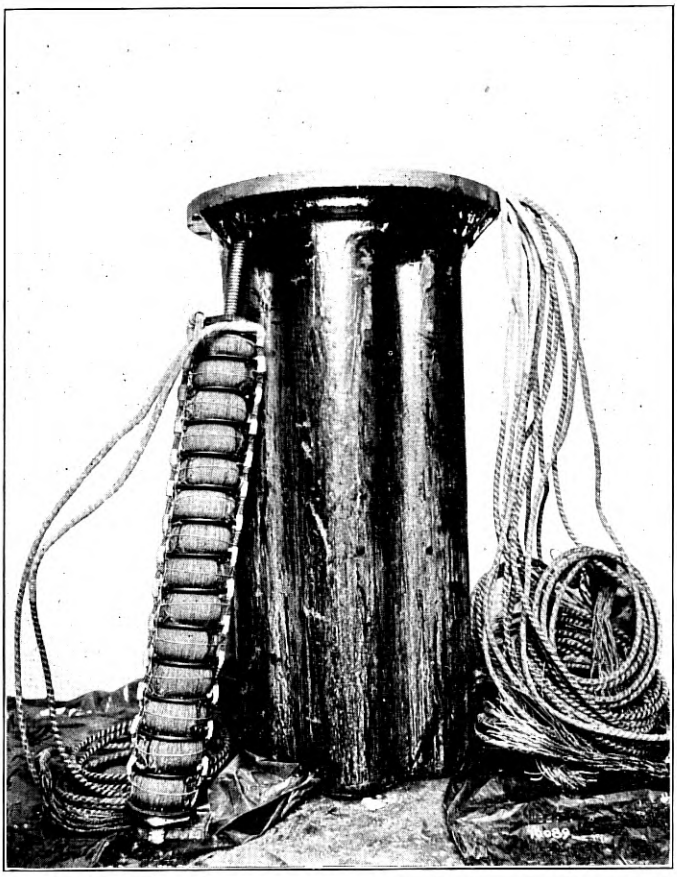


Fig. 19—Loading Coils on Spindle, Iron Loading Coil Case and Spindle Cables

cable circuits alone of No. 16 and No. 19 A. W. G. could be satisfactorily operated for distances less than 100 and 60 miles, respectively, and this is far short of our requirements in this case. In fact, we wish to operate some telephone circuits on these conductors and through this cable and future cables up to at least 1000 miles in length. This

can be accomplished by the use of telephone repeaters connected to the loaded conductors.

Telephone repeaters have been developed to a high state of perfection and are completely described in a paper presented by Messrs. Bancroft Gherardi and Frank B. Jewett at a joint meeting of the A. I. E. E. and the Institute of Radio Engineers in New York, October 1, 1919. Briefly, the purpose of a telephone repeater is to receive small telephone currents, amplify them and send them on, preserving

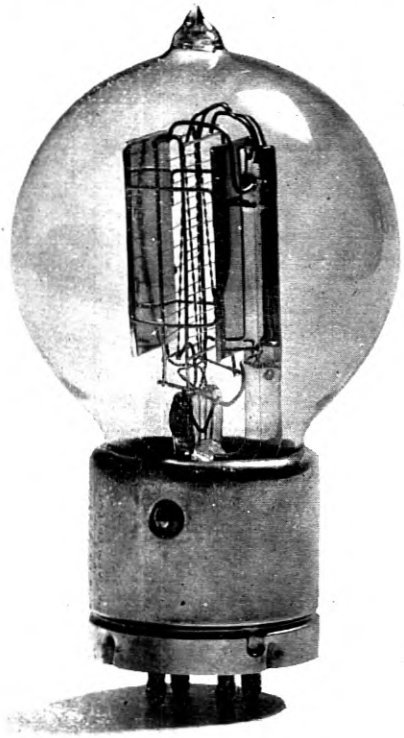


Fig. 20—Vacuum Tube

all the while the original wave shape. Therefore, if one or more telephone repeaters are properly inserted in circuits adapted to their use, the range of satisfactory transmission can be greatly extended. As many hundreds of vacuum-tube repeaters are in operation on the Philadelphia-Pittsburgh cable and connected cables, and as a great many more are planned for future installation, we will briefly consider the elementary features of some of the types of repeaters used.

Fig. 20 shows the structure of the vacuum tube which is an essential

element of this type of repeater. It is a small glass bulb with a vacuum that is as good as is practicable to obtain. In the tube is a filament which is heated to incandescence during operation,

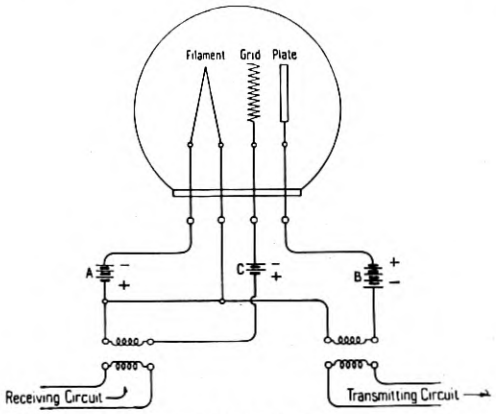


Fig. 21—Vacuum-Tube Repeater Element

and a grid and plate. The circuits directly associated with the tube are shown in more detail in Fig. 21, and this would constitute a device for amplifying currents from one direction. As is well understood, any change in the potential impressed on the grid causes a change in the current flowing in the plate-filament circuit. To obtain complete two-way repeater action two of these amplifier arrangements are combined with the circuits shown in Fig. 22.

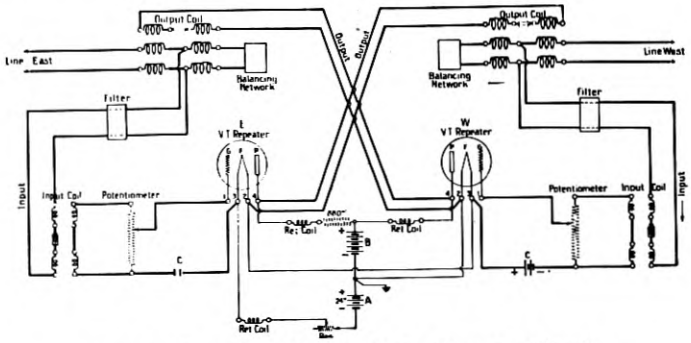


Fig. 22—Two-Way Vacuum-Tube Repeater Circuit

It will be noted that the line circuit from one direction, for instance, the one designated "line west," is connected through a three-winding transformer to a balancing network which is so made up as to balance

the line as nearly as possible at telephone frequencies. This balance is essential to proper repeater operation. The circuit arrangement is such that part of the incoming energy is diverted to that part of the circuit containing the input coil directly associated with this three-winding transformer. By the action of the vacuum-tube arrangement amplified energy is transmitted to the line east. That part of the original incoming energy from the line west that goes through the balancing network or the output coil is not, of course, transmitted along into the line east. The operation in the case of currents incoming from the line east is similar and it will be noted that the complete repeater circuit is made up of two symmetrical parts. This circuit arrangement constitutes what is known as a two-wire repeater and the apparatus is, of course, all closely associated in the same office.

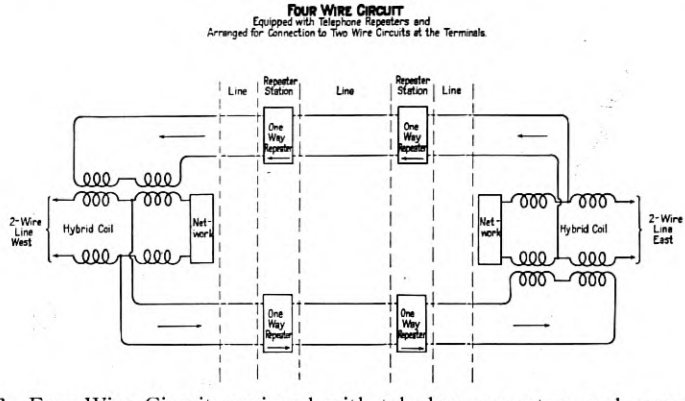


Fig. 23—Four-Wire Circuit equipped with telephone repeaters and arranged for connection to two wire-circuits at the terminals

Several of these repeaters may be inserted in tandem at appropriate points in a circuit, but there is a limit to the length of circuit that can be satisfactorily operated with this arrangement, this length depending upon the type of the facilities used. When longer circuits are required, a four-wire arrangement is used, as shown in Fig. 23. It will be noted that in this arrangement the three-winding transformers are not located in the same office but may be in offices several hundred miles apart. At each of the intermediate stations a vacuum-tube amplifier arranged for amplification in one direction only is connected to each of the two branches of the circuit. Two circuits are, of course, required between the terminals and these may be either physical or phantom circuits.

An advantage of this arrangement is that balancing networks are

not required at each repeater station and the general matter of balance and consequently good repeater operation in the circuit as a whole is greatly simplified. This arrangement can, therefore, be satisfactorily used for long circuits where two-wire operation might be impracticable,

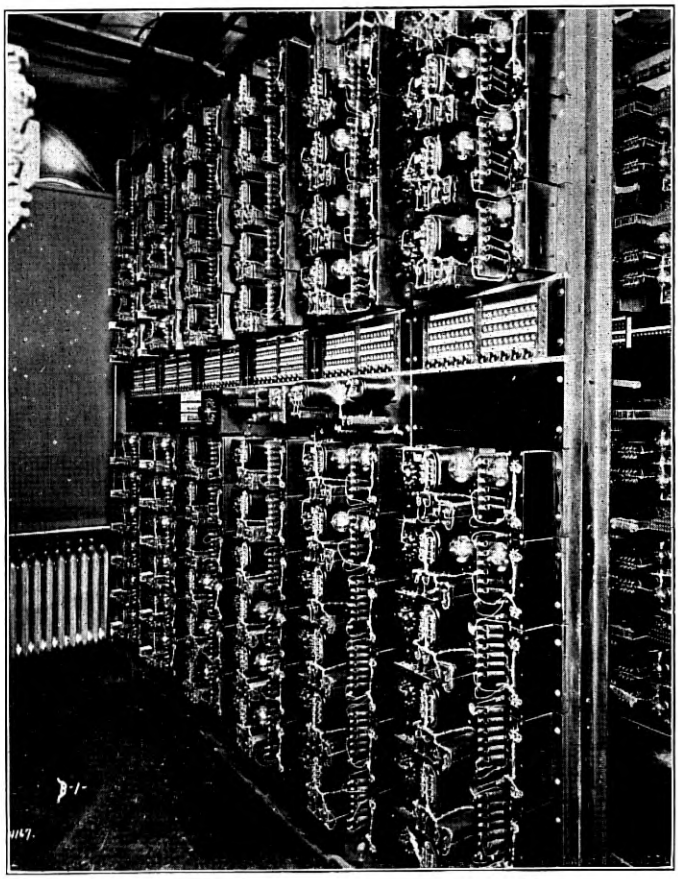


Fig. 24—Group of Repeaters at Reading, Pa.

and examples would be such circuits as New York-Pittsburgh or New York-Chicago.

Both of these types of circuits may be operated on No. 19 A. W. G. four-wire facilities which may be either physical or phantom circuits.

Fig. 24 shows a group of repeaters installed in the office at Reading, Pa., and Fig. 25 shows one of the four-wire repeater units in somewhat greater detail.

Line Impedance

In order that networks may be used to balance the lines for repeater operation, it is necessary as a practical proposition that the impedance characteristics of the lines be fairly uniform over the range of telephone frequencies. The solid line in Fig. 26 shows the resistance component of the impedance of a No. 19 loaded cable circuit with all loading coils in place. The solid line in Fig. 27 shows the

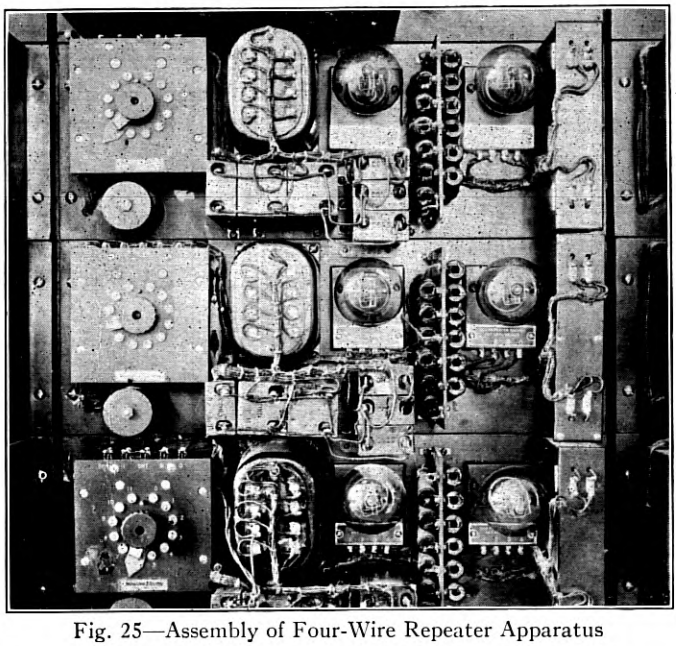


Fig. 25—Assembly of Four-Wire Repeater Apparatus

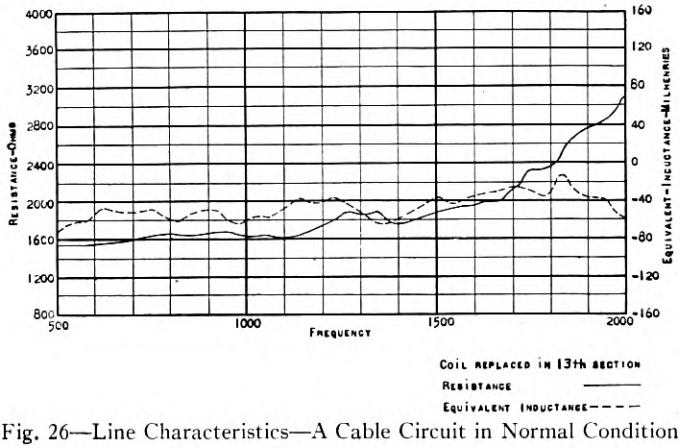


Fig. 26—Line Characteristics—A Cable Circuit in Normal Condition

resistance component found in impedance measurements on the same circuit with one coil omitted at the thirteenth loading point from the end at which the tests were made. It will be noted that in the latter case the characteristics of the circuits vary greatly with frequency. It would therefore be very difficult as a practical proposition to build up a network that would balance lines in this condition, and such variations in the electrical characteristics of a circuit impair the quality of telephone transmission, as the currents of different frequencies are differently affected. The necessity for careful maintenance work in promptly replacing loading coils which may become defective or preventing other irregularities from creeping into the plant will therefore be clear.

Transmission Regulation

The resistance of small-gage cable conductors is one of the important factors that determine the transmission losses of a circuit. The resistance of a No. 19 A. W. G. pair is about 88 ohms per mile so that

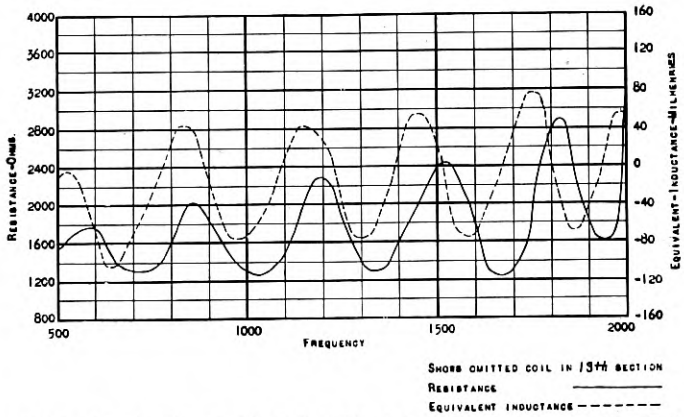


Fig. 27—Cable Circuit with Loading Coil Missing at Thirteenth Loading Point from Terminal

in a long circuit this factor of line resistance reaches considerable proportions. Now as most of the cable is aerial, the resistance of the conductors is of course affected by changes in temperature both daily and seasonal and the transmission losses vary accordingly. These changes in transmission values are of such magnitude that automatic transmission regulators are being provided for certain groups of longer circuits. All changes in the transmission equivalents of the circuits from whatever causes must be carefully watched and necessary adjustments made or the service will be seriously affected.

TELEGRAPH

In the section between Philadelphia and Pittsburgh practically all of the existing long aerial wire circuits are composited, that is, they are arranged for simultaneous telephone and telegraph operation. The telegraph circuits thus obtained are generally used in furnishing what is sometimes called "leased wire" service. The ground return system providing either full duplex or single-line operation is used and the line currents average about 75 milliamperes. This grounded telegraph system cannot be used where simultaneous telephone and telegraph service is desired on loaded cable circuits of the length involved in this cable, and as a part of the work of carrying out the comprehensive toll cable plans of the Bell system, a new telegraph system had to be developed. It was found preferable to use a metallic return circuit and to limit the line current to a value between 3 and 5 milliamperes in order to prevent serious interference to the telephone circuits due to the "flutter effect,"[2] Morse thump, and mutual interference between telegraph circuits. Morse thump results when the composite sets, that is, the apparatus used for separating the telephone and telegraph currents, do not completely prevent the latter from entering the telephone circuit, thus causing interference. The telegraph repeaters are located at about 100-mile intervals on the No. 19 circuits and at somewhat less frequent intervals on No. 16 circuits. The telegraph apparatus is of course located in the same buildings that are used to house the telephone repeaters, and on the Philadelphia-Pittsburgh cable telegraph repeaters will be located initially at Philadelphia, Harrisburg, Bedford and Pittsburgh.

TEST BOARDS

All of the conductors in the cables are carried into stations located at about 50-mile intervals and apparatus is provided in these stations for making regular tests to ascertain the condition of the cable and to locate trouble quickly. At these offices the different kinds of operating apparatus are also connected to the cable conductors; examples of this apparatus are phantom repeating coils, composite sets to permit simultaneous telephone and telegraph operation, telegraph repeaters, telephone repeaters and associated balancing equipment, signaling apparatus, and where required, the switchboards necessary for making the telephone connections involved in furnishing service. It is necessary that this apparatus which is installed in large quantities

[2] Paper by Martin and Fondiller in JOURNAL OF A. I. E. E., February, 1921.

be systematically arranged and facilities provided for making quick changes in the circuit arrangement. The circuits are wired through jacks installed in groups in test boards for this purpose and to facilitate testing. One of these boards is illustrated in Fig. 28. This particular board is located in one of the larger offices. The test boards in one of the repeater stations, such as Bedford, would consist of a smaller

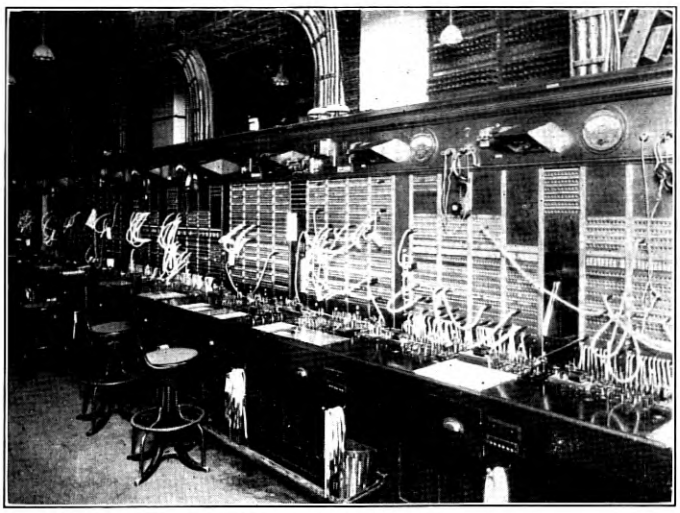


Fig. 28—Test Boards

number of positions. A position is three feet in length. In Fig. 28 each position bears a number.

Stations and Power Plants

Telephone repeaters of either the two-wire or four-wire type are connected to the circuits at approximate intervals of either 50 or 100 miles, depending upon the type of facilities which it is economical to use in the different circuits and the kind of service for which a given circuit is intended. As mentioned above, telegraph repeaters are installed at about 100-mile intervals. At some of these points existing offices are used while in a number of cases it was necessary to erect buildings for the sole purpose of housing the repeaters, testing apparatus and other equipment associated with the cable. For example, new buildings of fire-proof construction were erected at Shippensburg, Bedford and Ligonier. Fig. 29 is a view of the building at the latter point and the other two buildings are similar to this one, the dimensions being about 50 by 80 feet. Power plants are installed in these build-

ings to furnish current of the proper characteristics for operating the apparatus, and storage batteries are provided to insure uninterrupted service. As an indication of the size of these plants the 24-volt storage batteries installed for the initial load at Bedford have a capacity of 2240 ampere-hours and this provides about one day's reserve. The capacity can, of course, be increased as repeaters are added from time to time when additional circuits are needed. Storage batteries of smaller sizes supplying current at potentials of 30, 120 and 130 volts are also provided.

Examples of Circuit Arrangements

Fig. 30 shows two possible methods of building up a Philadelphia-Pittsburgh terminal circuit and Fig. 31, a method of building up a New York-Pittsburgh terminal circuit. In all three cases these telephone circuits are intended to have a transmission equivalent of about 12 miles of standard cable. Some Philadelphia-Pittsburgh

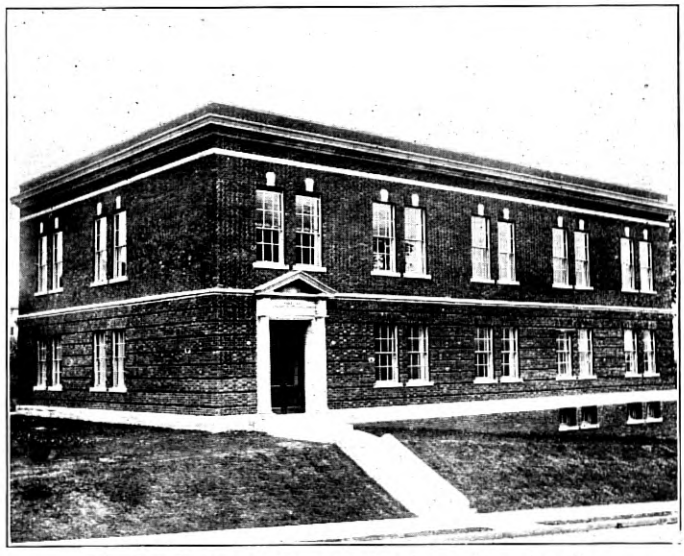


Fig. 29—Test and Repeater Station at Ligonier, Pa.

terminal circuits of the first type have been in everyday operation for several months, but it is not the most economical arrangement that it is possible to obtain for general use in providing this or similar service. It will be noted that No. 19 four-wire facilities are used between Philadelphia and Harrisburg and four-wire repeaters are located at these two points. At Harrisburg the four-wire circuit is

connected to a No. 16 two-wire circuit with a two-wire repeater at Bedford. This arrangement was used in order to start service through the cable with the facilities available, but it is intended later on to use the arrangement shown in example No. 2.

In example No. 2, No. 16 heavily loaded conductors are used and two-wire repeaters are located at Reading, Shippensburg and Ligonier. The total transmission equivalent of this circuit without repeaters is

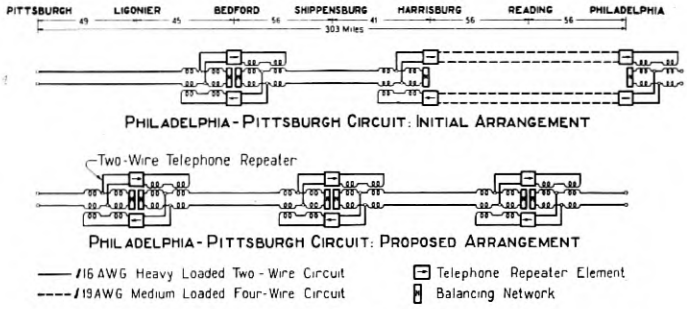


Fig. 30—Cable Circuit Arrangements

about 50 miles of standard cable so that in order to obtain a net equivalent of 12 miles for the circuit each of the three repeaters must give a transmission gain of nearly 13 miles of standard cable. This circuit could not of course be used for telephone purposes without repeaters.

The third example shows how it is expected to operate New York-Pittsburgh circuits intended for business between these two terminals.

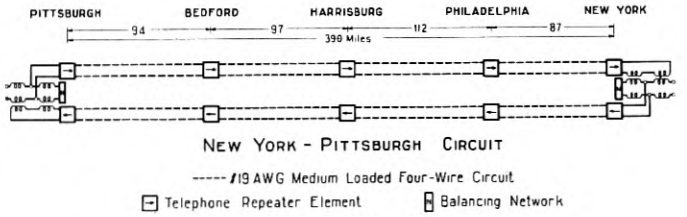


Fig. 31—Cable Circuit Arrangements

Four-wire No. 19 loaded cable facilities are used with four-wire telephone repeaters located at New York, Philadelphia, Harrisburg, Bedford and Pittsburgh.

Even with conductors of only two gages in the cable, it is clear that many different combinations of facilities can be built up into telephone circuits and an endeavor is always made to use the most eco-

nomical arrangement that will furnish the service required over each circuit group. The examples described above are of circuits used for business between the terminals indicated and if these circuits were to be connected to others extending to points considerable distances beyond these terminals different arrangements would be required. The cable conductors used in building up these telephone circuits can be composited and telegraph circuits are thus provided for simultaneous operation with the telephone circuits.

Conclusion

In the above discussion, an effort has been made to furnish some descriptive information regarding a complete cable system recently completed and now in successful operation between Philadelphia and Pittsburgh and designed for long-distance telephone and telegraph service. In one sense this discussion may be considered a report of the present status of the toll cable plant intended to connect Atlantic Seaboard cities with Chicago and other cities, and extensions are now under construction. However, most of the general methods which it is planned to use in these extensions are not expected to differ greatly from those described.

This cable system utilizes many new developments in the communication art and some of these, which have been briefly touched on here on account of their important application, have been described in more detail in previous papers. It is expected that more information regarding other specific developments which have contributed in an important way to the successful carrying out of this project or which may be utilized later on will be furnished in future papers.

An important feature of this cable project is the fact that while many new developments and practices are utilized, the design of the system as a whole is such as to fit in economically with existing wire and cable systems and proposed extensions.

Transmission Characteristics of the Submarine Cable[1]

By JOHN R. CARSON and J. J. GILBERT

SYNOPSIS: The present paper presents an extensive theoretical investigation of the impedance of the "sea return" of various types of submarine cables. In the case of the cables used for submarine telegraphy the impedance of the sea return has been practically negligible because of the low frequencies involved. For these low frequencies the cross-section of the return path is very large and its resistance low, even though the specific resistance of sea water is of the order of ten million times that of copper. As the frequency of the cable current is raised, however, the return currents crowd in nearer the cable and the resistance of the return path is increased. For frequencies in and above the telephone range, the return currents are forced into the steel armor wires around the cable and into the water just outside of the insulation. The small cross-section of the water involved and the loss in the armor wires cause the resistance of the return path to become a very large part of the total resistance of the circuit.

The present investigation led to the conclusion that the resistance of the return path could be greatly diminished by winding a low resistance conductor in the form of a copper tape immediately around the gutta percha insulation applied to the core of the cable. The concentric, cylindrical conductor thus formed lies within the armor wires but is not insulated from them and the sea water. Estimates of the sea return which would have been obtained in the Key West-Havana cable if no copper tape had been provided give values of 4, 6.5, and 8 ohms per nautical mile at 1,000, 3,000 and 5,000 cycles. The resistance actually obtained with the copper tape does not exceed 1.7 ohms at 5,000 cycles. The greater values would have increased the attenuation by approximately 30% at 1,000 cycles and by 50% at the two higher frequencies. The present cable permits of the operation of a carrier telegraph channel at 3,800 cycles, this lying above the range of telephone frequencies.

The paper gives a comparison of the theoretical conclusions with experimental data and the agreement is so satisfactory as to indicate that the theory is a reliable guide in the design of such a cable.—*Editor.*

I

THE transmission characteristics of a conducting system, such as a submarine cable circuit, are determined by its propagation constant, Γ, and characteristic impedance, K, which may be calculated for the frequency $p/2\pi$ from the formulas:

$$\Gamma = \sqrt{(R + ipL)\,(G + ipC)}, \qquad (1)$$

$$K = \sqrt{\frac{R + ipL}{G + ipC}},$$

where R, L, G and C are the four fundamental line parameters, resistance, inductance, leakance, and capacity, all per unit length. These formulas are rigorous for all types of transmission systems; but the determination of the line parameters is not always possible by elementary methods, and may indeed be a matter of considerable com-

[1] Reprinted from the *Journal of the Franklin Institute*, December, 1921.

plexity and involve rather difficult analysis. In the case of the submarine cable, exact formulas are available for calculating the capacity and leakage and the core impedance. Considerable uncertainty is introduced into the theory, however, on account of the lack of a method of determining the "return impedance," that is, the contribution of the "sea return" (sea water, armor wires, etc.) to the effective resistance and inductance of the circuit. An investigation of this problem was undertaken by the writers in connection with the research program of the American Telephone and Telegraph Company and the Western Electric Company.

The purpose of the present paper is to discuss transmission over the submarine cable, and, more particularly, to develop rigorous formulas for the calculation of the impedance of the return conductor of the cable. The results of theoretical calculations are then compared with actual experimental data; and the agreement between theory and experiment is so satisfactory as to indicate that the former is a reliable guide in the design and predetermination of the cable.

Besides providing a method for accurately calculating the transmission characteristics of a submarine cable, the present analysis leads to the following general conclusions:

(1) Contrary to usual assumption, the "sea return" impedance is by no means negligible. Even at quite moderate frequencies there is a considerable crowding of the return current into the immediate neighborhood of the cable, with a consequent rapid increase of the resistance and a corresponding decrease of the inductance of the circuit. Except at the lowest frequencies, therefore, the impedance of the "sea return" is a very important factor.

(2) The armor wires which surround the cable, and which are necessary for mechanical protection, have a very pronounced effect on the impedance of the sea return, and even at moderate frequencies may become the controlling factor. Their action is to screen the current from the sea water itself, and, as the frequency increases, to carry more and more of the return current, until it is almost entirely confined to the armor wires and excluded from the sea water.

(3) The rapid increase in the impedance of the armor wires with frequency, and their pronounced and even controlling effect on transmission makes a thorough-going study of their role in the electrical system a matter of first-class importance. Heretofore they appear to have been regarded only as a mechanical protection, and their effect on transmission has been ignored. The accurate method of calculating their impedance which is developed in the following pages is believed to have considerable value in this connection.

(4) At relatively high frequencies, the return impedance, and hence the attenuation and the distortion, may be very greatly decreased by a correctly designed thin metallic sheath concentric with the core, and in electrical contact with the armor wires. The very important action of such a sheath, even when extremely thin, does not appear to have been adequately recognized or studied. It is suggested that the introduction of such a sheath affords a means of greatly increasing the range of frequencies which the cable can transmit.

The general problem of determining the transmission characteristics of a system consisting of an insulated conductor surrounded by a concentric ring of armor wires immersed in sea water is of considerable difficulty, since in this case the propagated wave must be represented as a set of component waves centered upon or diverging from the axes of the core and of the individual armor wires. The problem was first simplified by replacing the ring of armor wires by a cylindrical sheath, thus giving circular symmetry to the structure. The analysis of this case, however, showed that the effect of the iron sheath replacing the armor wires was so pronounced as to make this simplifying assumption of doubtful validity. The general problem was therefore attacked, and rigorous methods developed for calculating the effect of the armor wires upon transmission. The results in this case differ markedly from those obtained for the case of a continuous iron sheath, which indicates that great caution must be used in making assumptions regarding the physical structure of the armoring.

The present paper follows rather closely the course of the writers' investigation. In Section II is analyzed the problem of transmission over a system consisting of n coaxial cylindrical conductors, which may be either in electrical contact at their adjacent surfaces or separated from each other by dielectric spaces. The outermost conductor, consisting of the sea water, is assumed to extend to infinity. This analysis is then applied, in Section III, to the case of a submarine cable which is armored with a continuous iron sheath. This problem is not only of interest in itself, but serves as a first approximation to the case of an actual cable, and gives a clear qualitative idea of the effect of the various factors on transmission. In Section IV the problem of the submarine cable armored with a ring of iron wires is attacked and solved by rigorous methods, and the theoretical results are then compared with experimental data.

II

The solution of the problem of transmission of periodic currents over a system comprising n coaxial cylindrical conductors consists

in finding the particular solution of Maxwell's equations which satisfies the boundary conditions—continuity of tangential electric and magnetic forces at the surfaces of the conductors. Let the common axis of the conductors coincide with the Z axis of a system of polar coordinates, R, Φ, Z, and let the electric and magnetic variables involve the common factor $exp\ (-\Gamma z + ipt)$, Γ is therefore the propagation factor characterizing transmission, and p is 2π times the frequency. This factor will not be explicitly written in any of the work that follows, but it will be assumed to be incorporated in each of the electric variables so that

$$\frac{\partial^n}{\partial z^n} = (-\Gamma)^n, \qquad \frac{\partial^n}{\partial t^n} = (ip)^n.$$

From symmetry, it is evident that the component of electric field intensity in the direction of ϕ vanishes, and that the magnetic lines of force are circles lying in planes perpendicular to the axis of the system, and centered on that axis. Also, the axial and radial electric forces are independent of ϕ. It can be shown that the radial component of electric field intensity *in the conductors* is negligibly small compared with the axial component. The latter, for a given conductor, is of the form $E\ exp(-\Gamma z + ipt)$, where E is a solution of the differential equation

$$\frac{\partial^2 E}{\partial r^2} + \frac{1}{r}\frac{\partial E}{\partial r} + (\Gamma^2 - 4\pi\lambda\mu ip)\,E = O. \tag{2}$$

Here λ and μ are the electrical conductivity and the magnetic permeability of the particular conductor, measured in absolute electromagnetic units, and E is a function of r alone.

For the frequencies in which we are interested it may be shown that $\Gamma^2/4\pi\lambda\mu p$ is exceedingly small, so that (2) may be written

$$\frac{\partial^2 E}{\partial r^2} + \frac{1}{r}\frac{\partial E}{\partial r} - 4\pi\lambda\mu ipE = O. \tag{3}$$

We will designate by the subscript j all quantities pertaining to the j^{th} conductor, counting from the axis. The solution of (3) for this conductor may then be written

$$E_j = A_j J_o(\rho_j) + B_j K_o(\rho_j), \tag{4}$$

where J_o and K_o are Bessel functions of zero order, A_j and B_j are arbitrary constants and

$$\rho_j = ri\sqrt{4\pi\lambda_j\mu_j pi} = r\alpha_j.$$

The magnetic field intensity can then be obtained from the curl law,

$$\mu \frac{dH}{dt} = \frac{dE}{dr},$$

which gives

$$H_j = \frac{\alpha_j}{\mu_j i p}\left[A_j J_o'(\rho_j) + B_j K_o'(\rho_j)\right], \tag{5}$$

where the prime indicates differentiation with respect to ρ_j. Taking the line integral of both sides of (5) around circular paths in conductor j lying close to the inner and outer surfaces of the cylinder we obtain

$$A_j J_o'(y_j) + B_j K_o'(y_j) = \frac{2\mu_j i p}{y_j}(I_1 + I_2 + \ldots + I_{j-1}), \tag{6}$$

$$A_j J_o'(x_j) + B_j K_o'(x_j) = \frac{2\mu_j i p}{x_j}(I_1 + I_2 + \ldots + I_j),$$

in which

I_j = current in the jth conductor,
$x_j = \alpha_j a_j$,
$y_j = \alpha_j b_j$,
a_j = external radius of jth conductor,
b_j = internal radius of jth conductor.

The values of the electric field intensity at the inner and outer surfaces of the j^{th} conductor can be written, from (4)

$$E_j' = A_j J_o(y_j) + B_j K_o(y_j),$$
$$E_j'' = A_j J_o(x_j) + B_j K_o(x_j).$$

Combining, in turn, each of these equations with relations (6) to eliminate A_j and B_j, we obtain

$$E_j' = Z_1' I_1 + Z_{j2}' I_2 + \ldots + Z_{jj}' I_j \tag{7}$$
$$E_j'' = Z_{j1}'' I_1 + Z_{j2}'' I_2 + \ldots + Z_{jj}'' I_j,$$

in which

$$Z_{jk}'' = 2\mu_j i p\left[\frac{1}{x_j}\frac{J_o(x_j) K_o'(y_j) - J_o'(y_j) K_o(x_j)}{J_o'(x_j) K_o'(y_j) - J'(y_j) K_o'(x_j)} - \frac{1}{y_j}\frac{J_o(x_j) K_o'(x_j) - J_o'(x_j) K_o(x_j)}{J_o'(x_j) K_o'(y_j) - J_o'(y_j) K_o'(x_j)}\right], \quad k \neq j$$

$$Z_{jj}'' = \frac{2\mu_j i p}{x_j}\left[\frac{J_o(x_j) K_o'(y_j) - J_o'(y_j) K_o(x_j)}{J_o'(x_j) K_o'(y_j) - J_o'(y_j) K_o'(x_j)}\right],$$

$$Z_{jk}' = 2\mu_j i p\left[\frac{1}{x_j}\frac{J_o(y_j) K_o'(y_j) - J_o'(y_j) K_o(y_j)}{J_o'(x_j) K_o'(y_j) - J_o'(y_j) K_o'(x_j)} - \frac{1}{y_j}\frac{J_o(y_j) K_o'(x_j) - J_o'(x_j) K_o(y_j)}{J_o'(x_j) K_o'(y_j) - J_o'(y_j) K_o'(x_j)}\right], \quad k \neq j \tag{8}$$

$$Z_{jj}' = \frac{2\mu_j i p}{x_j}\left[\frac{J_o(y_j) K_o'(y_j) - J_o'(y_j) K_o(y_j)}{J'_o(x_j) K'_o(y_j) - J_o'(y_j) K_o'(x_j)}\right].$$

We have now succeeded in expressing the electric forces in the conductors as linear functions of the currents $I_1 \ldots I_n$, the coefficients being of the nature of impedances, by a method which is simply an application of the principle of continuity of magnetic field intensity. The remaining boundary condition, continuity of the tangential component of electrical field intensity gives, where two consecutive cylinders are in electrical contact,

$$E'_{j+1} - E''_j = \left[Z'_{j+1,1} - Z''_{j1} \right] I_1 + \ldots + \left[Z'_{j+1,j} - Z''_{jj} \right] I_j + Z'_{j+1,j+1} I_{j+1} = 0. \tag{9}$$

This gives m relations between the n currents of the system, m being the number of contacts between successive cylinders. In the case where the j and $(j+1)$st conductors are separated by a layer of dielectric material, a relation between the boundary values of electric field intensity may be obtained as follows:

If E_r is the radial electric field intensity in the dielectric, then

$$V_{jj} = \int_{aj}^{bj+1} E_r dr$$

is the potential difference between the j and $(j+1)$st conductors, in the sense employed in ordinary circuit theory. If we now apply the law

$$\text{curl } E = -\mu \frac{dH}{dt}$$

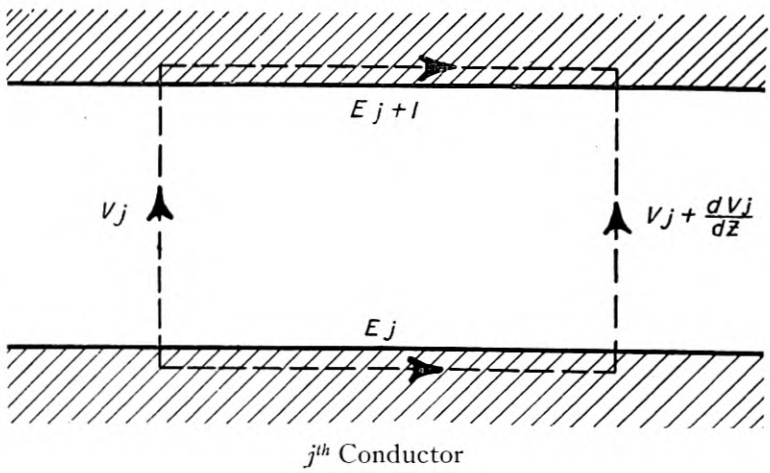


FIG. 1

to the elementary contour shown in Fig 1 we get

$$-\frac{\partial V_j}{\partial z} + E'_{j+1} - E''_j = \mu i p \Phi_j, \tag{10}$$

or

$$\Gamma V_j + E'_{j+1} - E''_j = \mu i p \Phi \tag{11}$$

where Φ_j is the magnetic flux threading the contour and is given by

$$\Phi_j = 2\,(I_1 + I_2 + \ldots + I_j) \log \frac{b_{j+1}}{a_j}.$$

From the law

$$\text{div } kE = 4\pi Q,$$

$$E_r = \frac{2}{k_j r}(Q_1 + Q_2 + \ldots + Q_j)$$

where Q_j is the charge on the j^{th} conductor and k_j is the dielectric constant of the medium, whence,

$$V_j = \frac{2}{k_j} \log \frac{b_{j+1}}{a_j}\,(Q_1 + Q_2 \ldots + Q_j). \tag{12}$$

Furthermore, the rate of gain of charge is

$$\frac{\partial}{\partial t}(Q_1 + Q_2 + \ldots + Q_j) = -\frac{\partial}{\partial z}(I_1 + I_2 + \ldots + I_j) - 4\pi(Q_1 + Q_2 + \ldots + Q_j)\, g_j/k_j, \tag{13}$$

where the last term represents the leakage current, g_j being the specific conductivity of the dielectric.

From (13) we have

$$\frac{1}{k_j}(4\pi g_j + ipk_j)\,(Q_1 + Q_2 + \ldots + Q_j) = \Gamma\,(I_1 + I_2 + \ldots + I_j,)$$

and substituting this value of $(Q_1 + Q_2 + \ldots + Q_j)$ in (12) gives

$$V_j = 2\,(I_1 + I_2 + \ldots + I_j)\,\frac{\Gamma}{4\pi g_j + ipk_j} \log \frac{b_{j+1}}{a_j} \tag{14}$$

and from this and (11)

$$-\left[\frac{\Gamma^2}{G_j + ipC_j} - ipL_j\right](I_1 + I_2 + \ldots + I_j) = E_j'_{+1} - E''_j \tag{15}$$

where

$$G_j = \frac{4\pi g_j}{2 \log \frac{b_{j+1}}{a_j}}, \quad C_j = \frac{k_j}{2 \log \frac{b_{j+1}}{a_j}}, \quad L_j = 2\mu_j \log \frac{b_{j+1}}{a_j}$$

Substituting the values of E_j'' and E_{j+1}' from (7) in (15) gives

$$-\left[\frac{\Gamma^2}{G_j + ipC_j} - ip\,L_j\right](I_1 + I_2 + \ldots + I_j) = (Z_{j+1,1}' - Z_{j,1}'')\,I_1 + \ldots - Z_{j+1,j+1}' I_{j+1}. \tag{16}$$

An equation of this sort may be obtained for each layer of dielectric and these combined with equations (9) and the condition that the electric field intensity in the sea water must vanish at infinity,

$$E_n' = Z_{n1}''\,I_1 + \ldots + Z_{nn}''\,I_n = O,$$

give n relations between $I_1 \ldots I_n$. In order that these shall be consistent, the determinant of the coefficients must vanish.

$$\begin{vmatrix} Z_{21}' - Z_{11}'', & Z_{22}', & 0 & - - - 0 \\ Z_{31}' - Z_{21}'', & Z_{32}' - Z_{22}'', & Z_{33}', & - - - 0 \\ - & - & - & - - - - \\ Z_{j+1,1}' - Z_{j,1}'' + Z_j, & Z_{j+1,2}' - Z_{j,2}'' + Z_j, & - & - - - - \\ - & - & - & - - - - \\ Z_{n1}'', & Z_{n2}'', & - & - - - Z_{nn}'' \end{vmatrix} = O \tag{17}$$

where

$$Z_j = \frac{\Gamma^2}{G_j + ipC_j} - ipL_j.$$

This is an equation in Γ^2 of degree equal to the number of dielectric layers; consequently, there are as many independent modes of propagation in the system as there are branches in the network of conductors.

From this point the method of determining the behavior of the system depends upon conditions in the particular problem. For the case where there are k dielectric layers separating the conductors into $k + 1$ groups the current on the j^{th} group may be written in the form

$$I_j = A_{j1}\,exp\,(-\Gamma_1 z + ipt) + \ldots + A_{jk}\,exp\,(-\Gamma_k z + ipt) + B_{j1}\,exp\,(\Gamma_1 z + ipt) + \ldots + B_{jk}\,exp\,(\Gamma_k z + ipt),$$

where $\Gamma^2_1 \ldots \Gamma^2_k$ are the k roots of the determinant (17) and $A_{j1} \ldots A_{jk}$, $B_{j1} \ldots B_{jk}$ are constants. These constants are not all in-

dependent, however, since, for each value of Γ, Γ_1 for instance, there exist k relations of the form (16) which the corresponding set of constants A_{11}, $A_{21}, \ldots A_{k1}$ must satisfy. The remaining $2k$ independent constants can then be determined from a knowledge of the conditions at the terminals of the conductors.

It is important to observe that the transmission characteristics of a system of coaxial conductors are influenced to a great extent by the manner of connecting the various members of the system. Anomalies in the impedance of a complicated network such as a submarine cable with several conducting sheaths in the return path, may often be traced to lack of proper connections between the sheaths, or to faulty joints.

III

The submarine cable armored with a continuous coaxial sheath, as shown in Fig. 2, is a particular case of the foregoing, and one which presents a clearer idea of the physical significance of the various steps in the general theory. There are only two groups of conductors, the

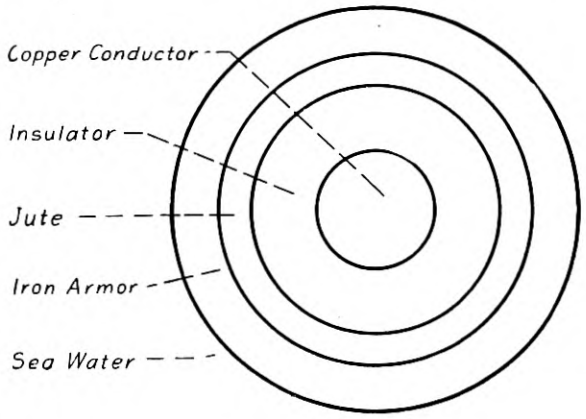


FIG. 2

first consisting of the core conductor, and the second comprising the iron sheath and the sea water, the two groups being separated by the insulating material and the layer of jute. Consequently, there is only one mode of propagation, and the analysis is considerably simplified.

The jute is assumed to contain sufficient sea water so that although it conducts practically no current axially, it maintains equality of potential between the outer surface of the gutta percha and the inner

surface of the iron sheath. Consequently equation (10) may be written

$$\frac{\partial V}{\partial z} - E_2' + E_1'' = - \mu i p \Phi = - i p L_{12} I_1, \tag{18}$$

where E''_1 and E'_2 are the values of electric field intensity at the outer surface of the core conductor and the inner surface of the iron, respectively, V is the potential difference between these two surfaces, and Φ is the magnetic flux threading unit length of the gutta percha and jute. Also, from (14)

$$- \frac{\partial V}{\partial z} = \frac{\Gamma^2}{G + ipC} I_1 \tag{19}$$

in which I_1 is the current in the core and

$$G = \frac{4\pi g_{12}}{2 \log \frac{b}{a_1}}, \qquad C = \frac{k_{12}}{2 \log \frac{b}{a_1}}, \tag{20}$$

where g_{12} and k_{12} are the electrical constants of the gutta percha, and b is the external radius of the core. It is evident, that G and C are respectively the leakage and capacity of unit length of the cable. Therefore, from (1),

$$\frac{\Gamma^2}{G + ipC} = R + ipL = Z, \tag{21}$$

where R and L are the resistance and inductance of unit length of the cable, including the sea return. Equation (18) may then be written

$$Z I_1 = E_1'' - E_2' + ipL_{12} I_1. \tag{22}$$

To determine Z we must express E''_1 and E'_2 as functions of I_1.

We have seen that

$$E_1'' = Z_1 I_1, \tag{23}$$

where Z_1 may be termed the "internal impedance" per unit length of this conductor. In fact, when we place $y_1 = o$ in (8) we obtain

$$Z_{11}'' = \frac{2\mu_1 ip}{x_1} \frac{J_o(x_1)}{J_o'(x_1)}, \tag{24}$$

which is the usual formula for the internal impedance of a cylindrical conductor.

Similarly

$$E_2' = - Z_2 I_1 \tag{25}$$

where Z_2 is the internal impedance of the return conductor, the

minus sign being due to the fact that the current in the return is in the negative direction of z.

Inserting (23) and (25) in (22) gives

$$Z = Z_1 + Z_2 + ipL_{12}.$$

The quantity Z_2 may be determined in the following manner. From (7) we have

$$E_2' = Z_{21}' I_1 + Z_{22}' I_2, \tag{26}$$

where I_2 is the current in the iron sheath. The value of this current can be found by applying the condition of continuity of electric field intensity at the common surface of the iron and the sea water, as in equation (9). This gives

$$Z_{21}'' I_1 + Z_{22}'' I_2 = Z_{31}' I_1 + Z_{32}' I_2 + Z_{33} I_3,$$

in which I_3 is the current in the sea water. From (8) it can be seen that $Z_{33} = O$, since $x_3 = \infty$, therefore

$$I_2 = \frac{Z_{31}' - Z_{21}''}{Z_{22}'' - Z_{32}'} I_1. \tag{27}$$

Substituting (27) in (26) gives

$$E_2' = \left[Z_{21}' + \frac{Z_{31}' - Z_{21}''}{Z_{22}'' - Z_{32}'} Z_{22}' \right] I_1,$$

and by comparison with (25) we have

$$Z_2 = - Z_{21}' - \frac{Z_{31}' - Z_{21}''}{Z_{22}'' - Z_{32}'} Z_{22}' \tag{28}$$

as the internal impedance of the return conductor. The resistance and reactance per unit length of this portion of the circuit are then represented by the real and imaginary parts of (28) respectively.

We may then determine R and L from the formula

$$Z = R + ipL = Z_1 + Z_2 + ipL_{12}, \tag{29}$$

where Z_1 and Z_2 are calculated from (23) and (28) and

$$L_{12} = 2 \log \frac{b_2}{a_1},$$

b_2 and a_1 being the inner radius of the iron and the outer radius of the core conductor, respectively.

For purposes of comparison, the return impedance is calculated for the case where the iron armoring is absent, the return current

being conducted by the sea water alone. As in the preceding case,

$$Z_1 = \frac{2\mu_1 ip}{x_1} \frac{J_o(x_1)}{J_o'(x_1)}$$

The expression for Z_2 simplifies considerably. The electric field intensity in the sea water may be written, from (4),

$$E_2 = B_2 K_o(\rho_2), \tag{30}$$

the term in J_o being absent in order to permit E_2 to vanish at infinity.

Also, from (6),

$$B_2 K_o'(y_2) = \frac{2\mu_2 ip}{y_2} I_1. \tag{31}$$

From (30) and (31) we have

$$E_2' = \frac{2\mu_2 ip}{y_2} \frac{K_o(y_2)}{K_o'(y_2)} I_1. \tag{32}$$

from which the return impedance can be written,

$$Z_2 = -\frac{2\mu_2 ip}{y_2} \frac{K_o(y_2)}{K_o'(y_2)} \tag{33}$$

We have then

$$Z = R + ipL = \frac{2\mu_1 ip}{x_1} \frac{J_o(x_1)}{J_o'(x_1)} - \frac{2\mu_2 ip}{y_2} \frac{K_o(y_2)}{K_o'(y_2)} + ipL_{12}.$$

The resistance and inductance of the sea return of a submarine cable were calculated from formula (28), employing the following values for the constants:

Copper
$$\begin{cases} a_1 = .226 \text{ cm.} \\ b_1 = 0 \\ \mu_1 = 1 \\ \lambda_1 = 6.06 \times 10^{-4} \end{cases}$$

Iron
$$\begin{cases} a_2 = .990 \text{ cm.} \\ b_2 = .737 \text{ cm.} \\ \mu_2 = 100 \\ \lambda_2 = 8 \times 10^{-5} \end{cases}$$

Sea Water
$$\begin{cases} a_3 = \infty \\ b_3 = .990 \text{ cm.} \\ \mu_3 = 1 \\ \lambda_3 = 5 \times 10^{-11} \end{cases}$$

The armoring was then assumed to be replaced by sea water, and the resistance and inductance of the cable were calculated from (33).

The results of the calculations are shown in the curves of Fig. 3.

It is evident from these curves that the effect of the iron armoring is to increase considerably the impedance of the return path. The

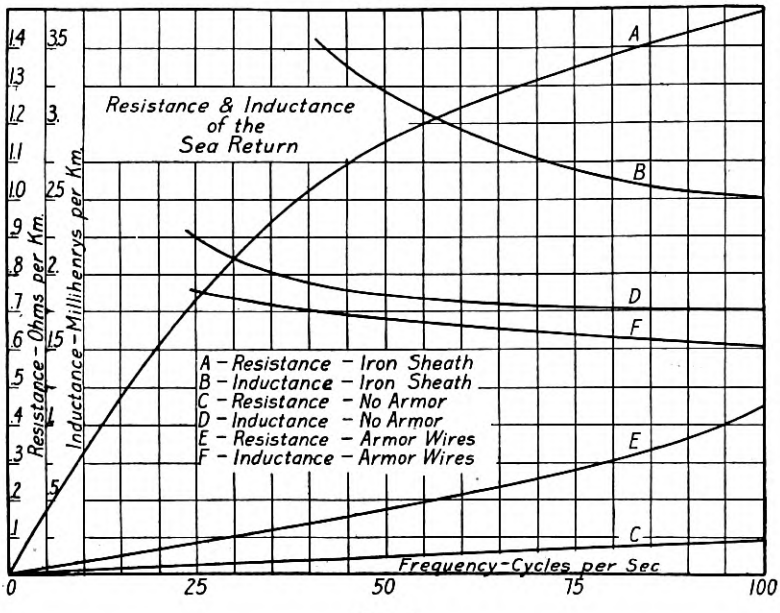


FIG. 3

physical explanation of this fact is that the iron acts as a shield to screen from the sea water the electromagnetic effects of the current flowing in the cable conductor. Energy is dissipated in the armoring and is prevented from spreading out through the surrounding medium. The assumption that the armor wires could be replaced by a solid cylinder of iron is, therefore, subject to question, since it is possible that the larger surface area of the assemblage of armor wires, and the gaps between these wires may be effective in diminishing the energy dissipated in the armoring and consequently diminishing the screening effect. This problem is investigated in the following section.

IV

The physical system under consideration is shown schematically in cross-section in Fig. 4, and consists of an insulated conductor and

protective covering of jute, surrounded by a ring of N armor wires immersed in sea water. The method of solution is essentially similar to that given in the preceding pages, and consists in determining the values of electric field intensity at the outer surface of the core conductor and the inner surface of the return conductor, from which the internal impedances of the two conductors can be found.

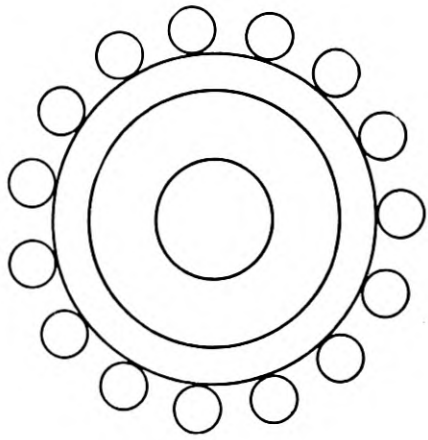


FIG. 4

The main difficulty in the analysis is caused by the lack of uniaxial symmetry in the return conductor. This was overcome by employing a method developed by one of the authors[2] in a study of transmission in parallel wires.

The electric field intensity in the sea water satisfies the differential equation

$$\frac{\partial^2 E}{\partial r^2} + \frac{1}{r}\frac{\partial E}{\partial r} + \frac{1}{r^2}\frac{\partial^2 E}{\partial \phi^2} - 4\pi\lambda\mu p i E = 0,$$

the solution of which is a Fourier-Bessel expansion,

$$E = A_0 K_0(r\alpha) + A_1 K_1(r\alpha)\cos\phi + A_2 K_2(r\alpha)\cos 2\phi + \ldots +,$$

r and ϕ being referred to the axis of the particular wire.

Assuming that the current distribution in the core conductor is independent of the angle ϕ, that is, neglecting the individual character of the armor wires only in their effect on the current distribution in the core, the effect due to the current in the core is represented

[2] "Wave Propagation over Parallel Wires; The Proximity Effect." John R. Carson, *Phil. Mag.*, vol. xli, p. 607 (1921).

by the first term of such a series, and the total field intensity may be written

$$E = A\,K_o(r\alpha) + \sum_{j=o}^{N-1} \sum_{s=o}^{\infty} B_s K_s(\alpha\rho_j) \cos s\phi_j, \qquad (34)$$

ρ_j and ϕ_j being referred to the axis of wire j, as shown in Fig. 5. That is, the resultant field is expressible as a set of waves centered on the axis of the cable and the axes of the N armor wires.

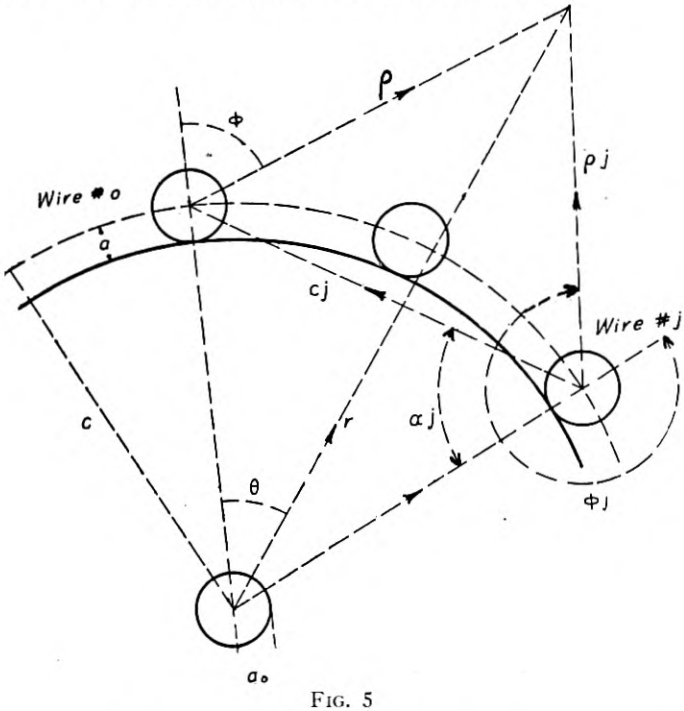


FIG. 5

In the neighborhood of the armor wires the arguments of the Bessel functions are sufficiently small [3] to permit of the approximations

$$K_o(\alpha\rho) = K - \log \rho,$$

where

$$K = 0.11593 \log \frac{1}{\alpha},$$

and

$$K_s(\alpha\rho) = \frac{1}{(-\alpha\rho)^s}.$$

[3] See Note I at end of paper.

The series (34) can, therefore, be written

$$E = A(K - \log r) + B_o(NK - \sum_{j=o}^{N-1} \log \rho_j) + \sum_{i=o}^{N-1} \sum_{s=1}^{\infty} B_s \frac{\cos(s\phi_j)}{\rho_j^s}, \quad (35)$$

in which B_s has absorbed the constant quantities. From this, the magnetic intensity in the sea water can be obtained by differentiation.

Inside any armor wire, at the surface, the field intensities are

$$E = C_o J_o(\xi) + C_1 J_1(\xi) \cos\phi + \ldots + C_n J_n(\xi) \cos n\phi + \ldots +, \quad (36)$$

$$H_\Phi = \frac{1}{a\mu i p}\left[C_o J_o'(\xi) + C_1 J_1'(\xi)\cos\phi + \ldots + C_n J_n'(\xi)\cos n\phi + \ldots + \right] \quad (37)$$

where $\xi = ai\sqrt{4\pi\lambda\mu p i}$,

λ and μ being the electrical conductivity and the magnetic permeability, respectively, of the material of the armor wire. The quantities a and ϕ are centered on the axis of the wire.

In order to determine the coefficients A, B_o, B_1, $-$, C_o, C_1, $-$ we make use of the fact that the electric and the magnetic field intensities are continuous at the surface of the wire. It is obvious, however, that nothing can be learned by equating (35) and (36) since they are formally dissimilar. We therefore transform [4] the various terms of (35) to a common axis which coincides with the axis of one of the armor wires, hereafter called wire "zero," and the electric field intensity in the sea water, close to the surface of the armor wire, is

$$\begin{aligned} E = {} & (A + NB_o)K - A\log c - B_o \log(ac_1c_2 \ldots c_{n-1}) - \Sigma_o \\ & + \left[q_1/\zeta - \zeta(A + S_{11}B_o) + \frac{\zeta}{1'!}\Sigma_1 \right] \cos\phi \\ & + \left[q_2/\zeta^2 + \frac{\zeta^2}{2}(A + S_{22}B_o) + \frac{\zeta^2}{2'!}\Sigma_2 \right] \cos 2\phi \qquad (38) \\ & \qquad - - - - - - - - - \\ & + \left[q_n/\zeta^n + \frac{(-\zeta)^n}{n}(A + S_{nn}B_o) - \frac{(-\zeta)^n}{n'!}\Sigma_n \right] \cos n\phi, \end{aligned}$$

where

$$\begin{aligned} \Sigma_0 &= S_{11}q_1 - S_{22}q_2 + S_{33}q_3 \ldots, \\ \Sigma_1 &= S_{02}q_1 - 2\,S_{13}q_2 + 3\,S_{24}q_3 \ldots, \qquad (39) \\ \Sigma_2 &= 1.2\,S_{13}q_1 - 2.3\,S_{04}q_2 + 3.4\,S_{15}q_3 \ldots, \\ \Sigma_3 &= 1.2.3\,S_{24}q_1 - 2.3.4\,S_{15}q_2 + 3.4.5\,S_{06}q_3 \ldots, \\ & \qquad - - - - - - - - \end{aligned}$$

[4] See Note II.

$$S_{pq} = \sum_{j=1}^{N-1} \frac{\cos pa_j}{\left(2 \sin \frac{j\pi}{n}\right)^q},$$
$$q_n = B_n/c^n, \tag{40}$$
$$\zeta = \frac{a}{c}.$$

The quantities ϕ and a have the same significance as in equations (36) and (37).

The tangential magnetic field intensity in the sea water at the surface of wire " zero " is, therefore,

$$H_\Phi = \frac{1}{ip}\frac{dE_z}{da} = \frac{1}{ipa}\Bigg[-B_o + \cos\phi\left[-q_1/\zeta - \zeta(A + S_{11}B_o) + \frac{\zeta}{1!}\Sigma_1 \right] \tag{41}$$
$$+ 2\cos 2\phi\left[-q_2/\zeta^2 + \frac{\zeta^2}{2}(A + S_{22}B_0) - \frac{\zeta^2}{2!}\Sigma_2 \right]$$
$$- - - - - - - - - -$$
$$+ n\cos n\phi\left[-q_n/\zeta^n + \frac{(-\zeta)^n}{n}(A + S_{nn}B_o) - \frac{(-\zeta)^n}{n!}\Sigma_n \right] - - - \Bigg]$$

To satisfy the condition of continuity of electric and magnetic field intensities at the surface of the armor wire it is necessary that the coefficients of the corresponding terms of (36) and (38) and of (37) and (41) be equal. This gives

$$C_o J_o(\xi) = (A + NB_o)K - A\log c - B_o \log(ac_1 \ldots c_n) - \Sigma_o, \tag{42}$$

$$C_n J_n(\xi) = q_n/\zeta^n + \frac{(-\zeta)^n}{n}(A + S_{nn}B_o) - \frac{(-\zeta)^n}{n!}\Sigma_n,\ n = 1, 2, \ldots \infty \tag{43}$$

$$\xi J'_o(\xi)\, C_o = -\mu B_o, \tag{44}$$

$$\xi J_n'(\xi)\, C_n = -n\mu\left[q_n/\zeta^n - \frac{(-\zeta)^n}{n}(A + S_{nn}B_o) - \frac{(-\zeta)^n}{n!}\Sigma_n \right].\ n = 1, 2, \ldots \infty \tag{45}$$

From these expressions the quantities $B_1 \ldots C_1 \ldots$ can be determined. Multiplying (43) by $n\mu$ and subtracting (45) gives

$$C_n = \frac{2n\mu q_n}{\zeta^n}\frac{1}{n\mu J_n(\xi) - \xi J_n'(\xi)}, \tag{46}$$

which expresses C_n in terms of q_n. Multiplying (43) by $\zeta J_n'(\zeta)$ and (45) by $J_n(\zeta)$, and subtracting gives

$$q_n = (-1)^n \lambda_n \xi^{2n} \left[\frac{1}{n} (A + S_{nn} B_o) - \frac{1}{n!} \Sigma_n \right], \quad n = 1, 2, \ldots \infty \tag{47}$$

where

$$\lambda_n = \frac{n\mu J_n(\xi) - \xi J_n'(\xi)}{n\mu J_n(\xi) + \xi J_n'(\xi)}. \tag{48}$$

From the infinite set of simultaneous equations (47) the infinitely many variables q_n may be determined in terms of A and B_o.[5]

We have thus determined the arbitrary constants $C_o \ldots C_n$ and $q_1 \ldots q_n$ (or $B_1 \ldots B_n$) as functions of A and B_o. It remains to express the latter quantities in terms of physical quantities. If I_1 is the current in the armor then $\frac{I_1}{N}$ is the current in a single wire. Integrating (41) completely around the armor wire "zero" gives, therefore,

$$2pi \frac{I_1}{N} = - B_o. \tag{49}$$

Similarly, if I_o is the current in the core conductor, we find

$$2piI_o = - A. \tag{50}$$

We can, therefore, express all the arbitrary constants as linear, homogeneous functions of I_o and I_1.

To determine the relation between these currents, we have from (49) and (44),

$$CJ_o(\xi) = \frac{ZI_1}{N}, \tag{51}$$

where

$$Z = \frac{2\mu ip}{\xi} \frac{J_o(\xi)}{J_o'(\xi)}.$$

Substituting (49), (50) and (51) in (42) gives

$$\frac{Z}{N} I_1 = -2ip(I_o + I_1) K + 2ipI_o \log c + 2ip \frac{I_1}{N} \log(ac_1 \ldots c_n) \tag{52}$$

$$- (S_{11} q_1 - S_{22} q_2 + S_{33} q_3 - \ldots),$$

from which, since $q_1 \ldots q_n$ are functions of I_1 and I_o, the ratio I_o/I_1 can be obtained.

[5] See Note III.

Having shown that the constants A, $B_o \ldots$ of the series (35) are proportional to I_o, we can express the electric field intensity at the inner surface of the return conductor in the form

$$E_2 = - Z_2 I_o.$$

The computation of Z_2 is facilitated by transforming the terms of (35) to the axis of the core conductor [6] and placing $r = c - a$. We thus obtain

$$E_2 = -Z_2 I_o = (A + NB_o)K - A\log(c-a) - NB_o \log c - N(q_1 - q_2 + q_3 \ldots) + (\text{terms containing } \cos\theta, \cos 2\theta, \text{ etc., as factors}). \quad (53)$$

We have, by applying the curl law to an elementary contour which links the core conductor and the return,

$$\frac{\partial V}{\partial z} - E_1 + E_2 = - ip\Phi_{12}, \quad (54)$$

where

$$E_1 = Z_1 I_o = \frac{2\mu_o ip}{\xi_0} \frac{J_o(\xi_o)}{J_o'(\xi_o)} I_o,$$

$$\Phi_{12} = L_{12} I_o = 2 I_o \log \frac{c-a}{a_o}, \quad (55)$$

and

$$\xi_o = a_o i \sqrt{4\pi\lambda_o \mu_o ip},$$

λ_o and μ_o being the electrical constants of the core conductor and a_o its radius. The value given above for Φ_{12} holds only for the contour on which E_2 is independent of the angle θ, that is, when the terms of (53) that contain $\cos\theta$, $\cos 2\theta$, etc., vanish. The value of Z_2 to be used in (54) is therefore determined from

$$E_2 = - Z_2 I_o = (A + NB_o) K - A \log (c - a) - NB_o \log c - N(q_1 - q_2 + \ldots) \quad (56)$$

As before,

$$- \frac{\partial V}{\partial z} = (R + ipL) I_o, \quad (57)$$

where R and L are the resistance and inductance per unit length of the cable, including the sea return.

We have then from (54),

$$R + ipL = Z_1 + Z_2 + ipL_{12}, \quad (58)$$

from which R and L can be determined.

[6] See Note II.

The process of calculating the resistance and inductance of a submarine cable by the method just described may be summarized as follows:

(1) Determine from (47) the quantities $q_1 \ldots q_n$ in terms of A and B_o, and then in terms of I_1 and I_o by (49) and (50).

(2) Substitute these values of $q_1 \ldots q_n$ in (52) and obtain the ratio I_o/I_1.

(3) Substitute for A, B_o and $q_1 \ldots q_n$ in (56) their values in terms of I_o and I_1.

(4) Eliminate I_1 from these two relations, thus obtaining E_2 in terms of I_o. Then $Z_2 = -E_2/I_o$.

(5) Substitute this value of Z_2 and the value of Z_1 calculated from (55) in equation (58).

(6) The resistance and the inductance per unit length of the cable may then be determined from the real and imaginary parts of the latter equation.

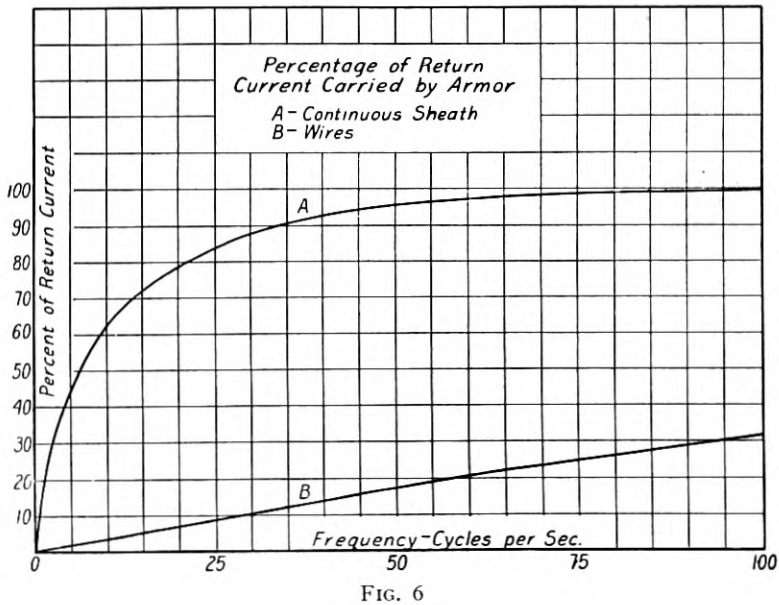


FIG. 6

The resistance and inductance of a cable of cross-section shown in Fig. 4 were computed by the method just described, the results being given by curves E and F of Fig. 3. The cable in this case is identical with that shown in Fig. 2 previously described, except that the continuous iron sheath has been replaced by fifteen wires. The

effect of the presence of the iron upon the resistance of the return conductor is still noticeable, although it is much less than in the case of the continuous iron sheath. The reason for this is evident after inspection of the curves of Fig. 6, which show the percentage of return current carried by the armor in the two cases. Especially at the lower frequencies, the return current is much more confined by the continuous sheath than it is by the wires.

As a check of the method, the resistance and inductance of the Seattle-Sitka cable of the United States Signal Corps were calculated for frequencies in the range 50 to 600 cycles per second, and

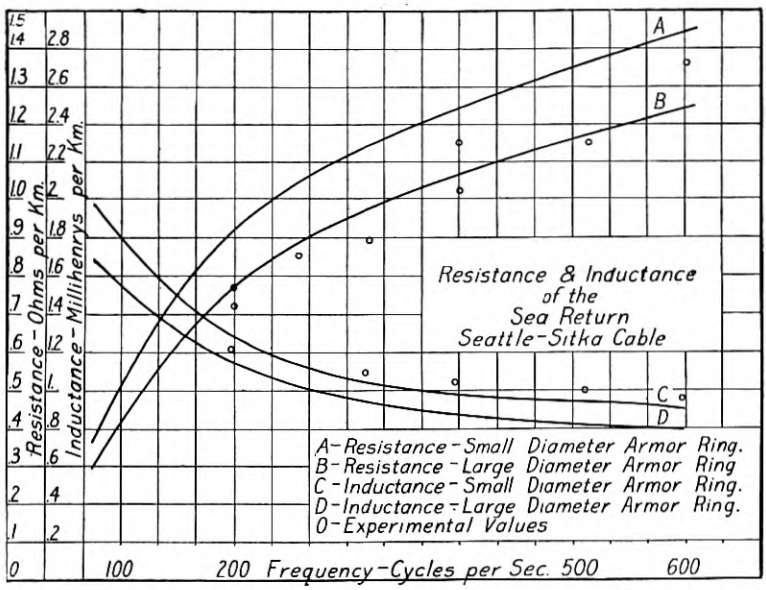


FIG. 7

the values so obtained were then compared with the results of measurements recently made upon this cable.[7] The constants used in the calculations were as follows:

Conductor	
Diameter	.216 cm.
Resistance per nautical mile	9 ohms
Rubber Insulation	
Outside diameter	.718 cm.
Capacity per nautical mile	.38 mf.
Armoring	
16 wires	each .242 cm. diameter
Outside Diameter of Cable	2.06 cm.

[7] "The Use of Alternating Currents for Submarine Cable Transmission," Frederick E. Pernot, *Jour. of the Franklin Institute*, vol. 190, p. 323, 1920.

Owing to lack of information concerning the mean radius of the ring of armor wires, two sets of data were computed employing the values $c = 0.6148$ and $c = 0.920$, which correspond, respectively, to zero and maximum separation of the armor wires.

The results of the calculations are shown in Fig. 7. The experimental values are indicated by small circles, and agree well with the theoretical values throughout the range of frequencies. The re-

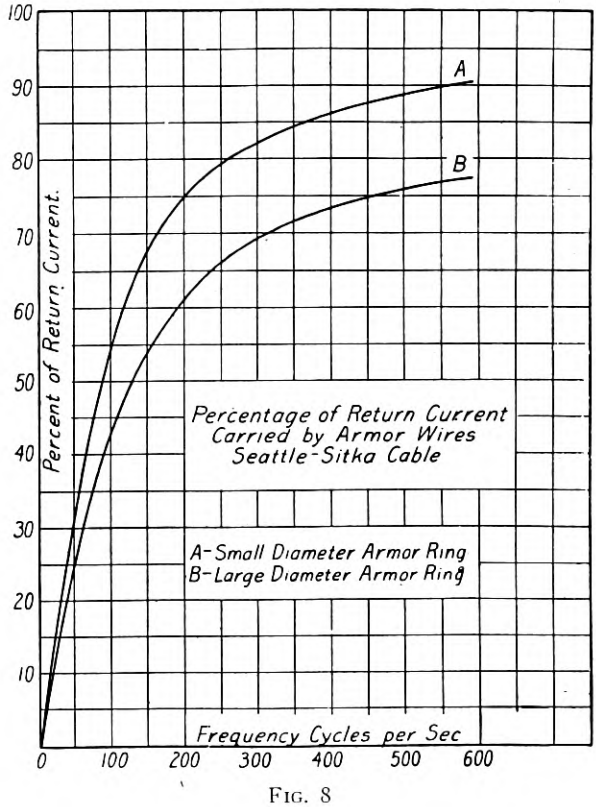


FIG. 8

sistance of the sea return increases most rapidly in the region of frequencies used in ordinary telegraphy, 0 to 100 cycles per second. In this range the inductance of the cable also has its greatest values, and these two effects have considerable influence in determining the transmission characteristics of the cable.

The percentage of the return current that is carried by the armor wires is shown in Fig. 8.

Conclusions

As was previously pointed out, the effect of the shielding action of the iron armor of a submarine cable is to diminish the electromagnetic field which is propagated through the sea water, and which gives rise to the return current. Combined with this effect is the shielding action of the sea water adjacent to the cable, upon the distant portions. The total shielding effect increases with the frequency until a point is reached where practically the whole of the return current is carried by the armor wires.

Several remedies have been suggested for diminishing the damping effect of the armor wires. It can be proved, for example, that for a given size of core and weight of armor, the number and size of armor wires can be chosen so as to give a minimum value of return impedance. A proper choice of the electrical constants of the material of which the armor is constructed would also be of advantage, since the return impedance is somewhat larger for iron than it is for material of higher or lower conductivity.

Another method of diminishing the return impedance, which has been used in practice, is to wrap the cable core with a number of concentric layers of conducting tape before it is covered with jute. The return current, as it crowds in toward the core with increasing frequency, will then have a path of comparatively low impedance, and at the higher frequencies only a small portion of the current will be carried by the armor wires and the sea water. The impedance of the return path can be calculated for this case by the methods given in the preceding pages. The following table compares the values of the resistance of the return conductor calculated by three different methods, and determined experimentally, for a cable provided with a brass tape 5 mils in thickness.

Resistance of Return Conductor—OHMS per Statute Mile

Frequency Cycles per Sec.	Approximate Method	Approx. Method[8] Corrected by Factor $\frac{2}{\pi}$	Exact Method	Experimental
3,000	4.00	3.15	2.87	2.92
10,000	4.90	4.25	4.45	4.60

The experimental values are the results of a series of measurements made by the Department of Development and Research of

[8] This is an empirical formula which has been found to be fairly close in most cases. The correction factor suggested itself in that it takes care of the increased surface of the armor wires, as compared with the corresponding continuous sheath.

the American Telephone and Telegraph Company upon the Victoria-Vancouver submarine cable. The calculated values were obtained by both the approximate and the exact methods, discussed in the preceding pages, in which the armor of the cable is treated, respectively, as a continuous sheath and as a ring of wires. The modifications which must be introduced to include the effect of the conducting tape are outlined in the discussion of the general theory. The agreement between the calculated and the measured values of return resistance proves that the method developed in the present paper is accurate even at the highest frequencies employed in telephony.

NOTE I—NOTE ON BESSEL FUNCTIONS

The Bessel Functions of zero order of the first and second kinds, $J_o(\rho)$ and $K_o(\rho)$, used in the preceding work are all to a complex argument $\rho = iq\sqrt{i}$ where q is a real number and $i = \sqrt{-1}$. The following formulas[9] may be used for determining the values of these functions:

$q < 0.1$

$$J_o(\rho) = 1 \qquad J'_o(\rho) = -\tfrac{1}{2}\rho$$

$$K_o(\rho) = \log_e \frac{2}{\gamma\rho} = .11593 - \log_e q - \frac{\pi i}{4}$$

$$K_o'(\rho) = -\frac{1}{\rho}$$

(*Jahnke u. Emde*, "Funktionentafeln," pp. 97, 98.)

$0.1 < q < 10$

The reports of the British Association for 1912 and 1915 give the values in this range of the functions ber q, ber′ q, bei q, bei′ q, ker q, ker′q, kei q, kei′q which are defined by the relations

$$J_o(iq\sqrt{i}) = \text{ber } q + i \text{ bei } q,$$
$$i\sqrt{i}\, J_o'(iq\sqrt{i}) = \text{ber}'q + i \text{ bei}'q,$$
$$K_o(iq\sqrt{i}) = \text{ker } q + i \text{ kei } q,$$
$$i\sqrt{i}\, K_o'(iq\sqrt{i}) = \text{ker}'q + i \text{ kei}'q.$$

[9] It is to be noted that this approximation for $K_o(\rho)$ differs from the expression used by J. J. Thomson, "Recent Researches in Electricity and Magnetism," p. 263. Thomson's formula (2) from which his approximation was derived, contains a number of errors and should read

$K_o(x) = (-C + \log 2i - \log x) J_o(x) - 2J_2(x) - \frac{1}{2}J_4(x) + \frac{1}{3}J_6)(x) \ldots\ldots\ldots$
where $C = .5772 \log = \log \gamma$.

$q > 10$

$$J_o\,(q\sqrt{-i}) = \frac{e^{q/\sqrt{2}}}{\sqrt{\pi q}}\left[\cos\left(\frac{q}{\sqrt{2}} - \frac{\pi}{8}\right) + i\sin\left(\frac{q}{\sqrt{2}} - \frac{\pi}{8}\right)\right]$$

$$J_o'\,(q\sqrt{-i}) = i\,J_o\,(q\sqrt{-i})$$

$$K_o(q\sqrt{-i}) = \sqrt{\frac{\pi}{2q}}\,e^{-q/\sqrt{2}}\left[\cos\left(\frac{q}{\sqrt{2}} + \frac{\pi}{8}\right) - i\sin\left(\frac{q}{\sqrt{2}} + \frac{\pi}{8}\right)\right]$$

$$K_o'\,(q\sqrt{-i}) = -\,i\,K_o\,(q\sqrt{-i})$$

Note II—Transformation of Fourier-Bessel Expansion

In problems involving Fourier-Bessel expansions it is sometimes necessary to transform quantities of the form

$$\frac{\cos s\phi_j}{\rho_j^s},\ \frac{\sin s\phi_j}{\rho_j^s},\ \log \rho_j,$$

from the system of coordinates ρ_j, ϕ_j to the systems ρ, ϕ or r, θ which are related as shown in Fig. 5.

The necessary formula may be derived as follows. We have

$$\frac{\cos s\phi_j + i\sin s\phi_j}{\rho_j^s} = \frac{e^{is\phi_j}}{\rho_j^s} = \left(\frac{e^{i\phi_j}}{\rho_j}\right) = \frac{1}{Z_j^s},$$

where Z_j is the conjugate of the vector $Z'_j = \rho_j\epsilon^{i\phi_j}$. Similarly we may write

$$Z = \rho\epsilon^{i(\phi-\pi+2a_j)},$$

$$C_j = c_j\,\epsilon^{(\pi+a_j)}$$

The vectors Z_j, Z and C, as may be seen from Fig. 5, have the lengths ρ_j, ρ and c, respectively, and the directions indicated by the arrows.

By vector addition,

$$Z'_j = Z + C_j$$

whence

$$Z_j = Z' + C'_j,$$

where Z' and C'_j are the conjugates of Z and C_j respectively.

By expansion

$$\frac{1}{Z_j^s} = \frac{1}{(Z' + C'_j)^s} = \frac{1}{C_j'^s}\left[1 - \frac{s}{1}\frac{Z'}{C'_j} + \frac{s\,(s+1)}{1.2}\frac{Z'^2}{C'^2} - \frac{s\,(s+1)\,(s+2)}{1.2.3}\frac{Z'^3}{C_j'^3} + - -\right]$$

We have $$\frac{1}{C_j'^s} = \frac{\epsilon^{is(\pi+a_i)}}{c_j^s} = (-1)^s \frac{\epsilon^{isa_j}}{c_j^s},$$

and $$\frac{Z'^n}{C_j'^n} = \frac{\rho^n}{c_j^n}\epsilon^{in(2\pi-\phi-\alpha_j)} = \frac{\rho^n}{c_j^n}\epsilon^{-in(\phi+\alpha_j)}$$

Therefore

$$\frac{\epsilon^{si\phi_j}}{\rho_j^s} = \frac{1}{Z_j^s} = \frac{(-1)^s}{c_j^s}\left[\epsilon^{isa_j} - \frac{s}{1}\frac{\rho}{c_j}\epsilon^{-i(\phi-a_j(s-1))} + \frac{s(s+1)}{1.2}\frac{\rho^2}{c_j^2}\epsilon^{-i(2\phi-a_j(s-2))} - \right]$$

Equating the real and imaginary parts gives

$$\frac{\cos s\phi_j}{\rho_j^s} = \frac{(-1)^s}{c_j^s}\left[\cos sa_j - \frac{s}{1}\frac{\rho}{c_j}\cos(\phi - a_j[s-1]) + \frac{s(s+1)}{1.2}\frac{\rho^2}{c_j^2}\cos(2\phi - a_j[s-2]) - - - - - -\right].$$

$$\frac{\sin s\phi_j}{\rho^s j} = \frac{(-1)^s}{c_j^s}\left[\sin sa_j + \frac{s}{1}\frac{\rho}{c_j}\sin(\phi - a_j[s-1]) + \frac{s(s+1)}{1.2}\frac{\rho^2}{c_j^2}\sin(2\phi - a_j[s-2]) - - - - - - -\right].$$

Similarly

$$\log Z_j = \log(C_j' + Z')$$

$$= \log C_j' + \frac{Z'}{C_j'} - \frac{1}{2}\frac{Z'^2}{C_j'^2} + \frac{1}{3}\frac{Z'^3}{C_j'^3} - - - - - - - -$$

$$= \log C_j' + \sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\frac{\rho^n}{c_j^n}\epsilon^{in(\phi-a_j)}.$$

Equating real and imaginary parts we have

$$\log \rho_j = \log c_j + \frac{\rho}{c_j}\cos(\phi - a_j) - \frac{1}{2}\frac{\rho^2}{c_j^2}\cos 2(\phi - a_j) + \ldots +,$$

$$\phi_j = \frac{\rho}{c_j}\sin(\phi - a_j) - \frac{1}{2}\frac{\rho^2}{c_j^2}\sin 2(\phi - a_j) + \ldots +.$$

The following formulas may be derived in a similar manner:

$$\frac{\cos s\phi_j}{\rho_j^s} = \frac{(-1)^s}{c^s}\left[1 + \frac{s}{1}\frac{r}{c}\cos(\theta - \gamma_j) + \frac{s(s+1)}{1.2}\frac{r^2}{c^2}\cos 2(\theta - \gamma_j) + \ldots +\right],$$

$$\frac{\sin s\phi_j}{\rho_j^s} = \frac{(-1)^{s+1}}{c^s}\left[\frac{s}{1}\frac{r}{c}\sin(\theta - \gamma_j) + \frac{s(s+1)}{1.2}\frac{r^2}{c^2}\sin 2(\theta - \gamma_j) + \ldots +\right],$$

$$\log \rho_j = \log c - \frac{r}{c}\cos(\theta - \gamma_j) - \frac{1}{2}\frac{r^2}{c^2}\cos 2(\theta - \gamma_j) - - - - - .$$

NOTE III—DETERMINATION OF q_1, q_2, ETC.

The equations (47),

$$q_n = (-1)^n \lambda_n \frac{\zeta^{2n}}{n} (A + S_{nn} B_o) - (-1)^n \lambda_n \frac{\zeta^{2n}}{n!} \Sigma_n \quad n = 1 - - \infty$$

are linear in the variables q_1, q_2 — , since

$$\Sigma_n = n! \, S_{n-1,n+1} \, q_1 - \frac{n!}{1!} S_{n-2,n+2} \, q_2 - - - .$$

The values q_1, q_2 . . . may be determined by a method of approximations, q_n being the limit of the sequence

$$q_n^{(0)} , q_n^{(1)} , q_n^{(2)} - - - - - ,$$

the successive terms of which are defined by the expressions

$$q_n^{(0)} = (-1)^n \lambda_n \frac{\zeta^{2n}}{n} (A + S_{nn} B_o),$$

$$- - - - - - - -$$

$$q_n^{(j+1)} = (-1)^n \frac{\zeta^{2n}}{n!} (A + S_{nn} B_o) - (-1)^n \frac{\zeta^{2n}}{n!} \Sigma_n (q^{(j)}),$$

where $\Sigma_n (q^{(j)})$ is the value of Σ_n when q_1, q_2 — are replaced by $q_1^{(j)}$, $q_2^{(j)}$ — — — .

This method, however, while formally simple and direct is not usually well adapted for numerical solution. For all sizes of armor wire and for frequencies of practical importance the argument ζ in the expression (48) is small compared with μ and the quantities,

$$\lambda_1, \lambda_2,$$

are all nearly unity. This suggests the use of the following method of solution of equations (47).

The solution of the auxiliary set of equations

$$p_1 = - \zeta^2 (A + S_{11} B_o) + \frac{\zeta^2}{1!} \Sigma_1(p),$$

$$p_n = (-1)^n \frac{\zeta^{2n}}{n} (A + S_{11} B_o) - (-1)^n \frac{\zeta^{2n}}{n!} \Sigma_n(p),$$

in the auxiliary variables p_1, p_2 — may be written,

$$p_1 = - \zeta^2 C_{11} (A + S_{11} B_o) + \frac{\zeta^4}{2} C_{12} (A + S_{22} B_o) + \ldots + ,$$

$$p_n = - \zeta^2 C_{n1} (A + S_{11} B_o) + \frac{\zeta^4}{2} C_{n2} (A + S_{22} B_o) + \ldots + ,$$

in which C_{11}, etc., are numerics. This solution is effected by retaining a finite number of equations and an equal number of variables and solving by the usual methods. It will be found that except in extreme cases, a very good approximation can be gotten by ignoring all the p's except the first four. The q's may then be obtained by the relation

$$q_n = p_n + d_n$$

d_n being defined by

$$d_n = (\lambda_n - 1)p_n - (-1)^n \lambda_n \frac{\zeta^{2n}}{n!} \Sigma_n(d).$$

This system is easily adapted to solution by successive approximations,

$$d_n = d_n^{(0)} + d_n^{(1)} + d_n^{(2)} + - - -$$

in which

$$d_n^{(0)} = \left(1 - \frac{1}{\lambda_1}\right) C_{n1} p_1 + \ldots + \left(1 - \frac{1}{\gamma_n}\right) C_{nn} p_n,$$

$$d_n^{(j+1)} = \left(1 - \frac{1}{\lambda_1}\right) C_{n1} d_1^{(j)} + \ldots + \left(1 - \frac{1}{\lambda_n}\right) C_{nn} d_n^{(j)},$$

C_{n1}, etc., being the numerical coefficients which appear in the expressions for p_1, p_2 . . . .

A very good approximation which holds in most cases is

$$d_n = (\lambda_1 - 1)\, C_{n1}\, p_1 + (\lambda_2 - 1)\, C_{n2}\, p_2 + \ldots + (\lambda_n - 1) C_{nn} p_n.$$

Analysis of the Energy Distribution in Speech[1]

By I. B. CRANDALL and D. MacKENZIE

SYNOPSIS: *The frequency distribution of energy in speech* has been determined for six speakers, four men and two women, for a 50-syllable sentence of connected speech, and also for a list of 50 disconnected syllables. The speech was received by a condenser transmitter whose voltage output, amplified 3,000 fold, was impressed on the grids of twin single stage amplifiers. The unmodified output of one of these amplifiers was measured by a thermocouple and was a known function of the total energy received by the transmitter, corrections being made for the slight variation with frequency of the response of the circuit. The output of the other amplifier was limited by a series resonant circuit to a narrow band of frequencies, the energy in this band being measured by a second thermocouple. The damping of the resonant circuit was so chosen that sufficient resolving power and sufficient energy, sensitiveness were obtained over the range from 75 to 5,000 cycles per second; and 23 frequency settings were made to cover this range. For each syllable simultaneous readings were recorded on the two thermocouples at each frequency setting. The consecutive syllables were pronounced deliberately by each speaker, maintaining as nearly as possible the normal modulation of the voice. Corrections were applied to offset the unavoidable variations in total energy incidental to repetition of a given syllable. 13,800 observations were made for all speakers. *The energy distribution curves* obtained are essentially the same for connected as for disconnected speech, and indicate that differences between individuals are more important than variations due to the particular test material chosen. A composite curve drawn from the individual curves shows a great concentration of speech energy in the low frequencies, a result which would not be expected from data previously published by others. The actual results contain a factor due to standing waves between the speaker's mouth and the transmitter, a complication always present in telephoning; this could not be eliminated.

The rate of energy output in speech for the normally modulated voice, was determined from the readings for total energy and was found to be about 125 ergs per second.

IN the study of speech and its reproduction by mechanical apparatus it is necessary to consider its composition from several different points of view. We desire first of all to know the actual frequency distribution of the total energy in speech, as well as the separate distributions for each individual sound. We also desire to know the apparent distribution of energy, that is, the distribution as perceived by the ear. Finally, we wish to know the importance of each frequency, that is, the contribution to "articulation" or "quality" in the exact reproduction of speech which can be traced to the energy of each elementary band of frequencies in the speech range. In all three cases certain frequency functions are used to represent these distributions. The advantage of considering these different frequency distribution functions separately has already been indicated by one of the present writers.[2]

[1] Reprinted from THE PHYSICAL REVIEW, N.S., Vol. XIX. No. 3, March, 1922.

[2] "The Composition of Speech," PHYS. REV., X, p. 74, 1917.

In our judgment the most important of these data of speech study is the *actual energy distribution*, considering speech as "a continuous flow of distributed energy," in accordance with the ideas expressed in the earlier paper. The present paper offers a determination of this fundamental factor.

To determine the energy distribution in speech to a high degree of accuracy it would be desirable to analyze a certain amount of connected speech and take a time average of the energy distribution of the whole. This is not feasible at the present time, but a very close approach to this result has been made. The method consists in analyzing the speech waves as impressed on a condense transmitter, using a tuned circuit to transmit narrow frequency bands of energy and pronouncing the separate syllables of the connected speech so slowly that the kick of a direct current galvanometer connected to an A. C. thermocouple can be separately read for each syllable. Using a suitable calibration for the whole apparatus, the magnitude of this kick can be interpreted in terms of the time integral of the energy at a particular frequency setting for each syllable. A mean of the readings for all the syllables in the "speech" at any frequency setting gives the relative energy at that frequency.

The present method is a modification of an earlier method in which approximate analyses of speech sounds were made, using a condenser transmitter, tuned circuit, an amplifying-rectifying circuit, and ballistic galvanometer. The method is, however, much improved as we now have very accurately calibrated condenser transmitters of better design,[3] and a great deal of care has been taken to calibrate the successive elements of the train of apparatus, and increase the resolving power.

Experimental Procedure

Sound waves emitted from the mouth of the speaker are allowed to fall upon the diaphragm of a condenser transmitter, connected in the conventional manner to the input of a three-stage amplifier. The output of this is impressed upon the input circuits of twin single stage amplifiers, potentiometers being interposed to permit regulation of the grid voltages of the twin amplifier tubes.

The output circuits of the fourth stage consist of the high windings of two step down ironclad transformers. These step down transformers have a voltage ratio of 11: 1 and are designed to work between impedances of 6,000 and 50 ohms. The low impedance winding of one of these transformers operates into a thermocouple heater of,

[3] The present design of the condenser transmitter and its calibration are fully treated in a paper by Dr. E. C. Wente which will appear shortly in this Journal.

roughly, 40 ohms resistance. The low side of the other transformer operates through a tuned circuit into a similar thermocouple heater.

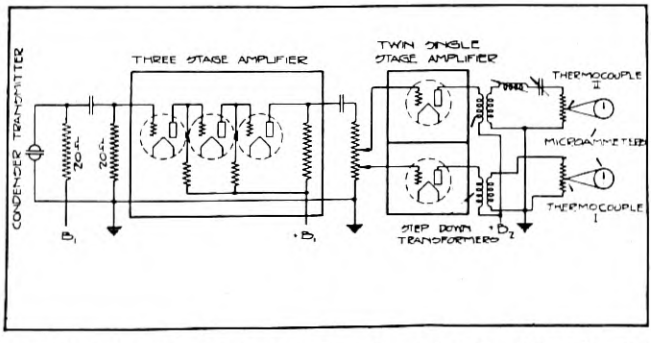


Fig. 1—Circuit Used for the Analysis of Speech. (The Usual Details of the Three-Stage Amplifier Are Not Shown)

The diagram of Fig. 1 exhibits the essential features of the electrical circuits just described.

When the diaphragm of the condenser transmitter is set in vibration by speech a current made up of a range of frequencies flows in the heater of thermocouple I., while the heater of thermocouple II is traversed only by such a band of frequencies as the resonant circuit allows. Fig. 2 shows a number of typical resonance curves obtained

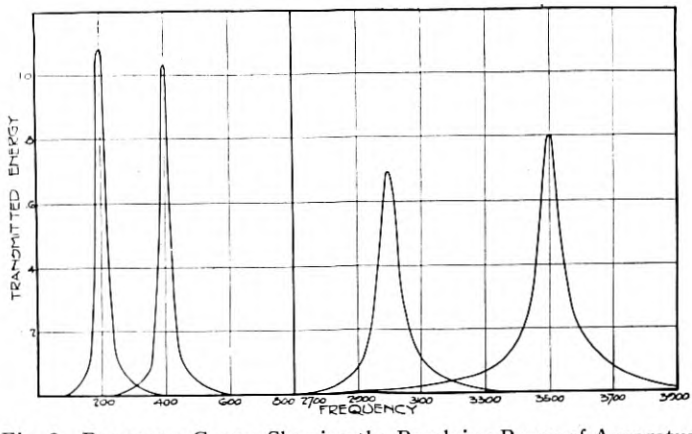


Fig. 2—Resonance Curves Showing the Resolving Power of Apparatus

in the course of calibrating this apparatus. These curves are such that the tuned circuit functions as a filter transmitter only a narrow region of frequencies. One side of the twin amplifier transmits the

entire electrical response of the system; the other side suppresses all save a band of frequencies, the center of this band being shifted by resetting the condenser and inductometer.

Having chosen for analysis a piece of connected discourse, the speaker utters the successive syllables separately but as nearly as may be with the same inflection and volume as if the syllables were continuously spoken. Two observers record the readings of microammeters in the couple circuits of the thermocouples. One of these instruments gives a deflection corresponding to the total energy of the syllable uttered; the deflection of the other instrument corresponds to the energy of the syllable lying within the limits of transmission of the tuned circuit.

Preliminary experiments were carried out to determine the relation between momentary deflection read on the microammeter, and the current momentarily flowing in the thermocouple heater. Currents of different values were caused to flow for intervals of time varying from 0.2 second to 1.2 seconds, and the deflections were found nearly proportional to the product of current squared and time interval; this proportionality was most nearly exact when the current was weak and the time intervals short. For all cases likely to be duplicated in the speech analysis work the error might be taken as about 5 per cent, a quantity small in comparison with the inevitable uncertainties due to other causes.

Quite low damping is attained in the resonant circuit. The values of inductance used ranged from 0.20 to 0.66 henry and the total resistance of the circuit—transformer winding, inductometer coil, thermocouple heater—is of the order of 100 ohms. The damping thus ranges from 75 to 250.

The circuit is calibrated in the following manner:

A switch is so introduced that it is possible to include in series with the thermocouple the resonant circuit, or replace it by a non-inductive resistance whose value is approximately that of the A. C. resistance of the inductometer winding. With the tuned circuit excluded, an alternating current of suitable magnitude is caused to flow in the thermocouple heater; the tuned circuit is then substituted and the new value of the current observed, the input voltage remaining constant. The ratio of current squared "tuned circuit in" to current squared "tuned circuit out" is plotted against frequency, yielding a curve for energy transmission.

Twenty-three bands in all were considered adequate for the analysis of energy distribution in speech; the centers of these were at 75, 100, 200, 300 cycles, 400 to 3,200 cycles by steps of 200; 3,500, 4,000, 4,500,

5,000 cycles per second. Beyond 5,000 cycles per second, the energy is so low as to be impossible of measurement with the apparatus used. A Weston Type 322 microammeter recorded the couple current for the tuned circuit side of the twin single stage amplifier. With this instrument and the thermocouple used, 0.2 microampere in the couple circuit corresponds to one-quarter of a milliampere in the heater, and this is the lowest readable deflection of the Weston instrument.

Reduction of Observations

Three corrections have to be made, the first being the correction for varying volume.

Simultaneous observations are made, at each setting of the tuned circuit, of the filtered and the unfiltered energy of each syllable. It is not possible to utter a given syllable with the same intensity and at the same distance from the transmitter for every one of twenty-

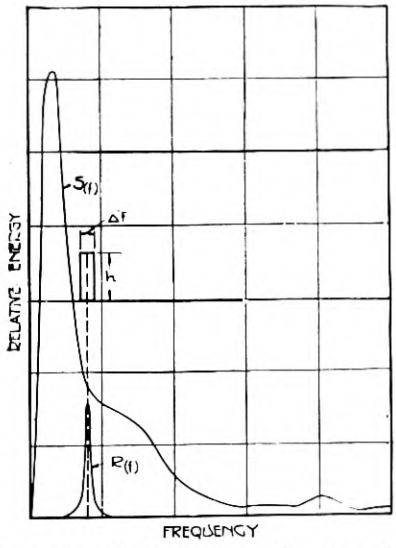


Fig. 3—Illustrating Correction of Observations, Necessary Because of Variation in Resolving Power with Frequency Setting

three times. Accordingly, the "unfiltered" readings are averaged and each of the filtered readings for each syllable reduced from the value actually observed to the value that would have been read had the volume and distance been such as to give the average "unfiltered" reading. This procedure is quite legitimate if it be granted possible to maintain a definite composition of the syllable in question throughout the changes of the tuned circuit setting.

A second correction was made for the varying area of tuned circuit curves.

In Fig. 3 let $S(f)$ be the speech spectrum determined by ideal methods; "R" the transmission curve of the tuned circuit, set for a resonant frequency f. An ideal transmission curve would be a rectangle when plotted in this figure, of height "h" and transmission range Δf.

The true amount of energy $S(f)$ associated with frequency f, and the experimentally determined value which we may call $(\overline{S}f)$ are connected by the relation

$$h\overline{S}(f)\Delta f = \int_f^{f+\Delta f} S(f)R(f)df$$

and if we make

$$h\Delta f = \int_f^{+\Delta f} R(f)df$$

we may take for all practical purposes $S(f) = \overline{S}(f)$, considering the narrowness of the transmission range. We must therefore find the factor $h\Delta f$, proportional to the area of each tuned circuit curve and divide the energy received through the filtered side by $h\Delta f$, in order to obtain $S(f)$. This treatment may be gone through for each syllable individually, but it is more convenient to sum the tuned circuit readings for all the syllables used, corrected one at a time for varying volume, and then apply the curve area correction to this sum.

A third correction was made for the varying frequency-sensitivity of the whole apparatus. Thus far we have discussed only the electrical energy in the output circuit of the fourth stage. It remains to show in what way this is related to the mechanical energy of the diaphragm, and this in turn to the incident sound energy.

The calibration of the circuit as a whole was made by introducing a small resistance carrying alternating current in series with the condenser transmitter, thus introducing a known potential drop in the undisturbed input mesh of the circuit.

An amplification curve is appended (A, Fig. 4) which gives to an arbitrary scale the ratio of volts output to volts input as a function of frequency, for the system as actually operated. The calibration of the condenser transmitter, shown in Fig. 4, Curve C, gives the open circuit voltage of the transmitter per unit pressure on the diaphragm as a function of frequency. The product of these curves is the volts output per unit alternating pressure on the diaphragm, and the square of this product, curve E is proportional to the electrical energy output per unit sound energy incident on the diaphragm, if we assume that the sound energy is proportional to the square of the alternating pressure. This point, however, requires some further discussion, which will be given later on.

It is plain from curve E that the response of the system is a maximum of frequencies in the neighborhood of 2,250 cycles. If, now, the observations already corrected for varying volume and for area of resonance curves, are subjected to further correction for the exaggeration of these frequencies, it is possible to draw a curve which shall

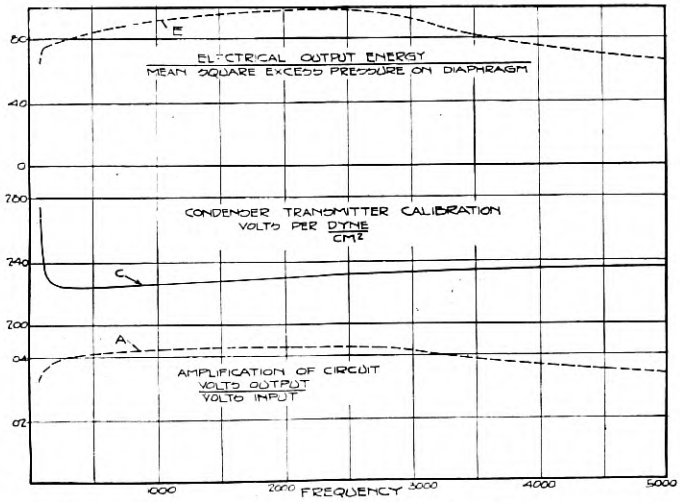


Fig. 4—Energy-Frequency Characteristics of the Apparatus

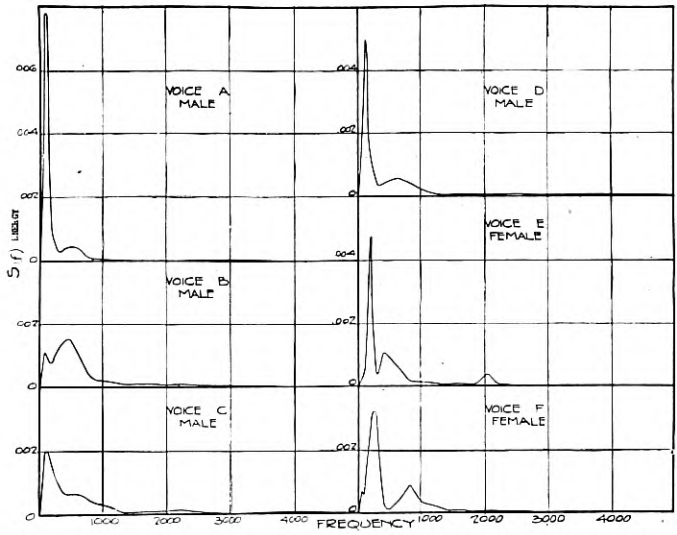


Fig. 5—Analysis of Individual Voices

exhibit the mean square of the excess pressure on the diaphragm, as a function of frequency in the voice exciting the vibration. We obtain this corrected curve by dividing the results, after the first and second corrections above have been made, by the ordinates of curve *E*.

OBSERVATIONS

In order to investigate the possibilities of this method it was decided to work with a rather short piece of connected speech, and to use a limited number of observers, on account of the large number of observations which are required for each separate syllable. With six speakers (four men and two women) each pronouncing the test sentence of fifty syllables for each of the twenty-three frequency settings, 6,900 separate observations were required. It is believed that representative results have been obtained from these observations, but if this is not the case then some method of graphical registration of the energy-time curve of speech for the different frequency settings must be applied in order to handle the vast amount of data involved in work on an appreciably larger scale.

The test sentence used was as follows:

"*Quite* four score and seven years ago our father brought forth on this continent, a *nice* new nation, conceived in liberty, and dedicated to the proposition that all men are created equal."

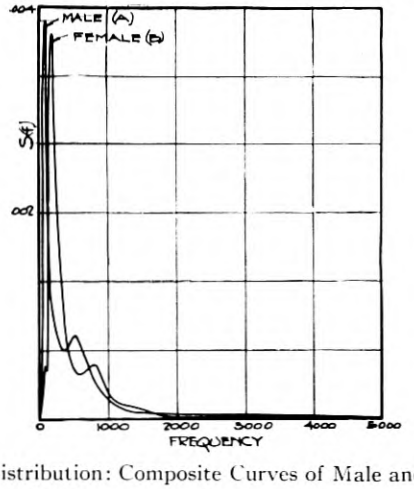


Fig. 6—Energy Distribution: Composite Curves of Male and Female Voices

The two *italicized* words were added to the first sentence of the "Gettysburg Address" in order to bring the total up to fifty syllables, and improve the balance between the vowel sounds.

The resulting speech-energy curves are shown in Figs. 5, 6 and 7,

plotted so that $\int_0^\infty S(f)\,df = 1$ in each case. In Fig. 5 the individual curves for each of the six speakers are shown on a small scale; in Fig. 6 the composite curve for the men and the composite curve for the women, drawn separately, and in Fig. 7 the composite curve for all six speakers, giving the data of curves 6*A* and 6*B* equal weight.

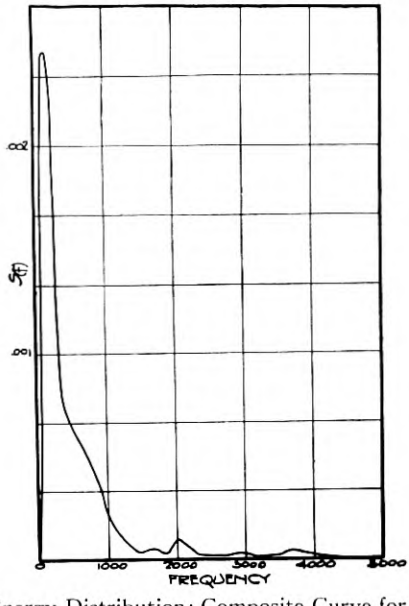


Fig. 7—Energy Distribution: Composite Curve for All Voices

These curves are very similar to a curve obtained by Dr. Fletcher of this laboratory, using block filters and based on the simple calling sentence "Now we're off on one." A general consideration of this fact and of the data shown leads us to believe that the differences between curves of this sort, made by the method described are due rather more to differences between the voices of the individual speakers than to the particular piece of connected speech which is chosen, provided the speech is of reasonable length. The differences between the different voices are so marked that we should expect them to remain even though we used as test material a connected speech ten or fifty times as long as the sentence used.

The Energy Distribution in Speech

An interesting comparison may be made between the curves shown for the energy distribution of "continuous speech" and certain speculative curves previously constructed to indicate the energy distribu-

tion. One of these curves is shown in Fig. 8. Curve *A* was constructed by one of the writers in 1916 in an attempt to synthesize the energy curve from the energy distributions of the vowel sounds, using the vowel analyses of Dr. Dayton C. Miller. Curve *C* is the composite "continuous speech" curve of Fig. 7. The vowel sounds analyzed by

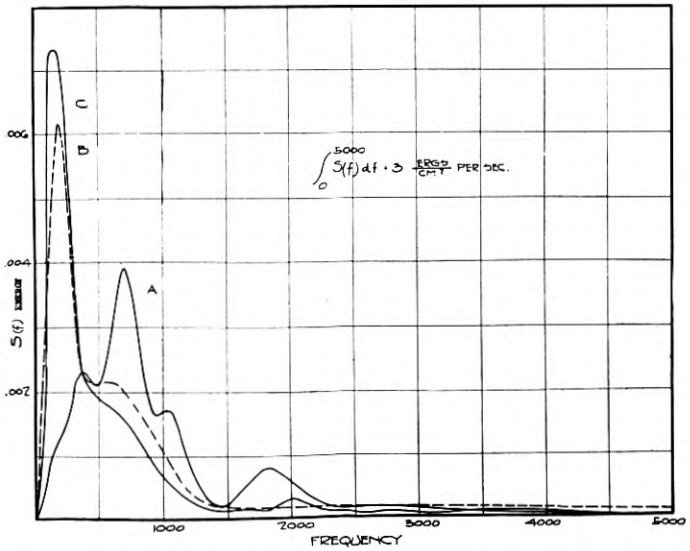


Fig. 8—Energy Distribution: *A*. Synthesized from Vowel Records of D. C. Miller (1916). *B*. Disconnected Speech Analysis of this Paper. *C*. Connected Speech Analysis of this Paper (from Fig. 7)

Miller were intoned and the vowel sounds analyzed by us were spoken, but Miller's work seemed to show that there was no essential difference between intoned and spoken vowel sounds. There is, however, a very noticeable difference between Curve *A* and Curve *C*, the energy in the fundamental tone of the speaker's voice coming out much more strongly in Curve *C*. We should expect that our improved apparatus would record the energy in the lower frequencies more correctly than the apparatus heretofore used but as we used different test material (connected speech instead of disconnected syllables or vowel sounds) it is not immediately evident which of these two factors is responsible for the differences between the *A* and the *C* curves.

In order to investigate this point more fully the testing routine for all six speakers was repeated, using instead of the fifty-syllable sentence, the fifty disconnected syllables of one of the standard articulation testing lists, as used by Dr. Fletcher in this laboratory. The results for energy distribution are shown in Fig. 9, Curve *A* being

the mean energy distribution for the four male speakers, using the syllables, while Curve *B* is the mean energy distribution of the two female speakers. Curves 9*A* and 9*B* may be compared with Curves 6*A* and 6*B* which represent the sentence of continuous speech. The two sets of curves are essentially the same as shown in Fig. 8, *C* and *B*

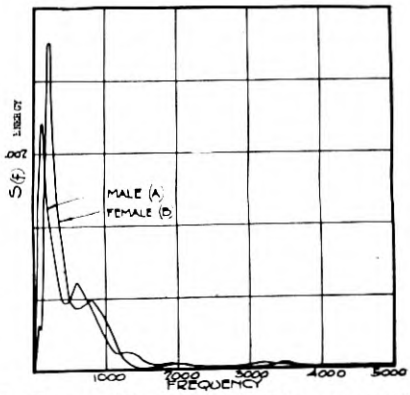


Fig. 9—Energy Distribution in Disconnected Speech

being respectively the composite curves for all speakers, using connected and disconnected speech.

Such small differences as exist between Curves *C* and *B* of Fig. 8 may probably be due to differences in the distribution of the vowel sounds in the connected and disconnected test material. This distribution is given in the following table:

Vowel Sounds	a	ā	á	e	ē	ér	i	ī	o	ō	ó	u	ū	ou	Total
In Sentence.....	6	6	3	7	3	3	7	2	2	5	3	0	2	1	50
In Syllabic List (No. 174).....	4	4	3	4	3	4	3	4	3	4	3	4	4	3	50

Key to Vowel Sounds: a, as in father (or o as in top); ā, as in tape; á, as in tap; e, as in ten; ē, as in team; ér, as in term; i, as in tip; ī, as in time; o, as in ton; ō, as in tone; ó, as in for; u, as in pull; ū, as in rule; ou, as in house

The similarity between Curves *C* and *B* of Fig. 8 is evidence of the general reliability of the method, and leads to two rather important conclusions.

In the first place, characteristic results have been obtained for a given set of speakers, using two different types of test materials. This

seems to show that the choice of test material does not require especial consideration, provided it is of sufficient length. It seems to be a matter of rather greater importance to increase the number of observers.

In the second place, it seems that for the actual energy distribution, the results previously obtained from the vowel analyses are definitely in error, in that they show relatively little energy associated with the lower voice frequencies.

Criticism of the Results

The foregoing treatment provides a curve showing the frequency distribution of the square of the excess pressure on the diaphragm.

In an undisturbed field of sound energy we have for the intensity

$$I = \frac{P^2}{2\rho a}$$

in which ρ is the mean density of the medium, a the velocity of sound in the medium and P the maximum excess pressure.

It remains for us to consider in how far the results obtained represent the frequency distribution of sound energy in speech.

Due to the fact that at frequencies where the sound wave-length is short and comparable with the diameter of the transmitter, considerable reflection takes place, and the pressure on the diaphragm is proportionately greater for these frequencies than for those which are not accompanied by strong reflection. In this respect again the higher frequencies provoke the greater response in the system.

The following experiment was tried to investigate this variation. A wall six feet square, with a central hole to fit over the condenser transmitter, was brought up to make the transmitter a part of a plane wall. The clearance around the periphery of the transmitter was tightly closed, and reflection was to be expected at all frequencies. Where total reflection takes place, a given quantity of sound energy results in twice the alternating pressure on the diaphragm as when no reflection occurs. That is, the resulting electrical energy observed should be four times as great for total reflection as for no reflection. The wall was expected to cause reflection at all frequencies, and the experiment consisted in reading the electrical response, with and without the wall, the condenser transmitter being exposed to tones of frequencies from 200 to 10,000 cycles per second under definite adjustments of the supply circuit of a receiver producing this tone. When the frequency is low, little reflection takes place from the transmitter standing alone, and bringing up the wall should cause a great increase in the response

of the system. At high frequencies the transmitter should reflect nearly as much alone as when part of a large wall, and the readings with and without the wall should be nearly equal. Plotting ratio of response without, to response with the wall was expected to yield a curve which could be used to make the final reduction of electrical output to incident sound energy, and so permit a more accurate determination of the spectrum of sound energy of the voice.

No consistent results were obtained after several trials and the experiment was abandoned. The failure is doubtless to be ascribed to standing waves, the character of which is very sensitive to the location in the room of the transmitter and the wall. This experiment is to be repeated under more favorable conditions when standing waves can be eliminated.

Thus, the curves finally obtained show no more than the frequency distribution of energy in speech in terms of the mechanical energy of a more or less ideal transmitter diaphragm. However, this information has its value because in any given configuration of transmitter, speaker, and room, there is a definite correspondence between the sound energy of the voice and the force acting on the diaphragm on which it falls, and in telephony at any rate it is this action on the diaphragm with which we are immediately concerned.

In conclusion we may give a determination of the total energy rate of speech, obtained as a by-product of the preceding investigation. Knowing the calibration of the system in absolute units, it is possible to determine the alternating pressure on the condenser transmitter diaphragm exposed to continuous speech from the normally modulated voice under the conditions of the experiment. Using the mean of the values obtained with 9 observers we find for the alternating pressure 11.3 dynes per sq. cm. (r.m.s.) for a distance of 2.5 cm. from mouth to diaphragm. This corresponds to an energy flow of 3.2 ergs per sq. cm. per second. Assuming that this energy flow is distributed uniformly over a hemisphere of 2.5 cm. radius, we may take 125 ergs per second as the total sound energy flow from the lips with the normally modulated voice.

The Nature of Speech and Its Interpretation[1]

By HARVEY FLETCHER

Introduction

VARIOUS phases of this subject have received serious study by phoneticians, otologists, and physicists. On account of its universal interest, it has received attention from men in many branches of science. In spite of the large amount of time devoted to the subject, the progress in understanding its fundamental aspects has been rather slow. At the present time the physical properties which differentiate the various fundamental speech sounds are understood in only a very fragmentary way. Some very interesting and painstaking work has been done on the physical analysis of vowel sounds, but the results to date are far from conclusive. Although several theories have been advanced to explain the way in which the ear interprets sound waves, they are still in the controversial stage.

The material which is presented here is the result of an investigation which has been carried on in the Research Laboratories of the American Telephone and Telegraph Company and Western Electric Company during the past few years.

To make a quantitative study of speech and hearing it is necessary to obtain the speech sounds at varying degrees of loudness and with definitely known amounts of distortion. The main reason why so few real results have been obtained in the investigation of speech sounds is due to the fact that it is extremely difficult to change the volume and distortion of these sounds by acoustic means. Due to recent developments in the electrical transmission of speech it is possible to produce the equivalent of these changes by electrical means. For this purpose a telephone system was constructed which reproduced speech with practically no distortion. It was arranged so that by means of distortionless attenuators the volume of reproduced speech could be varied through a very wide range, and so that by introducing various kinds of electrical apparatus the transmitted speech wave could be distorted in definitely known ways.

A method was developed for measuring quantitatively the ability of the ear to interpret the transmitted speech sounds under different conditions of distortion and loudness. By choosing these conditions properly, considerable information was gained concerning both speech and hearing. This indirect method of attack has a distinct advantage

[1] Presented at a meeting of the Electrical Section of the Franklin Institute held Thursday, March 30, 1922. Reprinted from the Journal of the Franklin Institute for June 1922.

for engineering purposes, in that it measures directly the thing of most interest, namely, the degrading effect upon telephone conversation of introducing electrical distortion into the transmission circuit. However, the application of the results is not limited to this particular field.

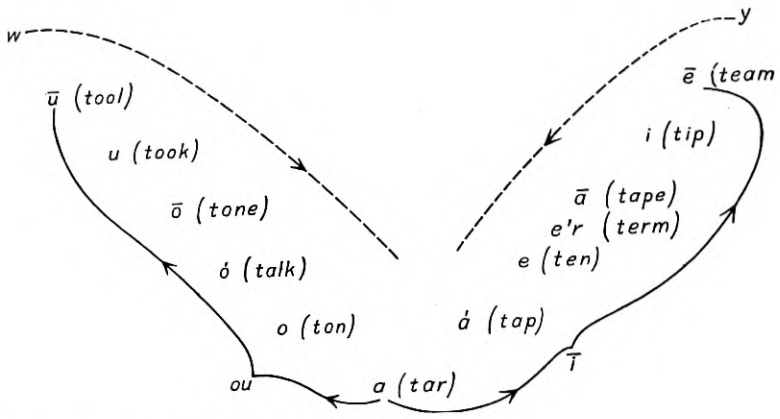
Method of Measuring the Quality of Speech

Briefly stated the method consists in pronouncing detached speech sounds into the transmitting end of the system and having observers write the sounds which they hear at the receiving end. The comparison of the called sounds with those observed shows the number and kinds of errors which are made. The per cent of the total sounds spoken which are correctly received is called the articulation of the system.

Table I.

Classification of the Speech Sounds.

Pure Vowels



Combinational and Transitional Vowels

w — y — ou — ī — h

Semi-vowels

l — r

Stop Consonants

Voiced	Unvoiced	Nasalized	Formation of Stop
b	p	m	lip against lip
d	t	n	tongue against teeth
j	ch	—	tongue against hard palate
g	k	ng	tongue against soft palate

Fricative Consonants

Voiced	Unvoiced	Formation of Air Outlet
v	f	lip to teeth
z	s	teeth to teeth
th (then)	th (thin)	tongue to teeth
zh (azure)	sh	tongue to hard palate

In order to understand the construction of the articulation lists and also to interpret the results of this investigation, I desire to give here a brief classification of the speech sounds, which is based upon the position of the various speech organs when the sounds are being produced. It is shown in the accompanying table (Table I).

The pure vowels are arranged in the vowel triangle, which is familiar to phoneticians. Starting with the sound ū the lips are rounded and there is formed a single resonant cavity in the front part of the mouth. Passing along the left side of the triangle from ū to a the mouth is gradually opened with the tongue lowered to form the successive vowels. Going along the right side of the triangle from a to ē, the tongue is gradually raised to the front part of the mouth forming two resonant chambers in the mouth cavity. An infinite number of different shadings of these vowels may be produced by placing the mouth in the various intermediate positions, but the ones which are shown were chosen as being the most distinct.

The sounds w, y, ou, ī and h are classed as combinational and transitional vowels. As the mouth is placed in the position to say ū and then suddenly changed so as to form any other vowel in the triangle, the result obtained is signified in writing by placing the letter w before the vowel. In a similar way we get the effect usually designated by y if the position of the vowel suddenly changes from ē to any other vowel. An infinite variety of dipthongs can be formed by changing the position of the mouth necessary to form one vowel to that to form another without interrupting the voice. The most distinct and principal ones used in our language are formed by passing from the sound a to either extreme corner of the triangle and are known as ou and ī. When a vowel commences a syllable it is formed by suddenly opening the glottis, permitting the air, which has been held in the lungs, to escape into the mouth, which is formed for the proper vowel. If the glottis remains open and the vowel is started by the sudden contraction of the lungs, we have the effect which is represented in writing by placing an h before the vowel. The sounds l and r are called semi-vowels because the voice train is partially interrupted, although the sound can be continued. The stop and fricative consonants are classified in a manner which is familiar to phoneticians.

It will be noticed that the markings are not those used in the international phonetic alphabet which were entirely too complicated for practical use. Only the bar and accent stroke are used. These can be written quickly and with little chance of error.

In order to pronounce these speech sounds properly, they must

be combined into syllables. For the purpose of this investigation they were combined into mono-syllables of the simple types consonant-vowel, vowel-consonant, and consonant-vowel-consonant.

To eliminate memory effects every possible combination of the sounds into these types of syllables was used unless there was a good reason for excluding it. The complete list contained 8700 syllables. For convenience of testing these syllables were divided into groups of fifty. Each group contained the same kind and number of syllable forms and an equal number of each of the fundamental vowel and consonant sounds.

TABLE II.

Speech-sound Testing List. List No. 160

	Speech-sound	Key-word		Speech-sound	Key-word
1	ha	ho(t)	26	gōb	go+b
2	hā	hay	27	shōl	shoal
3	wă	wa(g)	28	ros	rus(t)
4	wi	wi(th)	29	jod	ju(g)+d
5	vou	vow	30	bok	buck
6	ăr	air	31	zīk	z+(d)ike
7	ez	e(bb)+z	32	bīch	buy+ch
8	ūsh	you+sh	33	kīth	ki(te)+th
9	an	on	34	gīt	gui(de)+t
10	id	(l)id	35	yif	y+if
11	jouv	jow(l)+v	36	sin	sin
12	moush	mou(nd)+sh	37	tĕrm	term
13	rour	r+our	38	mĕrl	m+earl
14	zūth	z+(s)oothe	39	pĕrv	p+(n)erve
15	hūs	who+s	40	yēt	y+eat
16	chush	ch+(p)ush	41	bēl	b+eel
17	jum	j+(f)oo(t)+m	42	zef	ze(al)+f
18	thup	th+(s)oo(t)+p	43	weng	whe(n)+ng
19	fuch	foo(t)+ch	44	kev	k+ev(er)
20	wŏng	wa(ll)+ng	45	hăng	hang
21	chŏth	cha(lk)+th	46	păg	p+(r)ag
22	tŏj	ta(ll)+j	47	yās	y+ace
23	kŏg	k+aug(er)	48	dāp	d+ape
24	fōn	(tele)phone	49	yang	ya(cht)+ng
25	dōs	dose	50	lan	l+on

To illustrate the technique of articulation testing a sample list is given in Table II. In the first column the syllable is given in its phonetic form. A key-word showing how each syllable is pronounced is given in the second column. These syllables were written on cards which were shuffled each time before they were used, so that the order in which they were pronounced was entirely haphazard. One hundred and seventy-four similar lists were used in this work. In order to eliminate personal peculiarities, several

callers and observers were used. In Table III are shown the results obtained by an observer when this list was transmitted over a system which eliminated all frequencies above 1250 cycles per second.

TRANSMISSION BRANCH
ARTICULATION TEST RECORDING SHEET

WORD ARTICULATION 40 %

TITLE OF TEST 320311
CONDITION TESTED Low Pass Filter - 1250 ~
Attenuation = 5 napiers down
DATE 2-7-20 OBSERVER M.A
TEST No. 11 CALLER H.E.D
LIST No. 160

No.	OBSERVED	CALLED	ERRORS	No.	OBSERVED	CALLED	ERRORS
1	tan	te'rm	e'r-a m-n	26	zip	thup	th-z u-i
2	zit	gĭt	g-z ĭ-i	27	ko'd	to'j	t-k d-j
3	wa	wa'	a'-a	28	tish	chush	ch-t u-i
4	dāp	✓		29	yang	✓	
5	gō b	✓		30	zēt	zūth	ā-ē th-t
6	yis	yif	f-s	31	ref	ros	o-e s-f
7	māl	merl	e'r-ā	32	jum	✓	
8	thin	sin	s-th	33	jo'g	ko'g	k-j
9	zĭp	zĭk	k-p	34	jad	jod	o-a
10	jouv	✓		35	tūth	hūs	h-t s-th
11	yāt	yās	s-t	36	id	✓	
12	thou	vou	v-th	37	ha	✓	
13	bĭp	bĭch	ch-p	38	fōn	✓	
14	ha'ng	✓		39	ko'th	cho'th	ch-k
15	mĭs	moush	ou-ĭ sh-s	40	rour	✓	
16	dāch	dō s	ō-ā s-ch	41	an	✓	
17	kev	✓		42	bok	✓	
18	tig	pa'g	p-t a'-i	43	yēt	✓	
19	kĭs	kĭth	th-s	44	o'r	a'r	a'-o' y inserted
20	hā	✓		45	yēth	ūsh	ā-ē sh-th
21	weng	✓		46	wo'ng	✓	
22	dēl	bēl	b-d	47	kōv	perv	p-k e'r-ō
23	thich	fuch	f-th u-i	48	zēt	zēf	f-t
24	wif	wi	f inserted	49	lan	✓	
25	ez	✓		50	shōl	✓	

E.M.B TE.A

TABLE III.

The correct word is written opposite all of the syllables which were recorded incorrectly. The errors for each of the fundamental sounds were taken from this original sheet and recorded on an analysis

TABLE IV.

Summary Sheet — Average Errors above 5 %

TRANSMISSION BRANCH

ARTICULATION TEST ANALYSIS SHEET

TITLE OF TEST 320311

CONDITION TESTED Low Pass Filter — 1250 ~ Attenuation = 5 napiers down

DATE 6-22-20 OBSERVERS 8 Average of

LIST Nos. CALLERS 2

TEST Nos.

SOUNDS CALLED	SOUNDS RECORDED AS																					Omissions	Insertions	% Error
	b	ch	d	f	g	h	j	k	l	m	n	ng	p	r	s	sh	th	t	v	w	y	z			
b			10.3																				4.6		21.2
ch						4.9		8.7							3.2	7.3	4.3	24.7							54.4
d	7.8				11.9		3.3																		35.0
f						3.3									24.0	11.7	19.0						3.9		65.8
g			30.6				3.9																		38.4
h													4.2												13.5
j			15.3		4.0																				29.9
k													3.8					33.7							48.0
l																									3.9
m											2.9														6.8
n										31.7															36.4
ng																									3.6
p		3.0						24.4										22.2							55.0
r									10.4														5.5		18.5
s				6.2												21.5	24.4	7.0							64.9
sh				3.6											27.3		13.0								53.2
th		3.4		9.6											21.1	8.3		7.0				10.5	3.1		70.7
t		7.3				3.4		15.5					6.2												42.4
v																	10.7					15.1			32.9
w																									1.0
y														9.3											16.2
z			6.2				4.2									6.2	8.8		7.9						39.8

No. of times each sound is called

Total number of sounds called

Total number of errors

Consonant Articulation 65.8

Letter Articulation 78.2

Word Articulation 41.2

Summary Sheet — Average Errors above 2 %

TRANSMISSION BRANCH

ARTICULATION TEST ANALYSIS SHEET

TITLE OF TEST 320311

CONDITION TESTED Low Pass Filter — 1250 ~ Attenuation = 5 napiers down Average of

DATE 6-22-20 OBSERVERS 8

LIST Nos. CALLERS 2

TEST Nos.

Sounds Called	No. of Times Called	SOUNDS RECORDED AS															% ERROR
		a	ā	á	e	ē	ér	i	ī	o	ō	ó	u	ū	ou		
a				10.2													13.1
ā																	7.2
á		13.8			4.3				2.2			3.8					24.9
e			2.4	2.3						10.0							17.5
ē																	2.9
ér		3.2	38.8	2.7	4.4					3.6	3.8						60.3
i													5.3				8.4
ī																	4.2
o		4.6			24.2												31.9
ō																	3.9
ó		5.4								2.2							10.2
u								26.1		4.8							34.6
ū													2.8				11.8
ou																	5.6

Total number of sounds called

Total number of errors

Vowel Articulation 83.4

Letter Articulation 72.2

Word Articulation 41.2

sheet as shown in Table IV, for example it will be noticed that p was recorded as k 24.4 per cent, as p 45 per cent, and as t 22.2 per cent of the times called. On the other hand the sound w was only recorded incorrectly 1 per cent of the times called.

For this system the consonant articulation was 65.8 and the vowel articulation 83.4.

Description of the System for Reproducing Speech Sounds

The telephone system used in this investigation is probably more nearly perfect than any other which has yet been built. Its essential elements are a condenser transmitter to receive the speech waves and transform them into the electrical form, an amplifier for magnifying the intensity of the electrical speech currents, an attenuator for controlling the intensity, an equalizing network, and a receiver for delivering the speech to the ear. A schematic arrangement of the circuit is shown in Fig. 1.

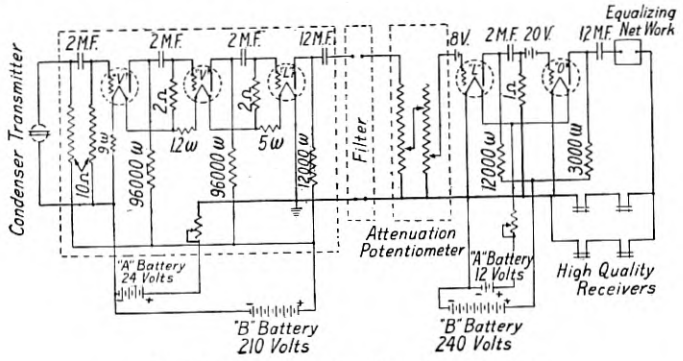


Fig. 1.—High Quality Telephone System

A detailed description of the construction and operation of the condenser transmitter has been given by Crandall and Wente and published in the *Physical Review*.[1] It is simply an air condenser, one of its plates being a flexible metal diaphragm.

A five-stage vacuum tube amplifier was used. Particular care was taken in coupling the stages together, so that the amplifier was practically free from frequency distortion.

The attenuator consisted of a potentiometer arrangement which could reduce the amplitude of the speech waves to approximately one-millionth of their maximum values.

The equalizing network was an arrangement of resistances, con-

[1] Crandall, *Phys. Rev.*, June, 1918; Wente, *Phys. Rev.*, July, 1917.

densers and inductance coils having a frequency selectivity which was the complement of that of the rest of the system.

The telephone receiver was a bipolar type having a special construction which was designed to broaden the range of frequency response.

The reproducing efficiency of the system from the mouth of the speaker to the ear of the listener for each frequency is shown in Fig. 2.

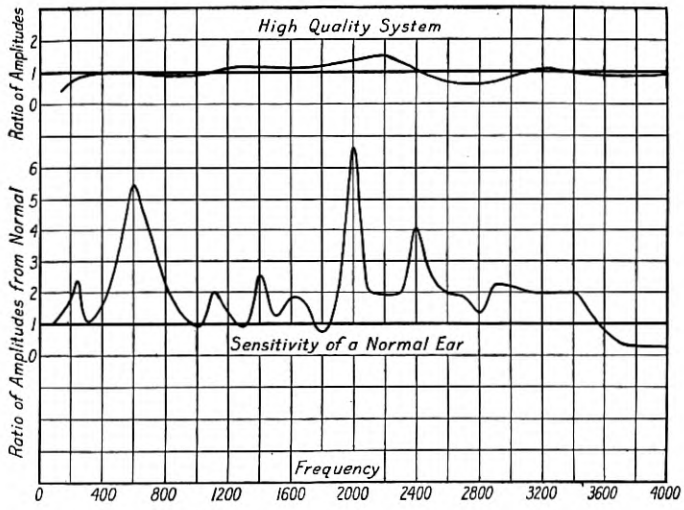


FIG. 2.

The pitch or frequency of the tone is given on the X axis. The ordinates represent amplitude ratios or the number of times the amplitude of the tone reaching the ear was greater than that which entered the transmitter. It will be seen that this high quality system has practically a uniform response for all frequencies throughout the speech range.

In order that its uniformity may be appreciated, a comparison curve is given. This curve shows the deviation in the sensitivity of a typical individual ear from the average sensitivity of a large number of ears. The ordinates represent the ratio of amplitudes at the various pitches which was necessary to bring the tone to the threshold of audibility. It is evident that this deviation is much larger than the departure of the high quality circuit from uniformity.

To show that this particular individual's curve is typical, the curves for both ears of 20 women are given in Fig 3. For convenience these curves are plotted on logarithmic paper. If an arithmetic

scale is used, all of the curves below the mean are crowded together in the small space between zero and one, and all those above the mean are stretched out from one to infinity. By using a logarithmic plot a symmetrical distribution is obtained. The method of obtain-

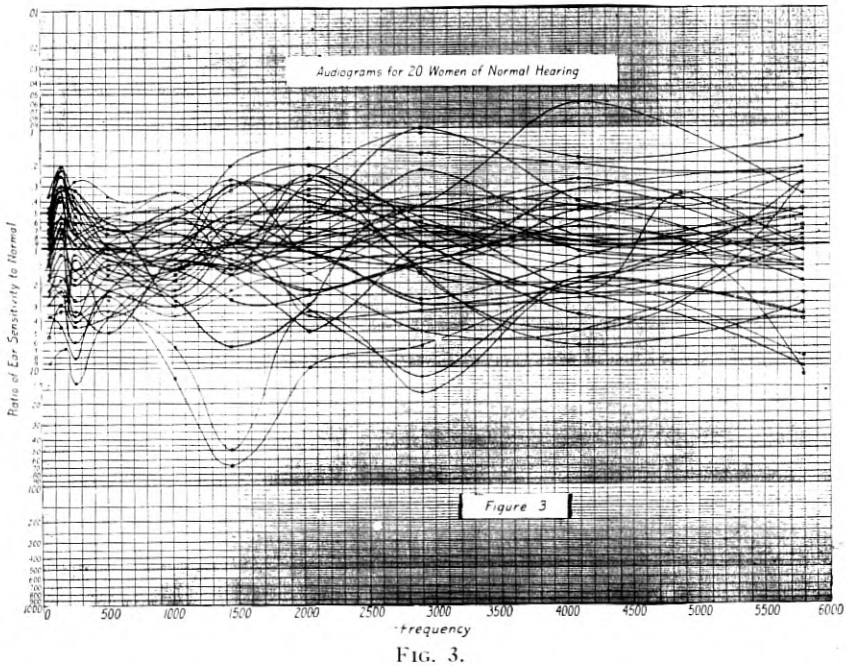


FIG. 3.

ing these ear sensitivity curves was fully described in a recent paper[2] given before the Natural Academy of Sciences.

It is interesting to note that they indicate that each individual has a hearing characteristic which is quite different from other individuals. Consequently speech sounds differently to different persons. Any distortions of the speech sounds will necessarily affect some persons differently from others. It is evident then that in discussing speech and hearing we must deal with statistical averages.

Experimental articulation tests showed that the ear interpreted the speech which was transmitted over this high quality system practically as well as that transmitted through the air. Some may wonder why such good quality is not furnished telephone users in commercial practice: Scientifically speaking, it is possible to furnish such quality, but it is evident that the equipment involved is so com-

[2] Fletcher and Wegel, *Proc. Nat. Acad. Science*, Vol. 8, No. 1, pp. 5–6, Jan., 1922.

plicated that such service would be altogether too costly for commercial use; people could not afford to pay for it.

The Relation Between the Volume and Articulation of Undistorted Speech

Articulation tests were made upon the high quality telephone system described above when it was set to deliver various intensities from the threshold of audibility to very large values. The results shown as syllable articulation values are given by the curve

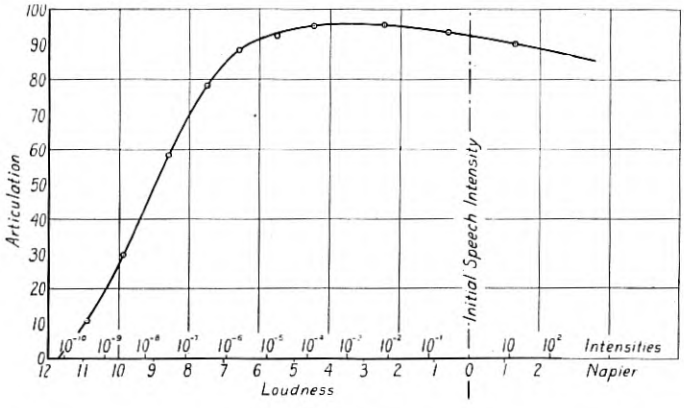


Fig. 4.

in Fig. 4. The abscissas in this curve represent loudness and are expressed as the natural logarithm of the number of times the speech wave amplitude has been decreased from the initial intensity at ½ inch in front of the mouth of the callers. This unit of loudness has never been given a name, and as a matter of convenience in this work it is called a napier. It will be noticed that when the volume is reduced 11½ napiers below the initial speech intensity the articulation becomes zero. This point also represents the value at which the speech becomes inaudible and corresponds to approximately 1/1000 dynes per square centimetre pressure variation against the ear drum. In energy units it is a reduction of ten billion times below the initial speech intensity. For very loud initial speech this point is shifted about 1 napier. For purposes of comparison the intensity reductions are also indicated on the loudness axis.

At 3 napiers below or at about 1/1000 of the initial speech intensity the articulation becomes a maximum. Louder speech than this seems to deaden the nerves so that a person makes a less accurate

interpretation of the received speech. These results were obtained in a room which was especially constructed to exclude outside noise. When noise is present at the receiving station the optimum loudness increases as the noise increases.

The articulation data were analyzed so as to show the errors of each of the fundamental sounds. The curves given in Fig. 5 show the results of this analysis. It will be noticed that the volume at which errors begin to be appreciable is different for the different sounds and is usually higher for the consonants than for the vowels.

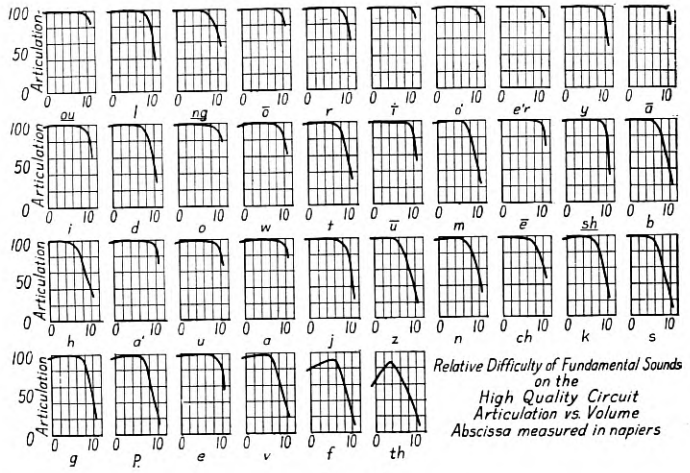


FIG. 5.

Within the precision of the test the intersection point on the X axis was the same for all the sounds, namely at 11.5 napiers.

It will be noticed that the consonants are usually harder to hear than the vowels. However, the speech sounds e and l, r, ng form notable exceptions to this general rule, since the former is among the most difficult, while the latter are among the very easiest speech sounds. The order in which the speech sounds are given here represents their relative difficulty of interpretation when received at average intensities. At all intensities, the sounds th, f and v are the most difficult. Z, h and s become very difficult at weak volumes. The sounds i, ou, er and ó are missed less than 10 per cent of the time, even with "very weak" intensity. At "average" volumes there are only three sounds more difficult than e while at "very weak" volumes there are 23 sounds more difficult. At very weak volumes l, which is the easiest sound at "average" volumes is missed three times as often as e.

We will now pass to a consideration of the effect of distortion upon the articulation of the sounds.

Description of Electrical Filters Used to Produce Distortion

In order to investigate distortion we would like to be able to take the train of speech waves going from the mouth to the ear and operate upon it in various ways such as eliminating frequencies in certain regions without marring or disturbing other frequencies. For ex-

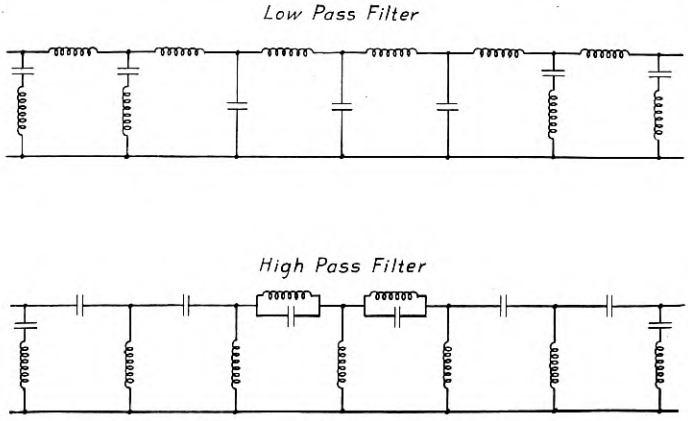


Fig. 6.

ample, if all frequencies above 1000 were eliminated, it would be possible to determine what intelligibility is carried by this range of frequencies.

Fortunately one of the recent electrical inventions is admirably adapted for this purpose, namely, the electrical wave filter invented by Dr. G. A. Campbell. This device was used extensively in this investigation.

The schematic circuit diagrams of the two types of filters which were used are given in Fig. 6.

This arrangement of coils and condensers produces an electrical conductor with the unusual properties that it transmits without appreciable diminution in amplitude any frequency between certain limits and reduces the amplitude of all frequencies outside these limits to less than 1/1000 of their original value. By varying the numerical values of the inductances and capacities this transmitted range can be placed at any desired position. In the arrangement which was used in the investigation these coils and condensers were

housed in two boxes. The switching mechanism was arranged so that by turning a dial the condensers and coils were connected in such a way that the filter transmitted different frequency bands.

In Fig. 7 are shown the transmission properties of the low pass filter when the dial is set to transmit frequencies from 0 to 1500. It is seen that for frequencies below 1400 the amplitudes of the transmitted tones are always greater than .8 of their initial values, while for frequencies above 1500 the amplitudes are decreased to less than .001 of their initial values. These electrical filters were connected into the high quality circuit between the third and fourth stages

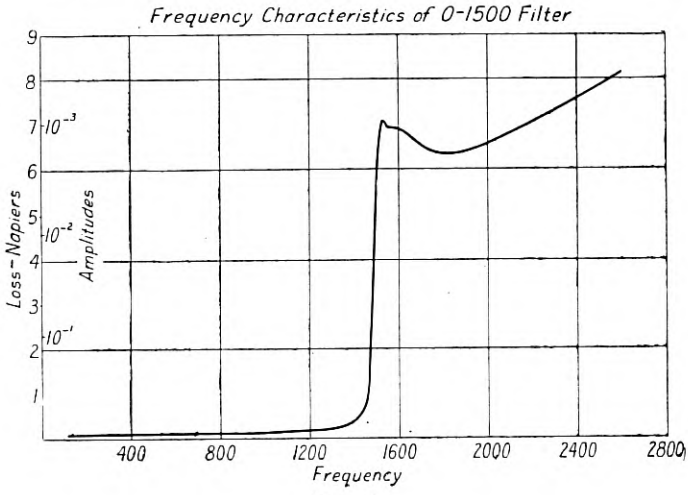


FIG. 7.

of the amplifier as indicated in Fig. 1. This combination formed a system which would pick up a complex sound wave and transmit faithfully to the ear those component frequencies in any desired region and eliminate all other frequencies.

RESULTS OF ARTICULATION TESTS WITH FILTER SYSTEMS

Articulation tests were made with these filter systems and the results analyzed as described above. In Fig. 8 the syllable articulation results are shown in graphical form. The ordinates for the solid curves represent the per cent of the articulation syllables called into the system which were correctly recorded at the observing end. The abscissas represent the so-called "cut off" frequency of the filter. For example on the curve labelled "Articulation L" the point (1000, 40) means that a system which transmits only frequencies

below 1000 cycles per second has a syllable articulation of 40 per cent. Similarly on the curve labelled "Articulation H" the point (1000, 86) means that a system which transmits only frequencies above 1000 cycles per second has a syllable articulation of 86 per cent. The dotted curves show the per cent of the total speech energy which is transmitted through the filter systems used in the articulation tests. These curves are derived from the results of Crandall and MacKenzie which were recently published.[3]

It will be seen that although the fundamental cord tones with their first few harmonies carry a large portion of the speech energy,

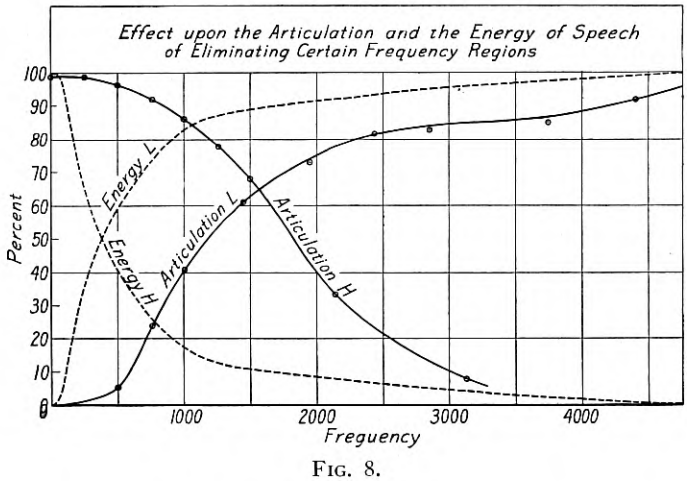


FIG. 8.

they carry practically none of the speech articulation. A filter system which eliminates all frequencies below 500 cycles per second eliminates 60 per cent of the energy in speech; but only reduces the articulation 2 per cent. A system which eliminates frequencies above 1500 cycles per second eliminates only 10 per cent of the speech energy, but reduces the articulation 35 per cent. A system which eliminates all frequencies above 3000 cycles per second has as low a value for the articulation as one which eliminates all frequencies below 1000 cycles per second. This last statement may appear rather astonishing since it is contrary to the popular notion of the relative importance of various voice frequencies from an interpretation standpoint.

The two solid curves intersect on the 1550 cycle abscissa and at 65 per cent articulation, which shows that using only frequencies

[3] See preceding paper.

above or frequencies below 1550 cycles an articulation of 65 per cent will be obtained. The two dotted curves necessarily intersect at 50 per cent.

The curves in Fig. 9 show how the articulation of some of the fundamental speech sounds was affected by eliminating certain frequency regions. The ordinate gives the number of times the sound was written correctly per 100 times called. As in Fig. 8 the

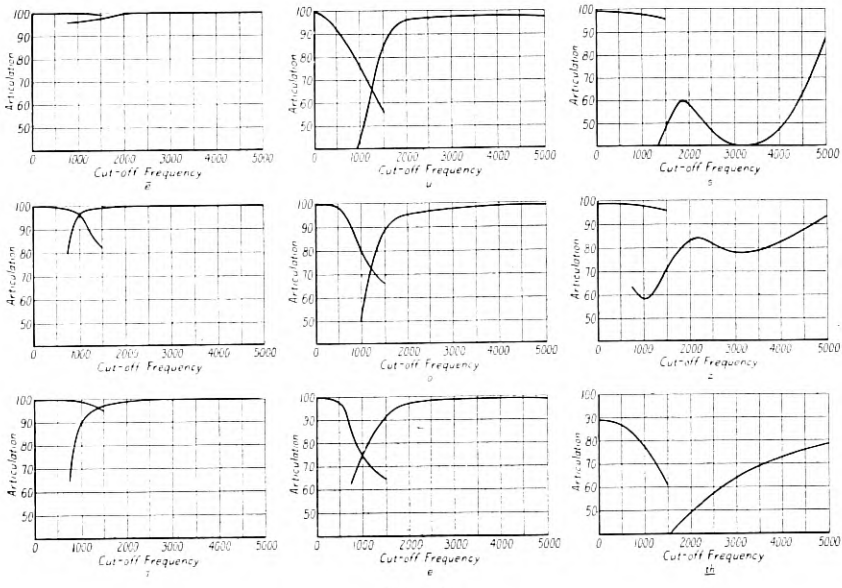


FIG. 9.

left hand curve shows the effect of eliminating all frequencies below and the right hand curve the effect of eliminating all frequencies above the frequency specified by the abscissa.

These nine speech sounds were chosen as representing three important classes. It is seen that the long vowels ē, l and ī can be transmitted with an error of less than 3 per cent when using either half of the range of frequencies. When using either frequencies from 0 to 1700 or from 1700 to infinity ē was interpreted correctly 98 per cent of the time. Similarly l was interpreted correctly 97 per cent of the time when using either the range from 0 to 1000 or 1000 to infinity, and ī 96 per cent of the time when using either the range from 0 to 1350 or from 1350 to infinity. The short vowels, u, o and e are seen to have important characteristics carried by frequencies below 1000. More than a 20 per cent error is made on any of these

three sounds when frequencies below 1000 are eliminated. The elimination of frequencies above 2000 produces almost no effect.

The fricative consonants s, z and th are seen to be affected very differently from those in the other two classes. These sounds are very definitely affected when frequencies above 5000 are eliminated. The sounds s and z are not affected by the elimination frequencies below 1500. It is principally due to these three sounds that the syllable articulation is reduced from 98 per cent to 82 per cent when frequencies above 2500 cycles are eliminated.

A more detailed analysis of the articulation results on all the speech sounds showing the kind as well as the number of errors will be given in a future paper.

Conclusion

In conclusion then we see that the intensity of undistorted speech which is received by the ear can be varied from 100 times greater to one-millionth less than the initial speech intensity without noticeably affecting its interpretation. The intensity must be reduced to one-ten-billionth of that initial speech intensity to reach the threshold of audibility for the average ear. Also it is seen that any apparatus designed to reproduce speech and preserve all of its characteristic qualities must transmit frequencies from 100 to above 5000 cycles with approximately the same efficiency. Although most of the energy in speech is carried by frequencies below 1000, the essential characteristics which determine its interpretation are carried mostly by frequencies above 1000 cycles. In ordinary conversation the sounds th, f and v are the most difficult to hear and are responsible for 50 per cent of the mistakes of interpretation. The characteristics of these sounds are carried principally by the very high frequencies.

It is evident that progress in the knowledge of speech and hearing has a great human interest. It will greatly aid the linguists, the actors, and the medical specialists. It may lead to improved devices which will alleviate the handicaps of deaf and dumb persons. Furthermore this knowledge will be of great importance to the telephone engineer, and since the telephone is so universally used, any improvement in its quality will be for the public good.

These humanitarian and utilitarian motives as well as the pure scientific interest have already attracted a number of scientists to this field. Now that new and powerful tools are available, it is expected that in the near future more will be led to pursue research along those lines.

The Contributors to this Issue

WILLIAM WILSON, Victoria University of Manchester, 1904–10; M.Sc., 1908; Cavendish Laboratory, Cambridge University, 1910–12, B.A., 1912; Lecturer in Physics, Toronto University, 1912–14; D.Sc. Manchester, 1913. Engineering Department Western Electric Company, 1914– . Dr. Wilson has published numerous papers on radio activity and thermionics and since 1917 has been in direct charge of vacuum tube design.

GEORGE A. CAMPBELL, B.S., Massachusetts Institute of Technology, 1891; A.B., Harvard, 1892; Ph.D., 1901; Göttingen, Vienna and Paris, 1893–96. Mechanical Department, American Bell Telephone Company, 1897; Engineering Department, American Telephone and Telegraph Company, 1903-1919; Department of Development and Research, 1919—; Research Engineer, 1908—. Dr. Campbell has published papers on loading and the theory of electric circuits and is also well-known to telephone engineers for his contributions to repeater and substation circuits. The electric filter which is one of his inventions plays a fundamental role in telephone repeater, carrier current and radio systems.

H. M. TRUEBLOOD, B.S., Earlham, 1902; Haverford, 1903; Massachusetts Institute of Technology, 1908–09; Ph.D., Harvard, 1913; aid and assistant United States Coast and Geodetic Survey, 1903–08; assistant in physics, Harvard, 1912–14; Joule-Thomson effect in super-heated steam; instructor and assistant professor electrical engineering, University of Pennsylvania, 1914–17; Department of Development and Research, American Telephone and Telegraph Comcompany, 1917—; work on inductive interference.

J. J. PILLIOD, E.E., Ohio Northern University, 1908; American Telephone and Telegraph Company, Toledo Home Telephone Company, and Union Switch and Signal Company, short periods, 1904–08; American Telephone and Telegraph Company, Long Lines Department, 1908–11; Engineering Department, 1912–13; Division Plant Engineer, Long Lines Department, 1914–17; Engineer of Transmission, 1918–19; Engineer, 1920–. As Engineer of the Long Lines Department, Mr. Pilliod has been in general charge of engineering work involved in the planning and installation of the newer sections of the cable project described.

JOHN R. CARSON, B.S., Princeton, 1907; E.E., 1909; M.S., 1912; Research Department, Westinghouse Electric and Manufacturing

Company, 1910–12; instructor of physics and electrical engineering, Princeton, 1912–14; American Telephone and Telegraph Company, Engeneering Department, 1914–15; Patent Department, 1916–17; Engineering Department, 1918; Department of Development and Research, 1919–. Mr. Carson's work has been along theoretical lines and he has published several papers on theory of electric circuits and electric wave propagation.

J. J. GILBERT, A.B., University of Pennsylvania, 1909; Harvard, 1910–11; Chicago, 1911–12; E.E., Armour Institute, 1915; instructor of electrical engineering, Armour, 1912–17; Captain Signal Corps, 1917–19; Engineering Department, Western Electric Company, 1919, since when he has worked on submarine cable problems.

I. B. CRANDALL, A.B., Wisconsin, 1909; A.M., Princeton, 1910; Ph.D., 1916; Professor of Physics and Chemistry, Chekiang Provincial College, 1911–12; Engineering Department, Western Electric Company, 1913–. Dr. Crandall has published papers on infra-red optical properties, condenser transmitter, thermophone, etc. More recently he has been associated with studies on the nature and analysis of speech which have been in progress in the Laboratory.

DONALD MACKENZIE, A.B., Johns Hopkins, 1908; A.M., 1911; Ph.D., 1914; assistant astronomy, 1914–17; associate physicist, Bureau of Standards, 1918–20; Engineering Department, Western Electric Company, 1920–.

HARVEY FLETCHER, B.S., Brigham Young, 1907; Ph.D., Chicago, 1911; instructor of physics, Brigham Young, 1907–08; Chicago, 1909–10; Professor, Brigham Young, 1911–16; Engineering Department, Western Electric Company, 1916–. The present paper by Dr. Fletcher gives some of the results of an investigation which is being made of the relation between the frequency characteristics of telephone circuits and the intelligibility of transmitted speech. Dr. Fletcher has also published on Brownian movements, ionization and electronics.

The Bell System Technical Journal

Devoted to the Scientific and Engineering Aspects of Electrical Communication

Published quarterly by the American Telephone and Telegraph Company, through its Information Department, in behalf of the Western Electric Company and the Associated Companies of the Bell System

Address all correspondence to the Editor
Information Department
AMERICAN TELEPHONE AND TELEGRAPH COMPANY
195 BROADWAY, NEW YORK, N. Y.

50c. Per Copy *$1.50 Per Year*

Vol. I NOVEMBER, 1922 No. 2

Physical Theory of the Electric Wave-Filter

By GEORGE A. CAMPBELL

NOTE: The electric wave-filter, an invention of Dr. Campbell, is one of the most important of present day circuit developments, being indispensable in many branches of electrical communication. It makes possible the separation of a broad band of frequencies into narrow bands in any desired manner, and as will be gathered from the present article, it effects the separation much more sharply than do tuned circuits. As the communication art develops, the need will arise to transmit a growing number of telephone and telegraph messages on a given pair of line wires and a growing number of radio messages through the ether, and the filter will prove increasingly useful in coping with this situation. The filter stands beside the vacuum tube as one of the two devices making carrier telegraphy and telephony practicable, being used in standard carrier equipment to separate the various carrier frequencies. It is a part of every telephone repeater set, cutting out and preventing the amplification of extreme line frequencies for which the line is not accurately balanced by its balancing network. It is being applied to certain types of composited lines for the separation of the d.c. Morse channels from the telephone channel. It is finding many applications to radio of which multiplex radio is an illustration. The filter is also being put to numerous uses in the research laboratory.

The present paper is the first of a series on the electric wave-filter to be contributed to the Technical Journal by various authors. Being an introductory paper the author has chosen to discuss his subject from a physical rather than mathematical point of view, the fundamental characteristics of filters being deduced by purely physical reasoning and the derivation of formulas being left to a mathematical appendix.—*Editor.*

THE purpose of this paper is to present an elementary, physical explanation of the wave-filter as a device for separating sinusoidal electrical currents of different frequencies. The discussion

will be general, and will not involve assumptions as to the detailed construction of the wave-filter; but in order to secure a certain numerical concreteness, curves for some simple wave-filters will be included. The formulas employed in calculating these curves are special cases of the general formulas for the wave-filters which are, in conclusion, deduced by the method employed in the physical theory.

All the physical facts which are to be presented in this paper, together with many others, are implicitly contained in the compact formulas of the appendix. Although only comparatively few words of explanation are required to derive these formulas, they will not be presented at the start, since the path of least resistance is to rely implicitly upon formulas for results, and ignore the troublesome question as to the physical explanation of the wave-filter. In order to examine directly the nature of the wave-filter in itself, as a physical structure, we proceed as though these formulas did not exist.

It is intended that the present paper shall serve as an introduction to important papers by others in which such subjects as transients on wave-filters, specialized types of wave-filters, and the practical design of the most efficient types of wave-filters will be discussed.[1]

Definition of Wave-Filter

A wave-filter is a device for separating waves characterized by a difference in frequency. Thus, the wave-filter differentiates between certain states of motion and not between certain kinds of matter, as does the ordinary filter. One form of wave-filter which is well known is the color screen which passes only certain bands of light frequencies; diffraction gratings and Lippmann color photographs also filter light. Wave-filters might be constructed and employed for separating air waves, water waves, or waves in solids. This paper will consider only the filtering of electric waves; the same principles apply in every case, however.

In its usual form the electric wave-filter transmits currents of all frequencies lying within one or more specified ranges, and excludes currents of all other frequencies, but does not absorb the energy of these excluded frequencies. Hence, a combination of two or more wave-filters may be employed where it is desired to separate a broad band of frequencies, so that each of several receiving devices is sup-

[1] I take pleasure in acknowledging my indebtedness to Mr. O. J. Zobel for specific suggestions, and for the light thrown on the whole subject of wave-filters by his introduction of substitutions which change the propagation constant without changing the iterative impedance.

plied with its assigned narrower range of frequencies. Thus, for instance, with three wave-filters the band of frequencies necessary for ordinary telephony might be transmitted to one receiving device, all lower frequencies transmitted to a second device, and all higher frequencies transmitted to a third device—separation being made without serious loss of energy in any one of the three bands.

By means of wave-filters interference between different circuits or channels of communication in telephony and telegraphy, both wire and radio, can be reduced provided they operate at different frequencies. The method is furthermore applicable, at least theoretically, to the reduction of interference between power and communication circuits. The same is true of the simultaneous use of the ether, the earth return, and of expensive pieces of apparatus employed for several power or communication purposes. In all cases the principle involved is the same as that of confining the transmission in each circuit or channel to those frequencies which serve a useful purpose therein and excluding or suppressing the transmission of all other frequencies. In the future, as the utility of electrical applications becomes more widely and completely appreciated, there will be an imperative necessity for more and more completely superposing the varied applications of electricity; it will then be necessary, to avoid interference, to make the utmost use of every method of separating frequencies including balancing, tuning, and the use of wave-filters.

Definition of Artificial Line

The wave-filter problem in this paper is discussed as a phase of the artificial line problem, and it is desirable to start with a somewhat generalized definition of the artificial line. The definition will, however, not include all wave-filters or all artificial lines, since a perfectly general definition is not called for here. Even if an artificial line is to be, under certain wave conditions, an imitation of, or a substitute for, an actual line connecting distant points, hardly any limitation is thereby imposed upon the structure of the device; an actual line need not be uniform but may vary abruptly or gradually along its length and may include two, three, four or more transmission conductors of which one may be the earth. Having indicated that wave-filters partake of somewhat this same generality of structure, the present paper is restricted to wave-filters coming under the somewhat generalized artificial line specified by the following definition:

An artificial line is a chain of networks connected together in sequence through two pairs of terminals, the networks being identical but other-

wise unrestricted. This generalized artificial line possesses the well-known sectional artificial line structure but it need not be an imitation of, or a substitute for, any known, real, transmission line connecting together distant points. The general artificial line is shown by Fig. 1 where N, N,. . . .are the identical unrestricted networks which may contain resistance, self-inductance, mutual inductance, and capacity.

In discussing this type of structure as a wave-filter, the point of view of an artificial line is adopted for the reason that it is advantageous to regard the distribution of alternating currents as being dependent upon both propagation and terminal conditions, which are to be separately considered. In this way the attenuation, or

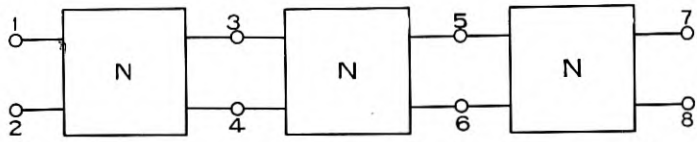


Fig. 1—Generalized Artificial Line as Considered in the Present Paper, where N, N, . . . are Identical Arbitrary Electrical Networks

falling off, of the current from section to section may be most directly studied. Terminal effects are not to be ignored, but are allowed for, after the desired attenuation effects have been secured, possibly by an increase in the number of sections to be employed.

The fundamental property of this generalized artificial line, which includes uniform lines as a special case, is the mode in which the wave motion changes from one section to the next, and may be stated as follows:

Wave Propagation Theorem

Upon an infinite artificial line a steady forced sinusoidal disturbance falls off exponentially from one section to the next, while the phase changes by a constant amount. Reversing the direction of propagation does not alter either the attenuation or phase change. When complex quantities are employed the exponential includes the phase change.[2] This theorem is proved, without mathematical equations, by observing

[2] This theorem is not new, but it is ordinarily derived by means of differential or difference equations whereas it may be derived from the most elementary general considerations, thus avoiding all necessity of using differential or difference equations, as illustrated in my paper "On Loaded Lines in Telephonic Transmission" (*Phil. Mag.*, vol. 5, pp. 313–331, 1903). In that discussion, as well as in this present one, it is tacitly assumed that the line is either an actual line with resistance, or the limit of such a line as the resistance vanishes, so that the amplitude of the wave never increases towards the far end of an infinite line.

that the percentage reduction in amplitude and the change in phase, in passing from the end of one section to the corresponding point of the next section, do not depend upon either the absolute amplitude or phase; they depend, instead, only upon the magnitudes, angles and interconnections of the impedances between the two points and of the impedances beyond the second point. These impedances are, since the line is assumed to be periodic and infinite, identically the same for corresponding points between all sections of the line, and, therefore, the relative changes in the wave will be identical at corresponding points in all sections. This proves the exponential falling off of the disturbance and the constancy of phase change; the ordinary reciprocal property shows that the wave will fall off identically whichever be the direction of propagation. By the superposition property it follows that the steady state on any finite portion of a periodic recurrent structure must be the sum of two equally attenuated disturbances, one propagated in each direction.

The fundamental wave propagation theorem may be generalized for any periodic recurrent structure irrespective of the number and kind of connections between periodic sections, provided the disturbance is such as to remain similar to itself at corresponding points of each of these connections.

Equivalent Generalized Artificial Line

Since, at a given frequency, any network employed solely to connect a pair of input terminals with a pair of output terminals may be replaced by either three star-connected impedances or three delta-connected impedances, the general artificial line of Fig. 1 may be

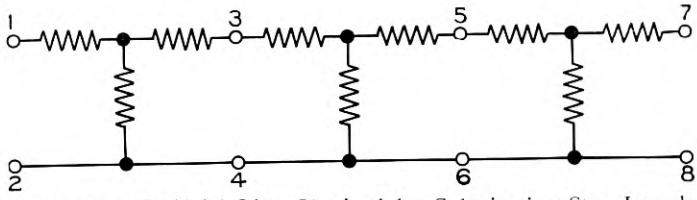


Fig. 2—Equivalent Artificial Line Obtained by Substituting Star Impedances

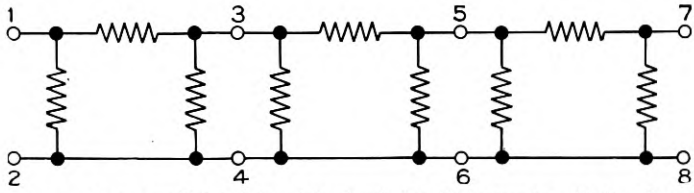


Fig. 3—Equivalent Artificial Line Obtained by Substituting Delta Impedances

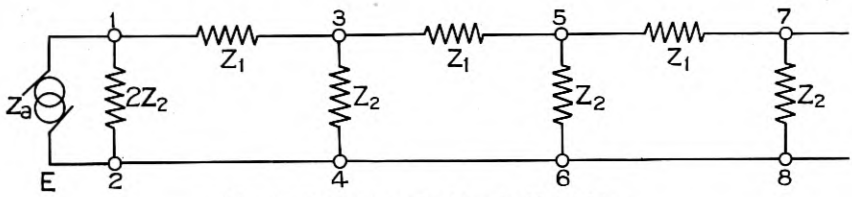


Fig. 4—Equivalent Ladder Artificial Line

replaced by the equivalent artificial line of either Fig. 2 or Fig. 3. By combining the series impedances in Fig. 2 and the parallel impedances in Fig. 3, the equivalent line in Fig. 4 is obtainable. The two ways of arriving at Fig. 4 give different values for the series and shunt impedances Z_1, Z_2, and different terminations for the line, but the propagation of the wave is the same in both cases, since the assumed substitutions are rigorously exact. While Fig. 4 may be considered as the generalized artificial line equivalent to Fig. 1, this requires including in Z_1 and Z_2 impedances which cannot always be physically realized by means of two entirely independent networks, one of which gives Z_1 and the other Z_2. This restriction is of no importance when we are discussing the behavior of the generalized artificial line at a single frequency; accordingly, the ladder artificial line is suitable for this part of the discussion. When we come to the more specific correlation of the behavior of the generalized artificial line at different frequencies, it will be found more convenient to replace the ladder artificial line by the lattice artificial line, which avoids the necessity of considering any impedances which are not individually physically realizable.

The equivalence between Figs. 1 and 4 is implicitly based upon the assumption that it is immaterial, for artificial line uses, what absolute potentials the terminals 1, 2; 3, 4; 5, 6; etc. have—this leaves us at liberty to connect 2, 4, 6, etc., together, so long as we maintain unchanged the differences in potential between 1 and 2, 3 and 4, etc. Instead of connecting 2 and 4 we might equally well connect 2 and 3, and then Z_1 would connect 1 and 4 as in Fig. 5; with these

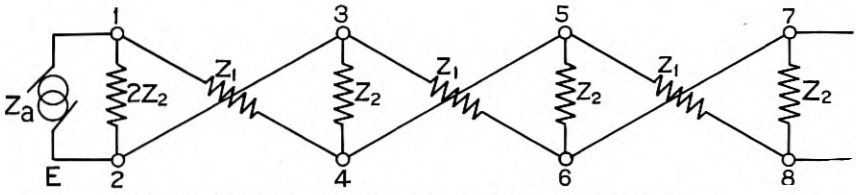


Fig. 5—Equivalent Artificial Line with Crossed Impedances

cross-connections the propagation still remains unchanged. We have again obtained Fig. 4 with no circuit difference except the interchange of terminals 3 and 7 with terminals 4 and 8; or, if this is ignored, a reversal in the sign of the current at alternate pairs of terminals. This shows that the reversal of the current in alternate sections of Fig. 4 may not be of primary significance, since networks which are essentially equivalent have reversed currents.

In order to deal, at the start, with only the simpler terminal conditions, we may consider the line to begin with only one-half of the series impedance Z_1, or only one-half of the bridged admittance $1/Z_2$. These mid-points are called the mid-series and mid-shunt points; knowing the results of termination at either of these points, the effect of termination at any other point may be readily determined. For Fig. 4 termination at mid-shunt has been chosen so that each section of the line adds a complete symmetrical mesh to the network.

An alternator, introducing an impedance Z_a, is shown as the source of the steady-state sinusoidal current in Fig. 4. Assume that the impedance Z_a is variable at pleasure, and that it is gradually adjusted to make the total impedance in the generator circuit vanish,—in this case no e.m.f. will be required to maintain the forced steady-state which becomes a free oscillation. If, in addition, it is assumed that the line has an infinite number of sections, this required value of Z_a will be the negative of the mid-shunt iterative impedance[3] of the artificial line, which will be designated as K_2. The first shunt on the line now includes $-K_2$ in parallel with $2Z_2$ so that its total impedance is, say, $Z' = -2Z_2K_2/(2Z_2-K_2)$. The infinite line with its first shunt given the special value Z' is thus capable of free oscillation.

It is possible to simplify this infinite oscillating circuit by cutting off any part of it which has the same free period as the whole circuit. The entire infinite line beyond the second shunt 3, 4 certainly has this same free period, provided its first shunt also has the impedance Z'. Conceive the shunt Z_2 at 3, 4 as replaced by the four impedances $2Z_2$, $2Z_2$, $+K_2$ and $-K_2$ all in parallel; the first and last, which together make the Z' required by the infinite line, leave $2Z_2$ and

[3] The "iterative impedance" of an artificial line is the impedance which repeats itself when one or more sections of the artificial line are inserted between this impedance and the point of measurement. It is thus the impedance of an infinite length of any actual artificial line, regardless of the termination of the remote end of the line. In general, its value is different for the two directions of propagation, but not when the line is symmetrical, as at mid-series and mid-shunt. The values at these points are denoted by K_1 and K_2. "Iterative impedance" is employed because it is a convenient term which is distinctive and describes the most essential property of this impedance; it seems to be more appropriate than "characteristic impedance," "surge impedance" and the other synonyms in use.

$+K_2$ in parallel, which have the impedance $Z''=+2Z_2K_2/(2Z_2+K_2)$. Removing Z' together with the infinite line on the right there remains on the left a closed circuit made up of the three impedances Z_1, Z' and Z'' in series.

After the division, the infinite line on the right will continue, without modification, to oscillate freely, since it is an exact duplicate of the original oscillating line, and so must maintain the free oscillation already started. Since it oscillates freely by itself, it had originally no reaction upon the simple circuit from which it was separated; this simple circuit on the left must thus also continue its own free oscillations without change in period or phase.

We might continue and subdivide the entire infinite line into identical simple circuits but it is sufficient to consider this one detached circuit, which is shown separately in two ways by Fig. 6, since from

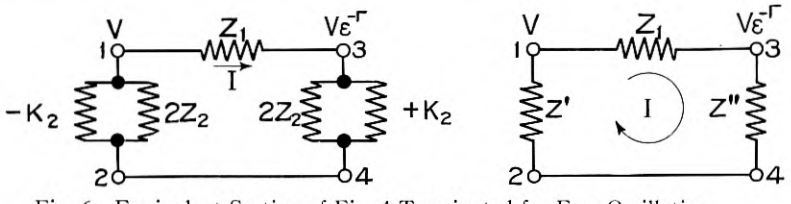


Fig. 6—Equivalent Section of Fig. 4 Terminated for Free Oscillation

its free oscillations the mathematical formulas for the steady-state propagation in the artificial line may be derived. This is deferred, however, until after the physical discussion is completed, so as to leave no room for doubt that the essentials of the physical theory are really deduced without the aid of mathematical formulas.

The generalized artificial line, if made up entirely of pure resistances, will attenuate all frequencies alike, and the entire wave will be in the same phase; this remains true, whatever be the impedance of the individual branch of the network, provided the ratio of the impedances of all branches is a constant independent of the frequency. This is precisely the condition to be avoided in a wave-filter; branches must not be similar but dissimilar as regards the variation of impedance with frequency. This calls for inductance and capacity with negligible resistance, so that there is an opportunity for the positive reactance of one branch to react upon the negative reactance of another branch, in different proportions at different frequencies. Assuming the unit network N of Fig. 1 to be made up of a finite number of pure reactances, the equivalent impedances Z_1 and Z_2 of Figs. 4 and 5 must also be pure reactances. Under this assumption

let us consider the free oscillations of Fig. 6; first, with K_2 assumed to be a pure reactance; second, with K_2 assumed to be a pure resistance; and third, in order to show that this third assumption is contrary to fact, with K_2 assumed to be an impedance with both resistance and reactance.

With K_2 a reactance, the circuit contains nothing but reactances, and free oscillations are possible if, and only if, the total impedance of the circuit is zero. The end impedances Z' and Z'' being different, the potentials at the ends of the mesh will be different, and this means that the corresponding wave on the infinite line will be attenuated, since the ratio between these potentials is the rate at which the amplitudes fall off per section.

With K_2 a pure resistance, a free oscillation is possible only if the dissipation in the positive resistance at the right end of the circuit is exactly made up by the hypothetical source of energy existing in the negative resistance $-K_2$ at the left end of the circuit. An exact balance between the energy supplied at one end and that lost at the other end is possible, since the equal positive and negative resistances K_2, $-K_2$ carry equal currents. This continuous transfer of energy from the left of the oscillating circuit of Fig. 6 to the right end is the action which goes on in every section of the infinite artificial line, and serves to pass forward the energy along the infinite line.

If K_2 were complex, $-K_2$ on the left of Fig. 6 and $+K_2$ on the right would not carry the same fraction of the circulating current I, since they are each shunted by a reactance $2Z_2$ which would allow less of the current to flow through $+K_2$ than through $-K_2$, if $2Z_2$ makes the smaller angle with $+K_2$, and vice versa. No balance between absorbed and dissipated energy is possible under these conditions when the equal and opposite resistance components carry unequal currents. A complex K_2, therefore, gives no free oscillation, and cannot occur with a resistanceless artificial line.

It is perhaps more instructive to consider the transmission on the line as a whole, rather than to confine attention exclusively to the oscillations of the simple circuit of Fig. 6 and so, at this point, without following further the conclusions to be drawn directly from this oscillating circuit, the fundamental energy theorem of resistanceless artificial lines will be stated, and then proved as a property of an infinite artificial line.

Energy Flow Theorem

Upon an infinite line of periodic recurrent structure, and devoid of resistance, a sinusoidal e.m.f. produces one of two steady states, viz.:

1. *A to-and-fro surging of energy without any resultant transfer of energy; currents and potential differences each attenuated from section to section, but everywhere in the same or opposite phase and mutually in quadrature, or,*

2. *A continuous, non-attenuated flow of energy along the line to infinity with no energy surging between symmetrical sections; current and potential non-attenuated, but retarded or advanced in phase from section to section, and mutually in phase at mid-shunt and mid-series points.*

The critical frequencies separating the two states of motion are the totality of the resonant frequencies of the series impedance, the anti-resonant frequencies of the shunt impedance, and the resonant frequencies of a single mid-shunt section of the line.

To prove the several statements of this theorem let us consider first the consequences of assuming that the wave motion, in progressing along the line, is attenuated, and next the consequences of assuming that the wave motion changes its phase. If the wave is attenuated, however little, at a sufficient distance it becomes negligible, and the more remote portions of the line may be completely removed without appreciable effect upon the disturbance in the nearer portion of the line. That part of the line which then remains is a finite network of pure reactances, and in any such network all currents are always in the same, or opposite, phase; so, also, are the potential differences; moreover, the two are mutually in quadrature; there is no continuous accumulation of energy anywhere, but only an exchange of energy back and forth between the inductances, the capacities and the generator. Continuously varying the amount of the assumed attenuation will cause a continuous variation in the corresponding frequency. The motion of the assumed character may, therefore, be expected to occur throughout continuous ranges or bands of frequencies and not merely at isolated frequencies.

The question may be asked—How far does the energy surge? Is the surge localized in the individual section, or does the surge carry the energy back and forth over more than one section, or even in and out of the line as a whole? To answer this question, it would be necessary, as we will now proceed to prove, to know something about the actual construction of the individual section. If each section is actually made up as shown in Fig. 6, and this is entirely possible in the present case (since only positive and negative reactances would be called for), then the section is capable of free oscillation, as explained

above, and the surging is localized within the section; twice during each cycle the amount of energy increases on the right and decreases on the left. But we do not know that the section is made up like Fig. 6; we only know that it is equivalent to Fig. 6 as regards input and output relations. As far as these external relations go, the actual network may be made exclusively of either inductances or capacities with the connections shown in Fig. 4 or with the cross-connections of Fig. 5, according as the current is to have the same or opposite signs in consecutive sections. In any network made up exclusively of inductances or of capacities, the total energy falls to zero when the current or the potential falls to zero, respectively. Twice, therefore, in every cycle the total energy surges into this line and then it all returns to the generator. With other networks, surgings intermediate between these two extremes will occur. The theorem, therefore, does not limit the extent of the surging.

Under the second assumption, the phase difference between the currents at two given points, separated by a periodic interval, is to be an angle which is neither zero nor a multiple of $\pm\pi$. The assumed difference in phase can only be due to the infinite extension of the artificial line since, as previously noted, no finite sequence of inductances and capacities can produce any difference in phase. That infinite lines do produce phase differences is well-known; in particular, an infinite, uniform, perfectly conducting, metallic pair shows a continuous retardation in phase. If the infinitely remote sections of the artificial line are to have this controlling effect on the wave motion, the wave motion must actually extend to infinity, that is, there can be no attenuation. The wave progressing indefinitely to infinity without attenuation must be supplied continuously with energy; this energy must flow along the entire line with neither loss nor gain in the reactances it encounters on the way. This continuous flow of energy can take place only provided the currents and potentials are not in quadrature; they may be in phase. In considering the free oscillations of Fig. 6 it was shown that K_2 is real if it is not pure reactance. That is, for the mid-shunt section the current and potential are in phase. It is easy to show that they are also in phase at the mid-series point which is also a point of symmetry.

This flow-of-energy state of motion thus necessarily characterizes a phase-retarded wave on a resistanceless artificial line, regardless of the amount of the assumed positive or negative retardation, which may be taken to have any value between zero and exact opposition of phase. Continuously varying the retardation throughout the 180 degrees will, in general, call for a continuous change in the frequency

of the wave motion. The second state of motion occurs, therefore, throughout continuous ranges or bands of frequencies.

No other state of motion is possible. With given initial amplitude and phase any possible wave motion is completely defined by its attenuation and phase change. All possible combinations of these two elements have been included in the two states, since the excluded conditions on each assumption have been included as a consequence of the other assumption. Thus, the exclusion of no attenuation in the first assumption was found necessarily to accompany the phase change of the second assumption; currents in phase or opposed, which were excluded from the second assumption, were found to be necessary features accompanying the first assumption. There remains only to consider the critical frequencies separating the two states of motion. At these frequencies there can be no attenuation and lag angles of multiples of $\pm\pi$, including zero, only. At symmetrical points the iterative impedance of the line must be a pure reactance to satisfy the first state of motion, and a pure resistance to satisfy the second state of motion. The only iterative impedances which satisfy these conditions are zero and infinity.

Some details relating to the pass and stop bands and the critical frequencies are brought together in the following table, where "stop ($\pm$)" refers to stop bands, the current being in phase or opposed in successive sections, and where γ and k refer to the line obtained by uniformly distributing $1/Z_2$ with respect to Z_1.

TABLE I.

For Ladder Artificial Line, Fig. 4

Band	Critical Frequency	Ratio $\frac{Z_1}{4Z_2}$	Uniform Line		Artificial Line			
			γ	k	Γ	$e^{-\Gamma}$	K_1	K_2
Stop (+)		>0	+real	imag.	+real	$0<<1$	imag.	imag.
	$Z_1 = 0$ $Z_2 = \infty$	0 0	0 0	0 ∞	0 0	1 1	0 ∞	0 ∞
Pass		$0>>-1$	imag.	+real	imag.	$e^{i\theta}$	+real	+real
	$Z_1 + 4Z_2 = 0$	-1	$i2$	$i2Z_2$	$i\pi$	-1	0	∞
Stop (−)		<-1	imag.	+real	$i\pi$ + real	$-1<<0$	imag.	imag.
	$Z_1 = \infty$ $Z_2 = 0$	$-\infty$ $-\infty$	∞ ∞	∞ 0	∞ ∞	0 0	∞ $\frac{1}{2}Z_1$	$2Z_2$ 0

It is not necessary to check the table item by item, many of which have already been proven, but it will be instructive to check some of the items by assuming that $Z_1/4Z_2$, called the ratio for brevity, is positive to begin with, and that a continuous increase in frequency reduces the ratio to zero and back through $\mp \infty$ to its original positive value. This cycle starts with a stop (+) band since the artificial line is in effect a network of reactances, all of which have the same sign; there is attenuation and the iterative impedances are imaginary. When the ratio decreases to zero, there must be either resonance which makes $Z_1 = 0$, or anti-resonance which makes $Z_2 = \infty$; in either case the artificial line has degenerated into a much simpler circuit; it is a shunt made up of all Z_2's combined in parallel, or a simple series circuit made up of all Z_1's, respectively; the iterative impedances are 0 and ∞, respectively; there is no attenuation in either case.

With a somewhat further increase of the frequency the ratio will assume a small negative value with the result that the artificial line will have both kinetic and potential energy. An analogy now exists between the artificial line and an ordinary uniform transmission line, which possesses both kinetic and potential energy, and is ordinarily visualized as being equivalent to many small positive reactances, in series, bridged, to the return conductor, by large negative reactances. The fact that uniform lines do freely transmit waves is a well-known physical principle, and it is not necessary to repeat here the physical theory of such transmission merely to show that the same phenomenon occurs with the identical structure when it is called an artificial line or wave-filter.

In order to determine just how far the ratio may depart from zero, on the negative side, without losing the property of free transmission, we look for any change in the action of the individual section of the artificial line which is fundamental; nothing less than a fundamental change in the behavior of the individual section can produce such a radical change in the line as an abrupt transition from the free transmission of a pass band to the to-and-fro surging of energy in a stop band. Now as the ratio is made more and more negative by the assumed increase of frequency, the value -1 is reached, at which frequency the symmetrical section (Fig. 6) of the artificial line is capable of free oscillation by itself. This is well recognized as a most fundamental change in the properties of any network, and it affords grounds for expecting a complete change in the character of the propagation over the artificial line. The change must be to a stop band with currents in opposite phase, since at resonance the potentials at the two ends of a section are in opposite phase.

Further increase in the frequency cannot make any change in the absolute difference in phase between the two ends of the other section, since opposition is the greatest possible difference in phase; the wave now adapts itself to increasing frequency by altering its attenuation.

Upon continuing the increase of frequency, so as to reduce the ratio to $-\infty$, we arrive at either anti-resonance corresponding to $Z_1 = \infty$ or resonance corresponding to $Z_2 = 0$; the artificial line has now degenerated into a row of isolated impedances Z_2, or into a series of impedances Z_1 short-circuited to the return wire; in either case the attenuation is infinite since no wave is transmitted. Passing beyond this critical frequency the ratio becomes positive, according to our assumption, and we are again in a stop (+) band.

While in this rapid survey of what happens during this frequency cycle little has been actually proven, it should have been made physically clear why abrupt changes in the character of the transmission occur at the frequencies making the ratio equal to 0, -1 or ∞, since the line degenerates into a simpler structure, or the phase change reaches its absolute maximum, on account of resonance, at these particular frequencies.

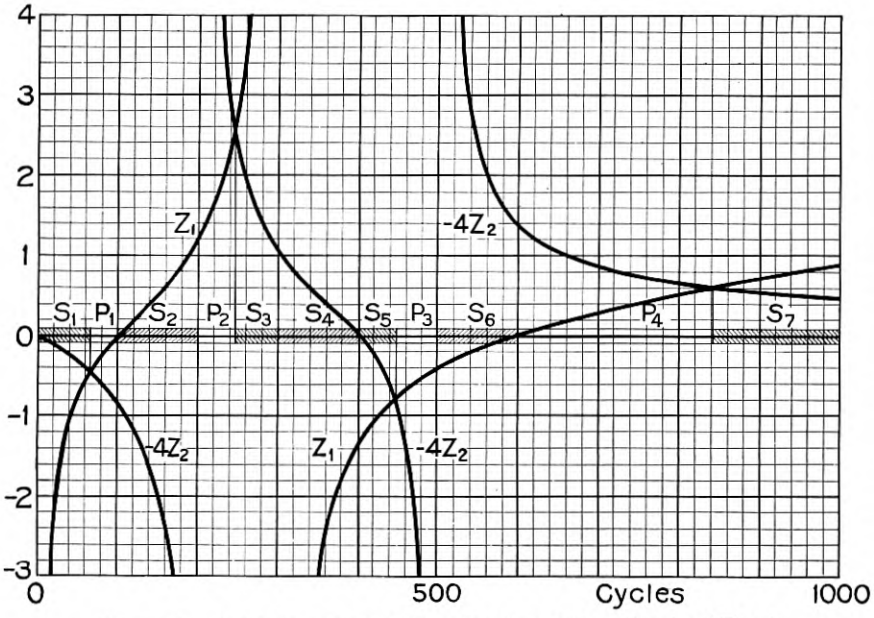


Fig. 7—Graph for Locating the Pass Bands and Stop Bands of Fig. 4
$Z_1 = -i(1^2 - x^2)(6^2 - x^2)/8x(3^2 - x^2)$,
$-4Z_2 = -i4x(4^2 - x^2)/(2^2 - x^2)(5^2 - x^2)$ and x = cycles/100

Information as to the location of the bands is often obtained most readily by plotting both Z_1 and $-4Z_2$, as illustrated in Fig. 7, and determining the critical frequencies by noting where the curves cross each other and the abscissa axis, as well as where they become infinite. Any particular band is then a pass band, a stop (+) band or a stop (−) band, according as Z_1, the abscissa axis, or $-4Z_2$ lies between the other two of the three lines. In Fig. 7 the pass bands are P_1, P_2, P_3, P_4; the stop (+) bands are S_2, S_4, S_6; and the stop (−) bands are S_1, S_3, S_5, S_7, and they illustrate quite a variety of sequences. By altering the curves the bands may be shifted, may be made to coalesce, or may be made to vanish.

Wave-Filter Curves

The pass band and stop band characteristics of wave-filters are concretely illustrated for a few typical cases by the curves of Figs. 8–13, which show the attenuation constant A, the phase constant B, and both the resistance R and reactance X components of the iterative impedance for a range of frequencies which include all of the critical frequencies, except infinity. The heavy curves apply to the ideal resistanceless case, while the dotted curves assume a power factor equal to $1/(20\pi)$ for each inductance which is a value readily obtained in practice. This value is, however, not sufficiently large to make these small scale curves entirely clear, since considerable portions of the dotted curves appear to be coincident with the heavy line curves; but this, as far as it goes, proves the value of the present discussion which rests upon a close approximation of actual wave-filters to the ideal resistanceless case.

The low pass resistanceless wave-filter, as shown by Fig. 8, presents no attenuation below 1,000 cycles; above this frequency the attenuation constant increases rapidly, in fact, the full line attenuation curve increases at the start with maximum rapidity, since it is there at right angles to the axis. The dotted attenuation curve, which includes the effective resistance in the inductance coils, follows the ideal attenuation curve closely, except in the neighborhood of 1,000 cycles, where resistance rounds off the abrupt corner which is present in the ideal A curve. The phase constant B is, at the start, proportional to the frequency, as for an ordinary uniform transmission line; its slope becomes steeper as the critical frequency 1,000 is approached where the curve reaches the ordinate π, at which value it remains constant for all higher frequencies. As shown by the dotted B curve, resistance rounds off the corner at the critical frequency, but

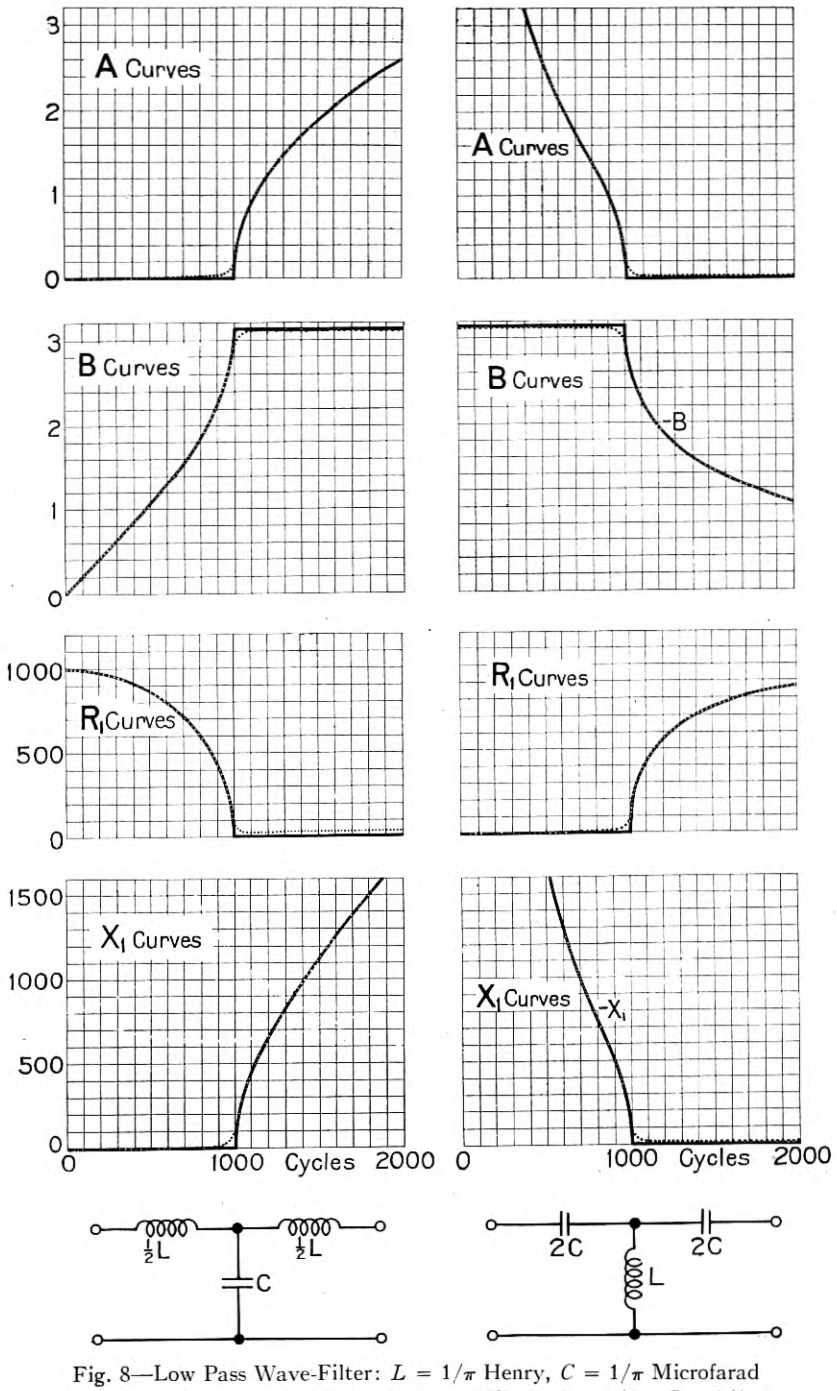


Fig. 8—Low Pass Wave-Filter: $L = 1/\pi$ Henry, $C = 1/\pi$ Microfarad
Fig. 9—Complementary High Pass Wave-Filter: $L = 1/4\pi$, $C = 1/4\pi$

otherwise leaves the curve approximately unchanged. The full line curves for R_1 and X_1 show that in the ideal case the iterative impedance is pure resistance and pure reactance in the pass and stop bands respectively, and that resistance smooths the abrupt transition at the critical frequency.

The high pass wave-filter shown by Fig. 9 passes the band which is stopped by the low pass wave-filter of Fig. 8, and vice versa. For this reason the two wave-filters are said to be complementary.

Another set of two complementary wave-filters is shown by Figs. 10 and 11, one of which passes only a single band of frequencies, not extending to either zero or infinity, while the other passes the remaining frequencies only. The single pass band of Fig. 10, embracing a total phase change 2π on the B curve, is actually a case of confluent pass bands, each of which embraces the normal angle π. The tendency of the two simple pass bands to separate, and leave a stop band between them, is shown by the hump in the dotted attenuation constant curve at 1,000 cycles. If, instead of the two simple bands having been brought together, one of them had been relegated to zero or infinity, the single remaining pass band would have exhibited the normal angular range π in the B curve, and there would have been no hump in the dotted A curve. The stop band of Fig. 11 also illustrates peculiarities which are not necessary features of a wave-filter with a single stop band in this position. This wave-filter is obtained from Fig. 7 by making all bands vanish except P_2, S_3, S_5 and P_3,—by extending P_2 to zero, P_3 to infinity, and making S_3 and S_5 coalesce, so that the attenuation becomes infinite in the stop band without passing from a stop $(-)$ to a stop $(+)$ band. The coalescing stop bands are responsible for the rapid changes in the B, R_1, and X_1 curves of Fig. 11 which would not have appeared if, in Fig. 7, the same pass band had been obtained by retaining P_1, S_2 and P_2 and making all other bands vanish.

An extreme case of complementary wave-filters is shown by Figs. 12 and 13, where no frequencies and all frequencies are passed respectively. The first result is obtained by combining inductances alone, which, as has been pointed out above, can give only an attenuated disturbance devoid of wave characteristics. The wave-filter shown for passing all frequencies has inductance coils in the line, and capacities diagonally bridged across the line. This wave-filter combines a constant iterative impedance with a progressive change in phase which is sometimes useful.[4] An outstanding char-

[4] A theoretical use of the phase shifting afforded by the lattice artificial line was made at page 253 of "Maximum Output Networks for Telephone Substation and Repeater Circuits," Trans. A. I. E. E., vol. 39, pp. 231–280, 1920.

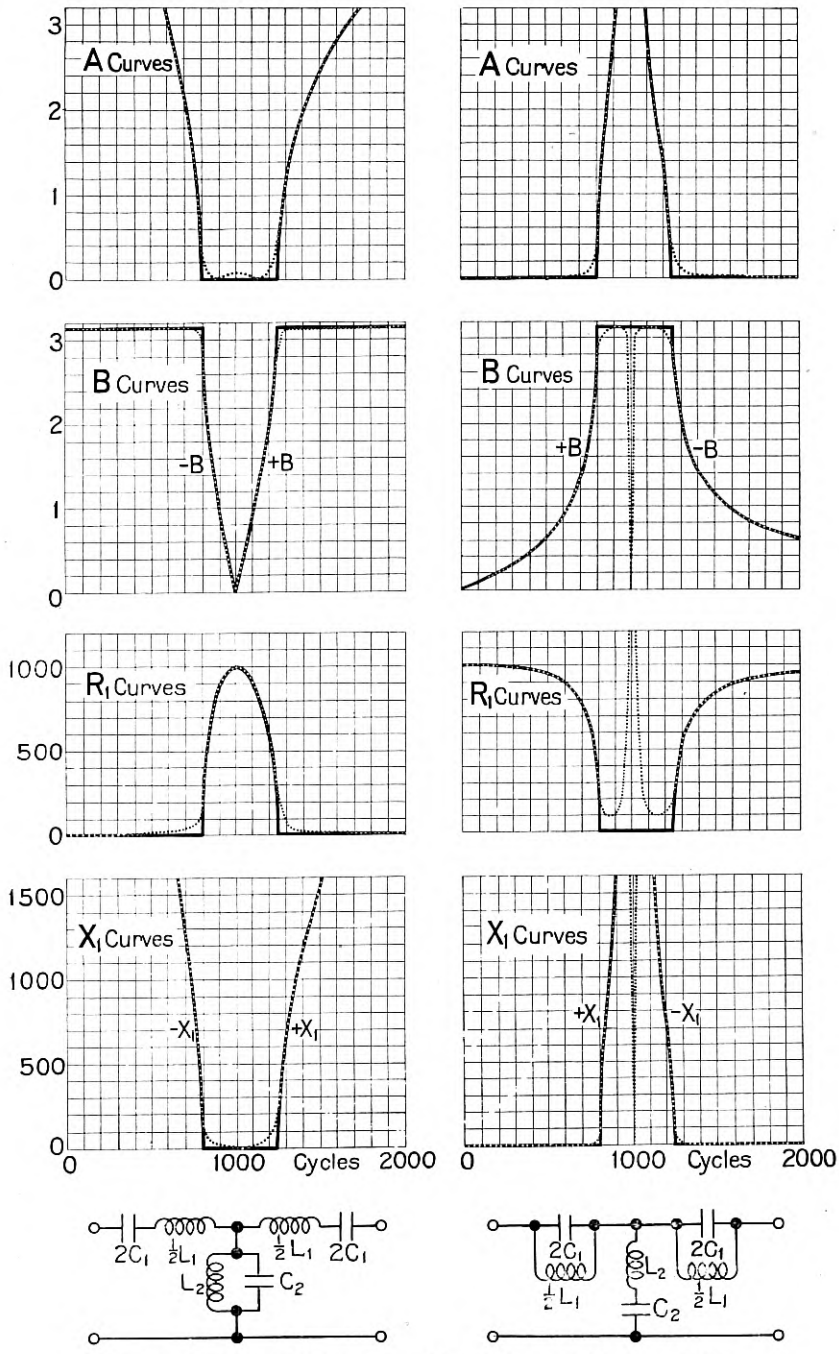


Fig. 10—Single Band Pass Wave-Filter: $L_1 = 20/9\pi$, $L_2 = 9/80\pi$, $C_1 = 9/80\pi$, $C_2 = 20/9\pi$

Fig. 11—Complementary High and Low Pass Wave-Filter: $L_1 = 9/20\pi$, $L_2 = 5/9\pi$, $C_1 = 5/9\pi$, $C_2 = 9/20\pi$

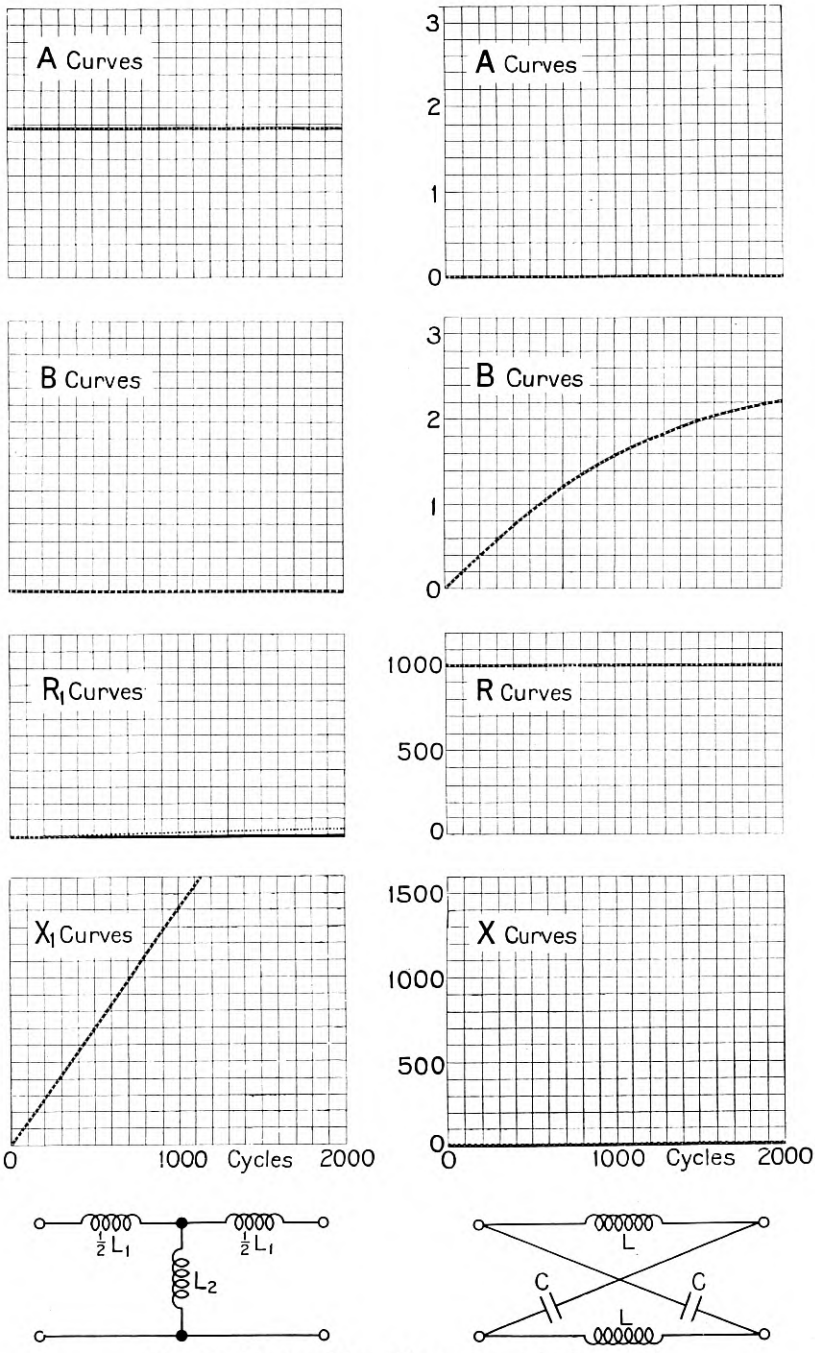


Fig. 12—No Pass Wave-Filter: $L_1 = 1/\pi$, $L_2 = 1/4\pi$

Fig. 13—Complementary All Pass Wave-Filter: $L = 1/2\pi$, $C = 1/2\pi$

acteristic of this type of artificial line is that it has, for all frequencies, the same iterative impedance as a uniform line with the same total series and shunt impedances. This artificial line will be considered in more detail in the next section of this paper.

Lattice Artificial Lines

Up to this point we have considered the properties of artificial line networks which were supposed to be given. In practice the problem is ordinarily reversed, and we ask the questions: May the locations of the bands be arbitrarily assigned? May additional conditions be imposed? How may the corresponding network be determined, and what is its attenuation in terms of the assigned critical frequencies? These questions might be answered by a study of Fig. 7, in

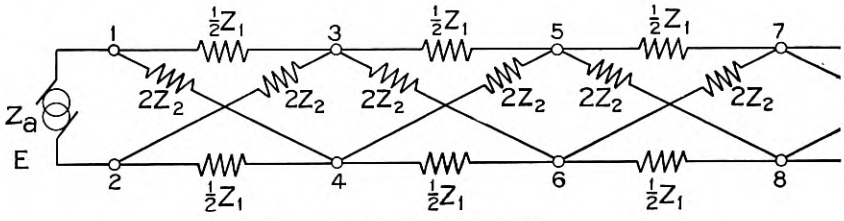


Fig. 14—Lattice Artificial Line

all its generality, but it seems simpler to base the discussion upon the artificial line shown in Fig. 14, which is to be a generalization of Fig. 13 to the extent of making the two impedances Z_1 and Z_2 any possible actual driving-point impedances. It is sometimes illuminating to regard this artificial line as a nest of bridges, one within another, as shown by Fig. 15.

On interchanging terminals 3 with 4 and 7 with 8 in Fig. 14 the network of lines remains unchanged; thus, Z_1 and $4Z_2$ may be interchanged in the formulas for the artificial line with no change in the result, except, possibly, one corresponding to a reversal of the current at alternate junction points. Another elementary feature of this artificial line is that it degenerates into a simple shunt or a simple series circuit at the resonant or anti-resonant frequencies, respectively, of either Z_1 or Z_2, and these are the critical frequencies, terminating the pass bands. At other frequencies, a positive ratio $Z_1/4Z_2$ must give a stop band, since the reactances are all of one sign. If a small negative value of this ratio gives free transmission, as we naturally expect, there will be identical transmission, except for a reversal of

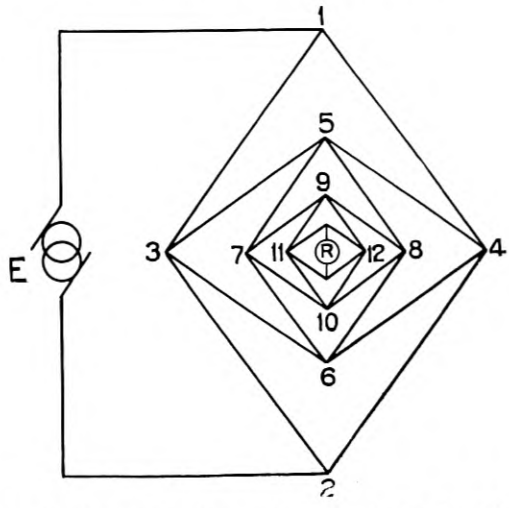


Fig. 15—Lattice Artificial Line Drawn to Show the Chain of Bridge Circuits

sign, when the ratio has the reciprocal value, which will be a large negative quantity, since we may always interchange Z_1 and $4Z_2$. The consequences of this and of other elementary properties of this artificial line are brought together in the following table:

TABLE II

For Lattice Artificial Line, Fig. 14

Band	Critical Frequency	Ratio $\frac{Z_1}{4Z_2}$	Uniform Line		Artificial Line		
			γ	k	Γ	$e^{-\Gamma}$	K
Stop (+)		$1>>0$	$2>>0$	imag.	+ real	$0<<1$	imag.
	$Z_1 = 0$ $Z_2 = \infty$	0 0	0 0	0 ∞	0 0	1 1	0 ∞
Pass		<0	imag.	+ real	imag.	$e^{i\theta}$	+ real
	$Z_1 = \infty$ $Z_2 = 0$	∞ ∞	∞ ∞	∞ 0	$i\pi$ $i\pi$	-1 -1	∞ 0
Stop (−)		<1	<2	imag.	$i\pi$ + real	$-1<<0$	imag.
	$Z_1 = 4Z_2$	1	2	$2Z_2$	∞	0	$2Z_2$

The cycle of bands: stop (+), pass, stop (−), adopted for the table, carries the attenuation factor $e^{-\Gamma}$ around the periphery of

a unit semi-circle; in the stop (+) band it traverses the radius from 0 to 1, in the pass band it travels along the unit circle through 180 degrees to the value -1, completing the cycle from -1 to 0 in the stop (−) band. In this cycle there are four points of special interest, corresponding to ratio values 1, 0, -1 and ∞, for which the wave is infinitely attenuated, unattenuated with an angular change of 0, of 90, and of 180 degrees, respectively. It is at the 90 degree angle that resonance of the individual section occurs; the iterative impedance is then equal to $2|Z_2|$.

Graph of the Ratio $Z_1/4Z_2$ for Fig. 14

If we plot Z_1 and $4Z_2$ the pass bands are shown by the points where the curves become zero or infinite, and the intersections of the two

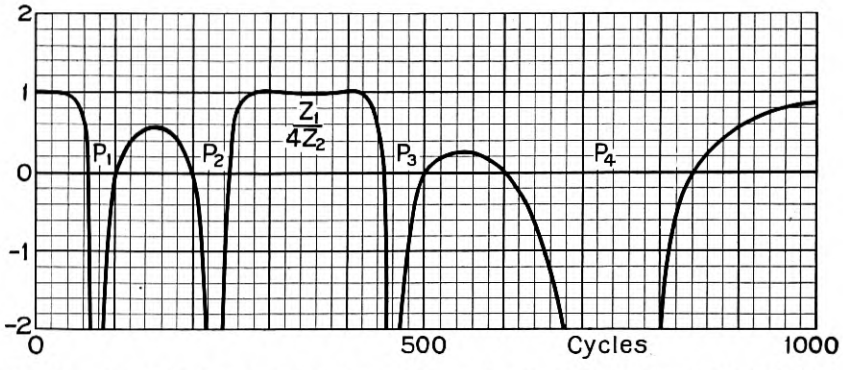


Fig. 16—Graph for Locating the Pass and Stop Bands of the Lattice Artificial Line, where $Z_1/4Z_2 = \left[(\lambda_1^2 - x^2)\ (\lambda_2^2 - x^2)^{-2}\ (\lambda_3^2 - x^2)\right]$ . . . , x = cycles/100, and the resonant roots x_1, x_3, . . . are 0.650, 1, 2, 2.452, 4.442, 5, 6, 8.476 and the double anti-resonant roots x_2, x_5, . . . are 0.766, 2.301, 4.585, 7.423

curves show the frequencies at which the attenuation becomes infinite. These intersections must be at an acute angle since each branch of the two curves has a positive slope throughout its entire length; for this reason it may be desirable to plot the ratio rather than the individual curves; this is especially desirable in cases where the two curves do not intersect, but are tangent. Fig. 16 is for a lattice network equivalent to two sections of the ladder type illustrated by Fig. 7, and so cannot include a stop (−) band. Accordingly, the ratio does not go above unity, although it reaches unity at the two frequencies 300 and 400, corresponding to the infinite attenuation where stop (−) and stop (+) bands meet in Fig. 7. It is also

unity at the extreme frequencies zero and infinity. The four pass bands have, of course, the same locations as in Fig. 7.

Multiplying the ratio by a constant greater than unity introduces stop (−) bands along with the stop (+) bands; multiplying it by a constant less than unity removes all infinite attenuations; these changes within the stop bands are made without altering the locations of the four pass bands.

Wave-Filter Having Assigned Pass Bands

In connection with practical applications we especially desire to know what latitude is permitted in the preassignment of properties for a wave-filter. If we consider first the ideal lattice wave-filter, its limitations are those inherent in the form which its two independent resistanceless one-point impedances[5] Z_1 and Z_2 may assume. The mathematical form of this impedance is shown by formula (7) of the appendix, which may be expressed in words as follows:

Within a constant factor the most general one-point reactance obtainable by means of a finite, pure reactance network is an odd rational function of the frequency which is completely determined by assigning the resonant and anti-resonant frequencies, subject to the condition that they alternate and include both zero and infinity.

The corresponding general expressions for the quotient and product of the impedances Z_1 and Z_2 are shown by formulas (8) and (9). Definite, realizable values for all of the $2n+2$ parameters and $2n+1$ optional signs occurring in these formulas may be determined in the following manner:

(a) Assign the location of all n pass bands, which must be treated as distinct bands even though two or more are confluent; this fixes the values of the $2n$ roots $p_1 \ldots p_{2n}$ which correspond to the successive frequencies at the two ends of the bands.

(b) Assign to the lower or upper end of each pass band propagation without phase change from section to section; this fixes the corresponding optional sign in formula (8) as + or −, respectively.

(c) Assign a value to the propagation constant at any one non-critical frequency (that is, assign the attenuation constant in a

[5] A one-point impedance of a network is the ratio of an impressed electromotive force at a point to the resulting current at the same point—in contradistinction to two-point impedances, where the ratio applies to an electromotive force and the resulting current at two different points.

stop band or the phase constant in a pass band); this fixes the value of the constant G and thus completely determines formula (8) on which the propagation constant depends.

(d) Assign to the lower or upper end of each stop band the iterative impedance zero; this fixes the corresponding optional sign in formula (9) as $+$ or $-$, respectively.

(e) Assign the iterative impedance at any one non-critical frequency (subject to the condition that it must be a positive resistance in a pass band and a reactance in a stop band); this fixes the constant H and thereby the entire expression (9) upon which the iterative impedance depends.

The quotient and product of the impedances Z_1 and Z_2 are now fully determined; the values of Z_1 and Z_2 are easily deduced and also the propagation constant and iterative impedance by formulas (11) and (12); Z_1 and Z_2 are physically realizable except for the necessary resistance in all networks.

These important results may be summarized as follows:

A lattice wave-filter having any assigned pass bands is physically realizable; the location of the pass bands fully determines the propagation constant and iterative impedance at all frequencies when their values are assigned at one non-critical frequency, and zero phase constant and zero iterative impedance are assigned to the lower or upper end of each pass band and stop band, respectively.

Lattice Artificial Line Equivalent to the Generalized Artificial Line of Fig. 1

Since any number of arbitrarily preassigned pass bands may be realized by means of the lattice network, it is natural to inquire whether this network does not present a generality which is essentially as comprehensive as that obtainable by means of any network N in Fig. 1, provided the generalized line is so terminated as to equalize its iterative impedances in the two directions. This proves to be the case.

If network N has identical iterative impedances in both directions, the lattice network equivalent to two sections of N is shown by Fig. 17; each lattice impedance is secured by using an N network; the N's placed in the two series branches of the lattice have their far terminals short-circuited so that they each give the impedance denoted by Z_0; the N's in the two diagonal branches have their far ends open and they each give the impedance denoted by Z_∞.

The lattice network of Fig. 18 has in each branch a one-point impedance obtained by means of a duplicate of the given network N and an ideal transformer. The two lattice branch impedances are $Z_q+Z_r=2Z_{qr}$ where the three impedances Z_q, Z_r, Z_{qr} are the effective self and mutual impedances of the network N regarded as a transformer. This lattice network has identically the same propaga-

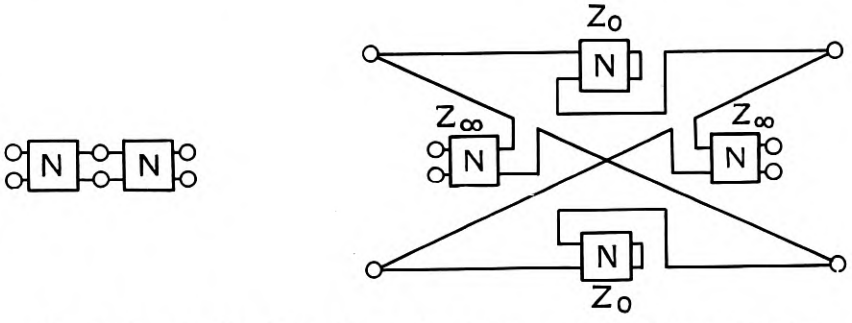


Fig. 17—Lattice Unit Equivalent to Two Sections of Fig. 1 Assumed to be Symmetrical

tion constant as the single network N shown on the left. Since the lattice cannot have different iterative impedances in the two directions, it actually compromises by assuming the sum of the two iterative impedances presented by N. A physical theory of the equival-

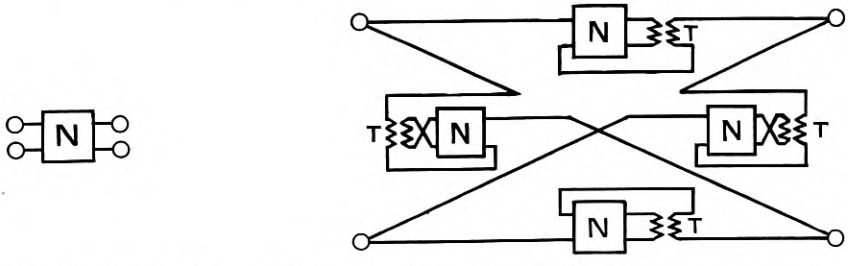


Fig. 18—Lattice Network Having the Same Propagation Constant as N and an Iterative Impedance Equal to the Sum of the Two Iterative Impedances of N

ences shown in Figs. 17 and 18 has not been worked up; the analytical proofs were made by applying the formulas given in the appendix under lattice networks.

Without going to more complex networks it is, of course, not possible to get a symmetrical iterative impedance, but that is not necessary for our present purposes where we are concerned primarily with the

propagation constant. It has now been shown with complete generality that:

The lattice artificial line, with physically realizable branch impedances, is identically equivalent in propagation constant and mean iterative impedance to the chain of identical physically realizable networks connected together in sequence through two pairs of terminals.

To complete this simplification of the generalized artificial line it is necessary to know the simplest possible form of the one-point impedances employed in the branches of the lattice network. The discussion of the most general one-point impedance obtainable by means of any network of resistances, self and mutual inductances, leakages and capacities will find its natural place, together with allied theorems, in a paper on the subject of impedances. For the present purpose it is sufficient to state:

The most general branch impedance of the lattice network may be constructed by combining, in parallel, resonant circuits having impedances of the form $R+iLp+(G+iCp)^{-1}$*; or they may equally well be constructed by combining, in series, anti-resonant circuits having impedances of the form* $\left[G+iCp+(R+iLp)^{-1}\right]^{-1}$.

Summary of Physical Theory

The wave-filter under discussion approximates to a resistanceless artificial line, and such an ideal artificial line is capable of two, and only two, fundamentally distinct states of motion. In one state the disturbance is attenuated along the line, and there is no flow of energy other than a back and forth surging of energy, the intensity of which rapidly dies out along the line. In the other state there is a free flow of energy, without loss, from section to section along the line, with no surge of energy between symmetrical sections. Each state holds for one or more continuous bands of frequencies; these bands have been distinguished as stop bands and pass bands.

A high degree of discrimination, between different frequencies, may be obtained, even if each section, taken alone, gives only a moderate difference in attenuation, by the use of a sufficient number of sections in the wave-filter, since the attenuation factors vary in geometrical progression with the number of sections.

Any number of arbitrarily located pass bands may be realized by means of the lattice artificial line; furthermore, the propagation constant at one frequency, and the iterative impedance at one frequency may both be assigned, while the location of zero phase con-

stant and zero iterative impedance at the lower or upper end of each pass band and stop band, respectively, is also optional. This completely determines the lattice artificial line. No additional condition, other than iterative impedance asymmetry, can be realized by replacing the lattice network by any four terminal network.

APPENDIX

Formulas for the Artificial Line

Formulas for the propagation constant and iterative impedance of the generalized artificial line, expressed in a number of equivalent forms, have already been given in my paper on Cisoidal Oscillations,[6] but it seems worth while to deduce the formulas anew here from the free oscillations of the detached unit circuit of Fig. 6, so as to complete the physical theory by deducing the comprehensive mathematical formulas by the same method of procedure.

Ladder Network Formulas

Notation:

Z_1, Z_2 = series impedance and shunt impedance of the section of Fig. 4, which is equivalent to the general network N of Fig. 1.

$\Gamma = A + iB$ = propagation constant per section.

K_1, K_2 = iterative impedances at mid-series and mid-shunt.

$\gamma = \alpha + i\beta = \sqrt{Z_1/Z_2}$ = propagation constant for uniform distribution of Z_1 and $1/Z_2$, per unit length.

$k = \sqrt{Z_1 Z_2}$ = iterative impedance of this same uniform line.

In Fig. 6, the current is indicated as I and the potentials at the ends of the section as $V, Ve^{-\Gamma}$. In order that the free oscillation may be possible the total impedance of the circuit $(Z_1 + Z' + Z'')$ must vanish; this determines the iterative impedance K_2. In addition to this condition it is sufficient to make use of two other simple relations: the proportionality of the potential drops in the direction of the current across Z' and Z'' to Z' and Z'', since they carry the same current (this determines the propagation constant Γ); and the equality of

[6] "Cisodial Oscillations," Trans. A. I. E. E., vol. 30, pp. 873-909, 1911. In the lowest row of squares of Table I, the iterative impedances and propagation constant of any network are given in five different ways, involving one-point and two-point impedances, equivalent star impedances, equivalent delta impedances, equivalent transformer impedances, or the determinant of the network. The only typographical errors in Table I appear to be the four which occur in the first, third and fifth squares of this row: in the values for K_q replace $(S_q - S_{qr})$ by $(S_q - S_r)$ and place a parenthesis before $U_q - U_r$); in the first value of K_r replace S_{qr} by S_{qr}^2; in the last value for Γ_{qr} add a minus sign so that it reads $\cosh^{-1}$.

K_1, the mid-series iterative impedance of the artificial line, to the total impedance on the right of the mid-point of the series impedance Z_1. These three relations, which can be written down at once, are:

$$Z_1 + Z' + Z'' = Z_1 - \frac{4Z_2K_2^2}{4Z_2^2 - K_2^2} = 0,$$

$$\frac{Ve^{-\Gamma}}{V} = -\frac{Z''}{Z'} = \frac{2Z_2 - K_2}{2Z_2 + K_2},$$

$$K_1 = \frac{1}{2}Z_1 + Z'' = \frac{1}{2}Z_1 + \frac{2Z_2K_2}{2Z_2 + K_2},$$

from which the formulas for Γ, K_1, and K_2, in terms of Z_1, Z_2, are found to be:

$$\Gamma = 2\sinh^{-1}\frac{1}{2}\sqrt{\frac{Z_1}{Z_2}} = 2\sinh^{-1}\frac{1}{2}\gamma, \tag{1}$$

$$\left.\begin{matrix}K_1\\K_2\end{matrix}\right\} = \sqrt{Z_1Z_2}\left(1 + \frac{Z_1}{4Z_2}\right)^{\pm\frac{1}{2}} = k\left(1 + \frac{1}{4}\gamma^2\right)^{\pm\frac{1}{2}} \text{ at mid } \left\{\begin{matrix}\text{series}\\\text{shunt,}\end{matrix}\right. \tag{2}$$

and the formulas for Z_1 and Z_2 in terms of Γ and K_1 or K_2 are likewise found to be:

$$Z_1 = 2K_1\tanh\frac{1}{2}\Gamma = K_2\sinh\Gamma, \tag{3}$$

$$Z_2 = K_1/\sinh\Gamma = \frac{1}{2}K_2\coth\frac{1}{2}\Gamma. \tag{4}$$

Formulas (3) and (4) are in the nature of design formulas in that they determine the impedance Z_1 and Z_2, at assigned frequencies, which will ensure the assigned values of Γ and K at these frequencies. In general, however, it would not be evident how best to secure these required values of Z_1 and Z_2; complicated or even impossible networks might be called for, even to approximate values of Z_1 and Z_2 assigned in an arbitrary manner. Fortunately, practical requirements are ordinarily satisfied by meeting maximum and minimum values for the attenuation constant throughout assigned frequency bands. Formulas (8) and (9) may be employed for this purpose as explained below.

It is convenient to have formulas (1) and (2) expressed in a variety of ways, since no one form is well suited for calculation throughout the entire range of the variables. Accordingly, the following analytically equivalent expressions are here collected together for reference:

$$\Gamma = i\,2\sin^{-1}\frac{\gamma}{2i} = i\cos^{-1}\left(1+\frac{\gamma^2}{2}\right), \tag{5}$$

$$= 2\sinh^{-1}\frac{\gamma}{2} = 2\tanh^{-1}\frac{\frac{\gamma}{2}}{\sqrt{1+\frac{\gamma^2}{4}}} = \cosh^{-1}\left(1+\frac{\gamma^2}{2}\right)$$

$$= 2\log\left[\frac{\gamma}{2}+\sqrt{1+\frac{\gamma^2}{4}}\right], \tag{5a}$$

$$= i\pi + 2\cosh^{-1}\frac{\gamma}{2i} = i\pi + \cosh^{-1}\left(-1-\frac{\gamma^2}{2}\right)$$

$$= i\pi + 2\log\left[\frac{\gamma}{2i}+\sqrt{-1-\frac{\gamma^2}{4}}\right], \tag{5b}$$

$$= i\frac{\pi}{2} - \sinh^{-1}\left(1+\frac{\gamma^2}{2}\right)i, \tag{5c}$$

$$= \gamma - \frac{1}{24}\gamma^3 + \frac{3}{640}\gamma^5 - \frac{5}{7168}\gamma^7 + \ldots, \text{ if } |\gamma| < 2, \tag{5d}$$

$$= 2\cosh^{-1} g + i\,2\sin^{-1}\frac{\beta}{2g}, \tag{5e}$$

where $\gamma = \alpha + i\beta$, $2g = \sqrt{\left(\frac{\alpha}{2}\right)^2+\left(1+\frac{\beta}{2}\right)^2}+\sqrt{\left(\frac{\alpha}{2}\right)^2+\left(1-\frac{\beta}{2}\right)^2}$,

$$= \cosh^{-1} h + i\cos^{-1}\frac{x}{h}, \tag{5f}$$

where $1+\frac{1}{2}\gamma^2 = x+iy$, $2h = \sqrt{(x+1)^2+y^2}+\sqrt{(x-1)^2+y^2}$,

$$\left.\begin{matrix}K_1\\K_2\end{matrix}\right\} = k\left(1+\frac{1}{4}\gamma^2\right)^{\pm\frac{1}{2}} = k\left(\cosh\frac{\Gamma}{2}\right)^{\pm 1} = k\left(\frac{\sinh\Gamma}{\gamma}\right)^{\pm 1}$$

$$= k\left(\frac{1}{2}\gamma\coth\frac{1}{2}\Gamma\right)^{\pm 1} \text{ at mid } \left\{\begin{matrix}\text{series}\\\text{shunt.}\end{matrix}\right. \tag{6}$$

The formulas leave indeterminate the signs of γ, k, Γ, and K, and also a term $\pm i2\pi n$ in γ and Γ. The signs are to be so chosen that the real parts are positive, or become positive when positive resistance is added to the system. The indeterminate $\pm i2\pi n$ can be made determinate only after knowing something of the internal structure of the unit network of which the artificial line is composed; the conditions to be met are—absence of phase differences when all branches of the unit network N of Fig. 1 are assumed to be pure

resistances and continuity of phase as reactances are gradually introduced to give the actual network.

Formula (5) is adapted for use in the pass bands, since the expressions are real when γ^2 is real, negative and not less than -4; similarly, formulas (5a) and (5b) are adapted for use in the stop ($\pm$) bands, that is, when γ^2 is positive and less than -4 respectively.

From the theory of impedances we know that any resistanceless one-point impedance is expressible in the form

$$Z = iD \frac{p \quad (p_2^2-p^2) \ldots (p_{2n-2}^2-p^2)}{(p_1^2-p^2)\ (p_3^2-p^2) \ldots (p_{2n-1}^2-p^2)} \tag{7}$$

where the factor D and the roots $p_1, p_2, \ldots p_{2n}$ are arbitrary positive, reals subject only to the condition that each root is at least as large as the preceding one. This enables us to write down the forms which the quotient and product of two resistanceless one-point impedances may assume, which are as follows:

$$\frac{Z'}{Z''} = G\left(\frac{p_1^2-p^2}{p_2^2-p^2}\right)^{\pm 1}\left(\frac{p_3^2-p^2}{p_4^2-p^2}\right)^{\pm 1} \ldots \left(\frac{p_{2n-1}^2-p^2}{p_{2n}^2-p^2}\right)^{\pm 1} \tag{8}$$

$$Z'Z'' = -H\left(\frac{p^2}{p_1^2-p^2}\right)^{\pm 1}\left(\frac{p_2^2-p^2}{p_3^2-p^2}\right)^{\pm 1} \ldots \left(\frac{p_{2n-2}^2-p^2}{p_{2n-1}^2-p^2}\right)^{\pm 1}(p_{2n}^2-p^2)^{\pm 1} \tag{9}$$

where G, H and the roots $p_1, p_2, \ldots p_{2n}$ are arbitrary positive reals, subject only to the condition that each root is at least as large as the preceding one, and the $2n+1$ and optional $\pm$ signs are mutually independent. Conversely, if the relations (8) and (9) are prescribed, then the required individual impedances Z' and Z'' are each of the form (7) and thus physically realizable.

If in formulas 1, 2, 5 and 6 we substitute for $Z_1/Z_2 = \gamma^2$ and $Z_1 Z_2 = k^2$ the right-hand side of formulas (8) and (9), respectively, we obtain formulas for the propagation constant and iterative impedance of an artificial resistanceless line in terms of frequencies at which the propagation constant becomes zero or infinite. Ordinarily, however, we are more interested in having expressions in terms of the frequencies which terminate the pass bands. To secure these the substitutions should be $4[8]/(4-[8])$ and $[9]\,(1-[8]/4)^{\pm 1}$, where [8] and [9] stand for the entire right-hand sides of formulas (8) and (9). This substitution amounts to obtaining the lattice network giving the required pass bands, and then transforming to the

ladder network having the same propagation constant and the same iterative impedance at mid-series or mid-shunt.

Lattice Network Formulas Fig. 14

The impedances of a single section between terminals 1 and 2, with the far end of the section 3 and 4 either short-circuited or open, are readily seen to be

$$Z_0 = \frac{2Z_1Z_2}{\frac{1}{2}Z_1 + 2Z_2}, \quad Z_\infty = \frac{1}{2}\left(\frac{1}{2}Z_1 + 2Z_2\right). \tag{10}$$

Since $\sqrt{Z_0Z_\infty}$ and $\sqrt{Z_0/Z_\infty}$ are the iterative impedance and the hyperbolic tangent of the propagation constant for any symmetrical artificial line, we have the following analytically equivalent formulas for the lattice network where $\gamma = \sqrt{Z_1/Z_2}$, and $k = \sqrt{Z_1Z_2}$ as for the ladder type.

Lattice Formulas

$$\begin{cases} \Gamma = 2\tanh^{-1}\frac{1}{2}\sqrt{\frac{Z_1}{Z_2}} = 2\tanh^{-1}\frac{1}{2}\gamma, & (11) \\ K = \sqrt{Z_1Z_2} = k. & (12) \end{cases}$$

$$\begin{cases} Z_1 = 2K\tanh\frac{1}{2}\Gamma, & (13) \\ Z_2 = \frac{1}{2}K\coth\frac{1}{2}\Gamma. & (14) \end{cases}$$

$$\Gamma = i2\tan^{-1}\frac{\gamma}{2i} = i\cos^{-1}\frac{1+\frac{1}{4}\gamma^2}{1-\frac{1}{4}\gamma^2}, \tag{15}$$

$$= 2\tanh^{-1}\frac{\gamma}{2} = 2\sinh^{-1}\frac{\frac{1}{2}\gamma}{\sqrt{1-\frac{1}{4}\gamma^2}} = \cosh^{-1}\frac{1+\frac{1}{4}\gamma^2}{1-\frac{1}{4}\gamma^2}$$

$$= \log\frac{1+\frac{1}{2}\gamma}{1-\frac{1}{2}\gamma}, \tag{15a}$$

$$= i\pi + 2\coth^{-1}\frac{\gamma}{2} = i\pi + \cosh^{-1}\frac{1+\frac{1}{4}\gamma^2}{-1+\frac{1}{4}\gamma^2}$$

$$= i\pi + \log\frac{1+\frac{1}{2}\gamma}{-1+\frac{1}{2}\gamma}, \qquad (15b)$$

$$= i\frac{\pi}{2} - \sinh^{-1}\frac{1+\frac{1}{4}\gamma^2}{1-\frac{1}{4}\gamma^2}\,i, \qquad (15c)$$

$$= \gamma + \frac{1}{12}\gamma^3 + \frac{1}{80}\gamma^5 + \frac{1}{448}\gamma^7 + \ldots, |\gamma| < 2, \qquad (15d)$$

$$= \frac{1}{2}\log\frac{\left(1+\frac{\alpha}{2}\right)^2+\left(\frac{\beta}{2}\right)^2}{\left(1-\frac{\alpha}{2}\right)^2+\left(\frac{\beta}{2}\right)^2} + i\tan^{-1}\frac{\beta}{1-\left(\frac{\alpha}{2}\right)^2-\left(\frac{\beta}{2}\right)^2}, \qquad (15e)$$

where $\gamma = \alpha + i\beta$.

In these formulas $Z_1/Z_2 = \gamma^2$ and $Z_1 Z_2 = k^2$ might be expressed in terms of the resonant and anti-resonant complex frequencies of Z_1 and Z_2, the frequencies being made complex quantities so as to include the damping. Where there is no damping, that is, where all network impedances are devoid of resistance, the simplified forms of these expressions are given by formulas (8) and (9). The use of these formulas for designing wave filters having assigned pass bands is explained at page 23.

The Binaural Location of Complex Sounds

By R. V. L. HARTLEY and THORNTON C. FRY

NOTE: Much has been written on the subject of the binaural location of pure tones but the case of complex sounds has received little attention in recent literature. The purpose of the present paper is to bring the discussion of complex sounds abreast of that relating to pure tones. Those who wish to acquaint themselves with the work on pure tones will be interested in reading the theoretical work of the authors and the experimental studies carried out by G. W. Stewart and students working under his direction. This work has been reported in various papers, most of which have appeared during recent years in the *Physical Review* and the *Physikalische Zeitschrift.*

A resume of the present paper is given by the authors in their concluding paragraph.—*Editor.*

THE need of determining the location of enemy submarines and aeroplanes during the war brought into use practical methods for locating a sound source which depend upon differences between the sound waves reaching the two ears. This stimulated a general study of the phenomena involved in binaural sound location. The foundation for this study had already been laid in the work of Lord Rayleigh and others, who, following more or less in his footsteps, had accumulated a considerable amount of information of both theoretical and experimental sorts. Of this information almost all that was of a theoretical nature and a considerable portion of the experimental kind dealt only with the location of *pure* tones, the more complicated and in some respects more important problem of *complex* sounds being almost entirely neglected. Such advances as were made in the theoretical aspects of the problem during the war were subject to the same restriction so that even to-day no comprehensive theory has been advanced which adequately covers the problem of the location of such sounds as occur in every-day life, and in the practical applications of binaural methods. However, the results obtained with pure tones can be made to throw considerable light upon the problem, and it is primarily from this standpoint that the following discussion is written.

It may be well at the outset to review some of the outstanding differences between the observed phenomena in the two cases. The accuracy of location is much less for pure tones, as is also the sense of definiteness of the sound image. The location of pure tones is almost wholly binaural as is evidenced by the inability of persons deaf in one ear to locate such a tone. With complex sounds not only is the location by binaural effects more accurate and definite, but also the observer is not dependent on these alone. Persons who are deaf

in one ear can locate familiar complex sounds almost as well as those with normal hearing.

Practically all theories of sound location start from the assumption that the listener subconsciously observes certain sound characteristics which depend upon the position of the source and forms a judgment of where the source must be by comparing these characteristics with information which he has stored up as a result of his past experience with cases in which the position of the source was known. In order to fix the position of the source he must assign to it three coördinates such as its distance and some two angles which define its direction. To do this he must be able to observe at least three independent properties of the sound which are functions of the position of the source. If fewer than three are available some difficulty in location is certain to arise. If more than three are available there is the possibility of a number of simultaneous independent determinations of the three coördinates.

If the sounds of every-day life were never distorted in transmission all of these determinations would yield the same set of coördinates and the only advantage which the listener would gain from the additional information available would lie in the fact that some one set might be peculiarly sensitive to slight differences in the position of the source, and therefore might lead to increased certainty on the part of the observer. Owing to reflection from the walls of buildings and the like, the sounds of every-day life seldom arrive undistorted, so that the observer must always be somewhat uncertain as to whether or not the coördinates of the sound source are actually those which he deduces from the properties of the sound wave as it reaches his ears. If enough properties are available to permit him to make two independent determinations he may use one of them to check the other, and if they agree he is justified in a feeling of increased certainty as to the accuracy of his judgment. The more independent determinations he can make the more checks he will be able to apply and consequently the more confident he will be.[1]

It should not be inferred, however, that it is only the sounds of the street which reach the observer in a distorted form. In a great many laboratory experiments the characteristics of the sounds have

[1] It is interesting to note in this connection that it is not surprising that an observer locates a complex tone with much greater certainty than a pure tone when we consider how rapidly the number of independent sets of data increases with increase in complexity of sound. We have already said that three independent properties are needed for the determination of the three coördinates of the source. Hence if only three are available, only one determination can be made and no checks are possible. On the other hand, if four are available, four groups of three each can be formed and therefore four separate determinations can be made. Similarly, 10 determinations can be made from 5 properties, 20 from 6, and 120 from 10.

been inconsistent, and in some cases they have not even corresponded to any actual source whatever. Under these circumstances, if an image is formed at all, some purely psychological factors must enter in. For pure tones it has been found possible to explain much of the experimental data obtained under circumstances such as this by assuming that the observer subconsciously judges one or more of the characteristics to be in error and applies such corrections as will make all of the data correspond to an actual source. As a criterion for determining which characteristics will be altered, it is assumed that, in general, those are chosen which require the smallest changes.

Let us now consider what characteristics are available for locating sounds of different kinds. A pure tone from a source at rest with respect to the observer has at any point only two physical characteristics which are subject to change with the position of the source. They are its amplitude and phase. Corresponding to each position of the source there is a particular amplitude and phase at each of the two ears so that a total of four properties—the loudness of the sound, the average phase, the difference in amplitude (which may conveniently be expressed as a ratio) and the difference in phase at the two ears—are available for determining the position of the source. It is inconceivable that the average phase can have anything to do with the location of the sound since it may be changed at will without altering the position of the source. The same remark applies to the loudness of the sound except in those instances where the observer is familiar with the source to such an extent as to know how loud it may be expected to be. Hence, if we restrict ourselves to the cases in which prejudicial information of this sort does not exist, we find that the observer has only two quantities from which he may deduce the position of the source. We should therefore expect that these two quantities would make it possible to locate the tone with respect to two coördinates only. This is found to be in general agreement with experiment, for most observers locate all sources of pure tones in the same horizontal plane with their heads and determine only the distance and angular departure from the median plane. If the source is more than a few yards away the intensity ratio and phase difference change very slowly with distance so that in this case even the sense of distance is not keen and a feeling of certainty exists with respect to the direction only.

In many experiments the tones at the two ears have been varied arbitrarily so as to give combinations having equal phases and unequal intensities or vice versa—combinations which cannot arise from actual physical sources in the absence of distortion. Under

these conditions the observer generally corrects one to a value consistent with the other except in extreme cases where the correction required for this purpose would be inordinately large. When this occurs he may either assume both to be correct and form two images—one based on the phase difference together with a mentally supplied intensity ratio consistent with it, and the other similarly derived from the observed intensity ratio—or he may fail to have a sense of location at all.

Before considering the available characteristics of complex sounds in general let us confine our attention for a time to those which are made up of a limited number of sustained pure tones such as an organ note with its series of overtones, or a group of tuning forks. Here the number of characteristics increases rapidly with the number of component tones. For each component tone there are two quantities: intensity ratio and phase difference. In addition, at either ear alone the relative intensities of any two of the tones changes with the position of the source, owing to the diffraction of the sound waves around the head being different for different frequencies. There are therefore as many of these observable intensity ratios as there are pairs of components. Similarly, for any two tones whose frequencies are commensurable, the relative phases of the two at the same ear depend upon the position of the source.

Not all of these characteristics are capable of contributing to binaural as distinct from monaural location. In fact, only the phase differences and intensity ratios of the separate components are binaural. A man who is deaf in one ear has available all of the relations between the intensities and phases of the various components at his normal ear. That these relations do actually contribute to sound location is supported by experimental evidence. Myers [2] found that, after familiarizing himself with a complex sound, a blindfolded observer could locate its position with considerable accuracy, even when it was moved about in the median plane, but that his accuracy could be destroyed by varying the relative intensities of the components.[3] It is not surprising then, that for complex sounds the accuracy is about the same whether the location is binaural or monaural.[4] The observed failure of monaural location in the case of a

[2] C. S. Myers, *Proc. Royal Soc.*, 1914, B 88,267.

[3] It should be noticed that this effect must have been purely psychological since it could be produced without moving the source at all. It therefore lends plausibility to the assumption upon which our theory is based: that when discordant or unusual stimuli are experienced, a mental readjustment of the stimuli is made in order to render them more nearly consistent with every-day experience.

[4] As shown by the experiments of Angell and Fite upon persons deaf in one ear. *Psychol. Rev.*, vol. 8, pp. 225–246, 1911.

pure tone follows directly from the absence of other frequencies with which the pure tone may be compared.

As we are here concerned with binaural phenomena we shall confine our attention to the relative phases and intensities at the two ears. The question at once arises: does the observer actually hear the different tones separately, and if so, does he assign a location to each separately?

To what extent the listener locates each component separately depends upon the ease with which the tones can be distinguished. The experiments which bear most directly upon this point are those in which the component tones at the two ears are arbitrarily adjusted to give values of phase difference corresponding to different locations. This is done under conditions where the location of each component separately is largely determined by the phase difference. More [5] experimented with two tones, transmitting them to the ears through tubes of adjustable lengths. This permitted him to change the phase difference at the two ears while keeping the intensities substantially equal. He observed the apparent location for various settings when each tone was applied by itself and when both were applied together, using forks of 256 and 320 cycles. With the paths equal the tones combined into a chord located in the median plane and the separate components could not be heard. With a setting for which the two components separately appeared on opposite sides of the head, one component was heard distinctly by the right ear only on the right side, and the other by the left ear only on the left side. At the same time the chord was heard rather indistinctly near the median plane but tending slightly toward the side of the lower tone.

Apparently the observer does not consciously separate the chord into its components unless he is forced to do so by some inordinate discrepancy between the positions of the images formed from them. There is no evidence in the case of equal paths to show that he did or did not subsconsciously locate the separate components and find them to be in agreement. In view of the second experiment it seems probable that he did. In this latter experiment he obviously found that the two components corresponded to different locations and assigned different sources to each. At the same time his experience told him that tones which would combine to form a musical sound generally have a common source. Hence he may have concluded subconsciously that the sound waves had probably been distorted in coming from a common source and so he corrected his observations

[5] Louis T. More: *Phil. Mag.* XVIII, 1909, p. 308.

on both tones to make them consistent and arrived at an image of the chord between the other two.

Similar results were obtained with forks of 256 and 384 cycles per second, except that in general the lower tone was completely blotted out. The higher tone was usually quite distinct and definitely located. The image of the chord was nearer to the image formed when the higher component was sounded by itself than to the image formed from the lower one alone. With settings for which the directions of the tones separately were the same, whether right, left, or middle, the upper tone disappeared leaving only the chord. In experiments with forks of 256 and 512 cycles it was difficult to distinguish the separate notes. With settings for which the two separately were on opposite sides the combination was on the side of the lower fork. This can be interpreted as meaning that the octave relationship is inherently difficult to resolve, or else that tones an octave apart so generally come from a common source that the observer was unwilling to make any other assumption.

Although the explanation of these results is not yet thoroughly understood, they show very definitely that in locating complex sounds made up of pure tones the observer does within limits locate the components separately. If they agree, a single image is formed; if they do not, he may either locate the tones separately or form a single compromise image or do both.

It is in this way that the theory developed for pure tones is applied to complex sounds made up of pure tones. The next step is to extend it so as to include complex sounds in general. To do this we must picture the observer as resolving each sound into sinusoidal components locating the components separately and forming one or more images based on a combination of the apparent sources as indicated by the separate components. While it is fairly easy to effect such a resolution mathematically it is somewhat less easy to interpret the result in a manner satisfactory to our intuitive conceptions of the phenomena involved; also, granted the theoretical possibility of the resolution, there remains the question of what physical or psychological limitations there may be to its application.

In view of the fact that a really pure component tone has no beginning or end, and no fluctuations in its amplitude, it is not at once apparent how a single discrete sound such as the bark of a dog can be resolved into components of that nature. However, if enough components are available it has been established beyond question that by properly choosing their frequencies, amplitudes, and phases, a

combination may be arrived at in which the algebraic sum of all the components is zero for all instants before and after the period occupied by the sound and equal to the instantaneous value of the sound wave for instants within that period. This combination is known to mathematicians as the Fourier Integral corresponding to the wave, and the formula for the phase and amplitude of each component sinusoid is known. It is an extension of the well known Fourier series expansion used for resolving sustained periodic disturbances.

The physical interpretation of this integral may be facilitated by reviewing the steps in its evolution from the Fourier series. It is well known that if the sound in question were repeated at regular intervals the resulting periodic wave could be resolved by Fourier analysis into a series of sinusoidal components, the frequencies of all of which are integral multiples of the frequency of repetition of the sound. Successive components therefore differ in frequency by an amount equal to this frequency of repetition. Now it is not essential that the repetitions of the sound follow each other immediately. Instead, they may be separated by intervals of silence. The effect of such silent intervals is to reduce the frequency of repetition and therefore also the fundamental frequency. As a result the component frequencies are brought closer together and the number within any particular frequency range is increased.

Suppose now that the interval between repetitions is indefinitely increased. As this is done the effect of any one occurrence of the sound becomes more and more independent of the others, and in the limit when the sounds next preceding and next following the one under consideration are infinitely far removed, we have the case of a discrete sound. As this limiting case is approached the fundamental frequency becomes smaller and smaller and the component frequencies, which are multiples of it, are separated by infinitesimal frequency differences. While the amplitude of each component also decreases, the number of components increases at such a rate that the aggregate energy of all the components within a given frequency range remains finite. In this way, the distribution of the sound energy over various frequencies—that is, the "energy spectrum"—can be obtained.

It is evident, then, that when an aperiodic complex sound is resolved mathematically there results an infinity of component tones, each having a characteristic intensity and phase. If an observer were capable of an equally complete resolution he would have at his disposal an infinity of sets of data from which an infinity of images could be formed. In the absence of distortion these should all coincide.

Practically, of course, no such refinement of resolution is possible. The ability to distinguish differences in pitch varies from person to person, but the minimum intervals employed in musical composition probably give a rough measure of the normal resolving power of the ear. Even with this limitation the broad sound spectrum, such as an irregular sound produces, is capable of yielding a very large number of separable components; and hence a large number of individual images. It is this fact—that with a very complex sound the number of independent determinations of the image is limited only by the resolving power of the observer—which makes his accuracy of binaural location as well as his sense of certainty much greater for such sounds than for pure tones.

So long as the images of all the components coincide, it is of little importance how fine the resolution is, for further refinement only serves to increase the sense of certainty by adding to the volume of accordant evidence. However, when the images are not in agreement the problem is more complicated and the degree of resolution becomes important. Here also purely physical considerations cease to be adequate and psychological factors must be considered similar to those involved in the location of a pure tone for which the intensity ratio and phase difference do not correspond to any actual source. When an observer is faced with discordant results he must make some subconscious judgment. For small discrepancies such as occur in every-day experience, he probably assumes those images which depart most from the rest to be misplaced because of distortion during transmission and so either corrects or ignores them. If the discrepancies are large he may find it difficult on the ground of experience to believe that so much distortion could occur. In such an event he will most likely form several images from different components or in extreme cases lose the sense of location altogether.

Bowlker found separate images to occur experimentally both for band music, which approaches a collection of tones and for the irregular barking of dogs. He placed tubes of unequal length to his two ears thereby upsetting the normal diffraction around the head and interposing a longer path on one side than on the other. Obviously, the distortion produced in this manner is of a type not likely to be met in every-day life and affects different frequencies in widely different fashions. He reports that when listening to "a band of three or four instruments played in the open—the notes will be found to be scattered over a wide range, most being to the side of the short tube, some being in front and some being to the side of the long tube. In listening with such a pair of tubes to two dogs furiously barking

the effect is at first quite alarming—one seems to be in the middle of a pack of dogs some of which are rushing viciously at one's throat."

An illustration of failure to form any image is found in a phenomenon observed in the use of binaural compensators for determining the direction of submarine sounds. The sound is picked up by two submarine telephone transmitters and led to the ears through independent paths. By adjusting the lengths of the paths the image can be shifted from side to side and for practical purposes the setting of the instrument is made by bringing the image exactly to the middle. A fairly definite sound image is formed, but observers report that part of the sound does not merge into this sound image and move in response to the adjustment, but instead appears as a diffuse background of noise.[6] This may be explained on the assumption that, while the images formed from most of the sound components agree sufficiently well that the observer corrects them to a single position, certain components are so distorted by resonance effects inherent in the apparatus that their images are scattered more or less at random. The lack of agreement among any considerable number of these prevents the formation of a second image and causes the sense of diffusedness.

As the distortion becomes still more extreme we should expect the experimental results to depend more and more upon the observer's power of resolution, for as the distortion is progressively increased a condition must finally be reached where the positions of the images are appreciably different for two components whose frequencies are so nearly alike as to make their recognition as separate tones difficult if not impossible. This condition actually occurred in an experiment of Baley's with a sound consisting of a mixture of sustained tones. Its effect on the listener is interesting from the standpoint of subconscious readjustment of discordant data.

Baley's [7] experiment consisted in applying a number of sustained tones to one ear of a musically trained observer and a number of different tones to the other ear, and testing his ability to assign them to their proper sides. So long as the intervals between the tones were fairly large, the observer never failed to locate them correctly. Considering the entire stimulus as a complex sound we may think of the observer as locating the tones individually and finding them to fall definitely into two groups whose images are located one at each ear.

[6] This interesting phenomenon was called to our attention by Mr. Richard D. Fay of the Submarine Signalling Corporation who tells us that it has been noted by a large number of observers.

[7] Stephan Baley: *Zeit. f. Psychol. u. Physiol.*, v. 70, 1914, p. 347.

However, when he used six tones which were separated from each other by a single tone interval, the separate components could not be distinguished and a painful sensation was produced. The observer was apparently faced with the situation that to make the observed intensity ratios and phase differences correspond to a single source would involve extremely large corrections in the observed data. On the other hand, his power of tone resolution was insufficient to separate the components and assign them to different sources. It is not surprising, then, that the difficulty manifested itself by painful sensations. While this illustration is taken from an extreme condition of laboratory experiment and may appear to have little bearing on the every-day location of sounds, it is really significant because of the manner in which it illustrates the importance of psychological factors in all cases in which the sound waves are distorted.

Resumé

In the foregoing discussion an attempt has been made to bring out the main features involved in extending the theory of the binaural location of pure tones to cover, qualitatively at least, the location of complex sounds. It has virtually been assumed that the latter involves three processes: first, the resolution of the sound into its component tones; second, the independent (generally subsconscious) location of each separate component; and third, the formation of a conscious judgment of the position of the source based on the locations of the individual images. The greatly increased amount of data available when the sound is complex has quite different effects on the final result according as the different images do or do not coincide. If they do, the accuracy of location and the sense of certainty are increased. If they do not, confusion arises, subconscious corrections are called for, and the final result is likely to depend very considerably on the psychological processes and individual prejudices of the particular observer.

The Heaviside Operational Calculus

By JOHN R. CARSON

SYNOPSIS: The art of electrical communication owes a great and increasingly recognized debt to Oliver Heaviside for his work in developing and emphasizing a correct theory of electrical transmission along wires and in particular for his insistance on the importance of inductance. His operational methods of solving the differential equations which are fundamental of the theory of electric circuits, although not widely known, are important. These methods are peculiarly applicable to many important problems of electrical transmission. The present paper, while theoretical in character, therefore deals with a subject of practical importance to the communication engineer.

Without attempting to give any adequate idea of the striking originality and ingenuity of Heaviside's methods, his operational calculus may be very briefly explained as follows. Problems in electric circuit theory are described by a set of differential equations involving the differential operator $\frac{d}{dt}$. These differential equations may be reduced formally to algebraic equations by replacing the differential operator by the symbol p and by this expedient a purely symbolic solution is obtained. This symbolic solution is called the *operational formula* of the problem.

In order to interpret the purely symbolic operational formula, Heaviside proceeded as follows: By direct comparison of the operational formula of specific problems with their known explicit solutions he was led to assign a definite significance to the operator p. Thereupon, he obtained by induction generalized specific criteria or rules for solving the operational formula.

The present paper, by attacking the problem from a different standpoint, shows that the Heaviside operational formula is a shorthand equivalent of an integral equation from which the methods and rules of his operational calculus are deducible.—*Editor.*

A VERY interesting and by no means the least valuable part of Heaviside's researches relates to operational methods of solving the differential equations of a class of physical problems of which electric circuit theory problems are typical; in fact Volume II of his Electromagnetic Theory is almost entirely devoted to this subject. The methods of solution which he originated and employed are of extraordinary directness and simplicity in a very large class of problems in applied mathematics. In fact it would be difficult to exaggerate the value of his work along this line, and nowhere is it more immediately and usefully applicable than in the theoretical problems of electro-technics.

Heaviside is, however, by no means easy reading and, in spite of the considerable number of published studies relating to his operational calculus, it is less generally understood and applied than its value warrants. The writer has had occasion to apply Heaviside's methods quite extensively in electrical problems and in the course of his study was led to a general formula which to him at least, has proved useful in interpreting and rationalizing the operational cal-

culus. This formula is presented and discussed in the present paper with the hope that it may be of service to students of Heaviside in understanding and applying his methods. The paper may also serve as a brief recapitulation of some of the outstanding methods applicable to the solution of problems in electric circuit theory.

The problems to which Heaviside applied his operational calculus relate to the oscillations of electrical and mechanical systems which can be described by a system of linear differential equations with constant coefficients or to such partial differential equations as *the telegraph equation*. The system is supposed to be set into oscillation *from a state of equilibrium* by suddenly applied forces and the operational formula gives the resulting behavior of the system.

The type of problem to which the operational calculus is applicable and Heaviside's method of solution may be illustrated in a sufficiently general manner for didactic purposes in connection with the solution of the system of equations

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= F_1(t) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + \ldots + a_{nn}x_n &= F_n(t) \end{aligned} \tag{1}$$

where the coefficients ajk are in general polynominals of the form

$$a_{jk} = \alpha_{jk} + \beta_{jk}\frac{d}{dt} + \gamma_{jk}\frac{d^2}{dt^2} + \ldots$$

and the α, β, γ coefficients are constants.

We are concerned with the determination of the variables $x_1 \ldots x_n$ as functions of the independent variable t for the following boundary conditions; the known functions $F_1 \ldots F_n$ *and the variables* $x_1 \ldots x_n$ *are identically zero for values of the independent variable* $t \angle 0$. In other words, the system is initially in a state of equilibrium when the "forces" $F_1 \ldots F_n$ are applied. These boundary conditions are extremely important in physical problems.

Owing to the linear character of the equations we may without loss of generality set all the F functions equal to zero except one, say $F_1(t)$, and write

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= F(t) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + \ldots + a_{nn}x_n &= 0. \end{aligned} \tag{3}$$

The solution of equations (3) for the prescribed boundary conditions may be made to depend on the auxiliary equations in the auxil-

iary variables $h_1 \ldots h_n$,—

$$\begin{aligned} a_{11}h_1 + \ldots + a_{1n}h_n &= 1 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \quad (t \geq 0) \\ a_{n1}h_1 + \ldots + a_{nn}h_n &= 0. \end{aligned} \tag{4}$$

The function on the right hand side, written, in accordance with the Heaviside notation, as unity is identically zero for $t<0$ and unity for $t \geq 0$ and $h_1 \ldots h_n$ are identically zero for $t<0$.

It may then be shown that [1]

$$x_j = \frac{d}{dt}\int_0^t F(t-y)\, h_j(y)\, dy, \qquad (j = 1, 2, \ldots n) \tag{5}$$

so that the solution of (3) depends entirely on (4).

Equations (4) formulate the problem actually dealt with by Heaviside who did not explicitly consider the more general equations (3). His method of attack was as follows; Writing p^n for the differential operator d^n/dt^n equations (4) become formally algebraic and yield a purely symbloic solution

$$h_j = \frac{1}{H_j(p)}. \tag{6}$$

Equation (6) is the Heaviside operational formula; as it stands, however, it is purely symbolic and the problem remains to find the significance of the equation and to deduce therefrom the value of $h = h(t)$ as a function of t.

Heaviside's method from this point on was one of pure induction. From the known solution of specific problems he inferred general rules for expanding and interpreting the operational formula: the body of rules thus developed for solving the operational equation may be appropriately termed the Heaviside Operational Calculus.

The contribution of the present paper to the theory of the Heaviside operational calculus depends on the following proposition and its immediate corollary.[2]

The differential equations (4), *subject to the prescribed boundary conditions, may be written as*:

$$\begin{aligned} h_j &= 0, \qquad \text{for } t<0 \text{ and } j = 1, \ldots n, \\ \frac{1}{pH_j(p)} &= \int_0^\infty e^{-pt} h_j(t)\, dt, \qquad \text{for } t \geq 0. \end{aligned} \tag{7}$$

The integral equation is an identity for all positive real values of p *and consequently determines* $h_j(t)$ *uniquely.*

[1] This formula has been established in previous papers. It is briefly discussed in Appendix I.

[2] See Appendix I.

It follows as an immediate corollary that the Heaviside operational equation

$$h = 1/H(p) \tag{8}$$

is merely a short-hand or symbolic equivalent of the integral equation

$$\frac{1}{pH(p)} = \int_0^\infty e^{-pt} h(t)\, dt. \tag{9}$$

The significance of the operational equation and the rules of the Heaviside operational calculus are therefore deducible from the latter equation. The whole problem is thus reduced to the purely mathematical problem of solving the integral equation.

It should be remarked in passing that, while the Heaviside operational calculus has been elucidated in connection with the solution of a set of differential equations involving a finite number of variables, it is not so limited in its applications. It is applicable also when the number of variables is infinite and to such partial differential equations as *the telegraph equation*. The foregoing theorem applies also to all such physical problems where an operational formula $h = 1/H(p)$ is derivable.

Before discussing the solution of the integral equation (9) and deducing therefrom some of the rules of the operational calculus, a simple but interesting and instructive example of the way the opertional formula is set up will be given.

Consider a transmission line of infinite length along the positive x axis and let it have a distributed inductance L and capacity C per unit length. Let a unit voltage be applied to the line at the origin $x = 0$ at time $t = 0$; required the line current I and voltage V at any point x at any subsequent time t.

The differential equations of the problems are

$$L\frac{\partial}{\partial t} I = -\frac{\partial}{\partial x} V,$$

$$C\frac{\partial}{\partial t} V = -\frac{\partial}{\partial x} I.$$

Replacing $\frac{\partial}{\partial t}$ by p, we get

$$I = \sqrt{\frac{C}{L}}\, e^{-\frac{px}{v}} V_0,$$

$$V = e^{-\frac{px}{v}} V_0,$$

where $v = 1/\sqrt{LC}$ and V_0 is the line voltage at $x = 0$.

Now by the conditions of the problem V_0 is zero before, unity after time $t = 0$; hence the foregoing equations are operational formulas and by (9)

$$\frac{1}{p}\sqrt{\frac{C}{L}}\; e^{-\frac{px}{v}} = \int_0^\infty e^{-pt} I_x\,(t)\,dt,$$

$$\frac{1}{p}\, e^{-\frac{px}{v}} = \int_0^\infty e^{-pt} V_x\,(t)\,dt.$$

The solutions of these equations are obviously

$$\begin{aligned} I_x &= 0 && \text{for } t<x/v, \\ &= \sqrt{\frac{C}{L}} && \text{for } t\geq x/v, \\ V_x &= 0 && \text{for } t<x/v, \\ &= 1 && \text{for } t\geq x/v, \end{aligned}$$

which are, of course, the well known solutions of the problem. The directness and simplicity of the solution from the definite integrals is, however, noteworthy.

By virtue of the foregoing analysis the Heaviside operational calculus becomes identical with the methods and rules for the solution of integral equations of the type

$$1/pH(p) = \int_0^\infty e^{-pt} h(t)\,dt \qquad (9)$$

to which brief consideration will now be given.

An integral equation is, of course, one in which the unknown function appears under the sign of integration; the process of determining the unknown function is the solution of the equation. Integral equations of the form of (9) were first employed by Laplace and may be referred to as equations of the Laplace type. More recently they have become of importance in the modern theories of divergent series and summability. The solution of a large number of integral equations of the Laplace type has been worked out; however the procedure is usually peculiar to the particular problem in hand. In this connection it is noteworthy that, from a purely mathematical standpoint, Heaviside's operational calculus is a valuable contribution to the systematic solution of this type of integral equations. That is to say, methods which he developed for the solution of his operational equation suggest systematic procedure in the solution of the integral equation (9), as might be expected from the relationship pointed out in the present paper.

As stated above a large number of infinite integrals of the type appearing in equation (9) have been worked out. Consequently the solution of (9) can frequently be written down by inspection. When this is not the case, however, the appropriate procedure is usually to expand the function $1/pH(p)$ in such a form that the individual terms are recognizable as identical with infinite integrals of the required type.

An interesting expansion of this kind and one which is applicable to a large number of physical problems is as follows:

Expand $1/pH(p)$ asymptotically in the form of the divergent series

$$1/pH(p) \sim \sum a_n/p^{n+1}.$$

This expansion is purely formal and the series is divergent. It is summable, however, in the sense that it may be identified with its generating function $1/pH(p)$. It is also summable in accordance with Borel's definition of the sum of a divergent series by the Borel integral [3]

$$\int_0^\infty dt\ e^{-pt} \sum a_n t^n/n!$$

This suggests that these two series are equal and consequently that

$$1/pH(p) = \int_0^\infty dt\ e^{-pt} \sum a_n t^n/n!$$

The solution is therefore

$$h(t) = \sum a_n t^n/n!$$

provided this series, which is called by Borel the associated function of the divergent expansion, is itself convergent. This is the case in all physical problems to which this form of expansion has been applied.[4]

The foregoing will be recognized as identical with Heaviside's power series solution, obtained by the empirical rule of identifying $1/p^n$ with $t^n/n!$ in the asymptotic expansion of $1/H(p)$.

Another form of solution of very considerable practical value depends on a partial fraction expression which can be carried out in a large number of physical problems. It is

$$1/pH(p) = a + b/p + c/p^2 + \sum A_k/(p - p_k)$$

[3] See Bromwich, Theory of Infinite Series, pp. 267–269.

[4] See Appendix II.

where

$$a = \left(1/pH(p)\right)_{p=\infty},$$

$$b = \left[\frac{d}{dp}\frac{p}{H(p)}\right]_{p=0},$$

$$c = \left[\frac{p}{H(p)}\right]_{p=0},$$

$$A_k = \frac{1}{p_k H'(p_k)},$$

and $p_1 \ldots p_n$ are the roots of $H(p) = 0$.

By virtue of this expansion [5] the solution is

$$h(t) = aP + b + ct + \sum \frac{e^{p_k t}}{p_k H'(p_k)},$$

where P denotes a "pulse" at the origin $t = 0$; that is,

$$P = \infty \quad \text{at } t = 0,$$
$$= 0 \quad \text{for } t > 0,$$
$$\int_0^\infty P\,dt = 1.$$

In the usual case where $a = c = 0$ and $b = 1/H(0)$, this reduces to

$$h(t) = 1/H(0) + \sum \frac{e^{p_k t}}{p_k H'(p_k)},$$

which will be recognized as the celebrated Heaviside Expansion Solution.

As illustrating the flexibility of the integral identity (9), another form of solution will be given which is often of value in practical problems where an explicit solution cannot be obtained. Suppose that $1/pH(p)$ can be written as

$$\frac{1}{pH(p)} = \frac{1}{pH_1(p)} \cdot \frac{1}{pH_2(p)}$$

and that functions $h_1(t)$ and $h_2(t)$ can be found which satisfy the equations

$$\frac{1}{pH_1(p)} = \int_0^\infty e^{-pt} h_1(t)\,dt$$

$$\frac{1}{pH_2(p)} = \int_0^\infty e^{-pt} h_2(t)\,dt$$

[5] The terms $a + c/p^2$ in this expansion were suggested by Dr. O. J. Zobel and must be included in a number of important problems in electric circuit theory.

Then the required function $h(t)$ is given by

$$h(t) = \int_0^t h_1(t - y)\, h_2(y)\, dy \tag{10}$$

by Borel's Theorem (Bromwich, Theory of Infinite Series, p. 280).[6]

As a final example of the foregoing discussion we shall consider a specific problem of some practical interest in itself and which involves Heaviside's so-called " fractional differentiation " and his resulting asymptotic solutions. The physical problem is as follows: a " unit-voltage " (zero before, unity after time $t = 0$) is applied through a terminal condenser C_0 to an infinitely long cable of resistance R and capacity C per unit length. Required the Voltage V at the cable terminals.

The operational formula of this problem is easily deduced; it is

$$V = \frac{\sqrt{p/a}}{1 + \sqrt{p/a}} \qquad \text{where } 1/\sqrt{a} = C_0 \sqrt{R/C}.$$

Consequently the integral equation can be written

$$\int_0^\infty e^{-pt} V(t)\, dt = \frac{1}{p} \frac{\sqrt{p/a}}{1 + \sqrt{p/a}}$$

$$= \frac{1}{p} \frac{1}{1 + \sqrt{a/p}}$$

Taking the last form of $1/pH(p)$, expanding asymptotically and recognizing that

$$1/p^{n+1} = \int_0^\infty e^{-pt} t^n/n!\, dt$$

$$1/p^n \sqrt{p} = \int_0^\infty e^{-pt} \frac{(2t)^n}{(2n-1)(2n-3)\ldots 1} \frac{dt}{\sqrt{\pi t}}$$

the resulting series solution can be recognized and summed as

$$V(t) = e^{at} - \sqrt{\frac{a}{\pi}}\, e^{at} \int_0^t \frac{e^{-ay}}{\sqrt{y}}\, dy$$

$$= \sqrt{\frac{a}{\pi}}\, e^{at} \int_t^\infty \frac{e^{-ay}}{\sqrt{y}}\, dy.$$

The last expansion by repeated integration by parts leads to the asymptotic series given by Heaviside. It is easy to show, also, that

[6] This formula is quite useful; it is applied in the solution of the last example of this present paper.

the series is truly asymptotic in the sense that the error is less than the last term included.

Another mode of procedure, however, suggests itself, which, by the aid of equation (10) gives the solution directly without series expansion. We have

$$\frac{1}{pH(p)} = \frac{1}{p-a} - \frac{1}{p-a}\sqrt{\frac{a}{p}},$$

$$= \int_0^\infty e^{-pt}\left[h_1(t) - h_2(t)\right] dt,$$

where

$$\frac{1}{p-a} = \int_0^\infty e^{-pt} h_1(t)\, dt$$

and

$$\frac{1}{p-a}\sqrt{\frac{a}{p}} = \int_0^\infty e^{-pt} h_2(t)\, dt.$$

Consequently $h_1(t) = e^{at}$, and since

$$\frac{1}{\sqrt{p}} = \int_0^\infty e^{-pt}\sqrt{1/\pi t}\; dt$$

it follows at once from (10) that

$$h_2(t) = \sqrt{\frac{a}{\pi}}\, e^{at}\int_0^t e^{-ay}\frac{dy}{\sqrt{y}}.$$

The solution $h(t) = h_1(t) + h_2(t)$ agrees with the preceding derived from the asymptotic expansion, and is considerably more direct and simple.

It is interesting to compare this solution with Heaviside's own operational solution (Electromagnetic Theory Vol. II, p. 40) which amounts to the following. The operational formula is written

$$V = \frac{p}{p-a} - \frac{1}{p-a}\sqrt{ap}.$$

The first term is discarded altogether and the second written as

$$V = \left(1 - \frac{p}{a}\right)^{-1}\sqrt{p/a}$$

$$= \left(1 + \frac{p}{a} + \left(\frac{p}{a}\right)^2 + \ldots\right)\sqrt{p/a}.$$

Identifying $\sqrt{p}$ with $1/\sqrt{\pi t}$ and p^n with d^n/dt^n the expansion becomes

$$V = \left(1 - \frac{1}{2}\left(\frac{1}{at}\right) + \left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\left(\frac{1}{at}\right) - \ldots\right)\sqrt{\frac{1}{\pi at}},$$

which agrees with the foregoing and is the actual asymptotic expansion.[7]

The foregoing discussion is sufficient, it is hoped, to show the place of the integral formula (9) in relation to the Heaviside operational calculus. It is believed to be particularly applicable in connection with a number of questions relating to divergent series and solutions which Heaviside's work has raised and which have received too little attention from mathematicians.

APPENDIX I

A proof of the integral formula

$$1/pH(p) = \int_0^\infty e^{-pt} h(t)\, dt \tag{9}$$

can be made to depend very simply on the formula

$$x(t) = \frac{d}{dt} \int_0^t F(t-y) h(y)\, dy. \tag{5}$$

This equation may be regarded as well established and can in fact be deduced in a quite general manner by synthetic arguments. It is derived and employed in papers by the writer (Trans. A. I. E. E., 1911, pp. 345–427, and Phys. Rev. Feb. 1921, pp. 116–134) and is deducible at once from the work of Fry (Phys. Rev. Aug. 1919, pp. 115–136).

On the basis of equation (5) the deduction of formula (9), in which, however, no pretense to rigor is made, proceeds as follows;

If the function $F(t)$ in equations (3) is set equal to e^{pt}, the complete solution (5) includes the particular solution [8]

$$e^{pt}/H(p)$$

which involves t only through the exponential term. The complete solution must, therefore, admit of reduction to the form

$$x(t) = e^{pt}/H(p) + y(t) \tag{a}$$

where $y(t)$ is the complementary solution.

[7] The procedure by which Heaviside arrived at the foregoing asymptotic solution is not, however, always so fortunate. For example if a terminal inductance is substituted for the terminal condenser of the preceding problem, precisely the same procedure gives an incomplete result. Heaviside recognized this and added an extra term without explanation (Elm. Th. Vol. II, p. 42) but his solution appears to be doubtful in the light of some recent work by the writer in applying the formula of the present paper to the same problem.

[8] Provided $H(p) \neq o$. This restriction is of no consequence in physical problems, where the roots of $H(p)$ are in general complex with *real part negative.*

Now equation (5) may be written, when $F(t) = e^{pt}$, as

$$x(t) = \frac{d}{dt} e^{pt} \int_0^t e^{-py} h(y)\, dy$$

$$= \frac{d}{dt} \left\{ e^{pt} \int_0^\infty e^{-py} h(y)\, dy - e^{pt} \int_t^\infty e^{-py} h(y)\, dy \right\}. \qquad \textbf{(b)}$$

Now the first term of the expression involves t only through the exponential term while the second term involves t through the lower limit of the integral which ultimately vanishes and therefore includes no term involving t only through the exponential. Consequently the first term of (b) is identifiable as the particular solution of (a) and by direct equation it follows that

$$1/pH(p) = \int_0^\infty e^{-py} h(y)\, dy \qquad \textbf{(9)}$$

which is the required formula.

The most important restriction which is implicit in the foregoing is that in splitting up the definite integral of (5) we have tacitly assumed that $h(t)$ is finite for all values of t; a restriction which is necessary in order that the infinite integral shall be convergent for all positive real values of p. This condition is satisfied in all physical problems and therefore introduces no practical limitation of importance.

However, even when this restriction does not hold formula (9) may be valid and uniquely determine $h(t)$ if p is restricted to values which make the infinite integral convergent, or when the problem is such that $e^{-pt}h(t)$ is an exact derivative. As an example, suppose that

$$1/H(p) = \frac{p}{p - a}$$

where a is a real positive quantity. It may be otherwise shown that $h(t) = e^{at}$ and formula (9) becomes

$$\frac{1}{p - a} = \int_0^\infty e^{-(p-a)t}\, dt$$

which is valid when $p > a$.

APPENDIX II

The discussion in the text does not pretend to be a proof of the power series expansion in any strict sense. A more satisfactory discussion proceeds as follows:

We assume that $1/H(p)$ can be formally expanded in the series

$$\sum_0^\infty a_n/p^n$$

We shall here introduce a necessary restriction on the function $1/H(p)$. It must include no function which is represented asymptotically by a series all of whose terms are zero; that is a function $\phi(p)$ such that the limit, as p approaches ∞, of $p^n\phi(p)$ is zero for every value of n. The function e^{-p} is such a function. (See Whittaker & Watson, p. 154.)

With this restriction understood, start with the integral (9) and integrate by parts; we get

$$\frac{1}{H(p)} = h(o) + \int_0^\infty e^{-pt}h^{(1)}(t)dt$$

where $h^{(n)}(t) = d^n/dt^n h(t)$.

Now let p approach infinity; in the limit the integral vanishes and

$$h(o) = 1/H(\infty) = a_0$$

from the asymptotic expansion.

Integrate again by parts; we get

$$p(1/H(p) - a_0) = h^{(1)}(o) + \int_0^\infty e^{-pt}h^{(2)}(t)dt.$$

Now let p again approach infinity; in the limit the integral vanishes, and the right hand side, by virtue of the asymptotic expansion, approaches the limit a_1, whence $h^{(1)}(o) = a_1$. Proceeding in this manner, repeated integrations by parts establish the relation $h^{(n)}(o) = a_n$. But provided the series is absolutely convergent, then

$$h(t) = \sum h^{(n)}(o)t^n/n!$$

$$= \sum a_n t^n/n!$$

which establishes the formula.

The power series solution is applicable to a large class of physical problems and has been rigorously established under certain restrictions by other methods than that employed above (see papers by Bromwich, *Phil. Mag*, May 1920, p. 407; Fry, *Phys. Rev.* Aug. 1919, p. 115; and the writer, Trans. A. I. E. E. 1919, p. 345).

On the basis of the preceding and with the aid of formula (10), expansions of the type

$$1/pH(p) \sim \frac{1}{\sqrt{p}} \sum b_n/p^{n+1} = \int_0^\infty e^{-pt} h(t)dt$$

which occur in physical problems, can be dealt with. For since

$$\sum b_n/p^{n+1} = \int_0^\infty dt\, e^{-pt} \sum b_n t^n/n!$$

and

$$\frac{1}{\sqrt{p}} = \int_0^\infty e^{-pt}\, dt/\sqrt{\pi t},$$

it follows from (10) that

$$h(t) = \frac{1}{\sqrt{\pi}} \int_0^t \frac{dy}{\sqrt{t-y}} \sum b_n y^n/n!$$

$$= \frac{1}{\sqrt{\pi t}} \sum \frac{b_n (2t)^n}{(2n-1)(2n-3)\ldots 1}.$$

The Physical Characteristics of Audition and Dynamical Analysis of the External Ear

By R. L. WEGEL

SYNOPSIS: This paper discusses some of the characteristics of the ear which have become important in the design and development of telephone apparatus and circuits. The field of audition, bounded by the curves of minimum and maximum loudness as functions of frequency, has been determined for a large number of ears, and the smaller included area most used in speech has been mapped. The nature of these fields in certain cases of abnormal hearing has also been determined and the conditions which must be observed in designing apparatus to satisfactorily relieve deafness are discussed.

The sensitivity of the ear is given in terms of the r. m. s. pressure measured by a calibrated condenser transmitter. It is printed out in the appendix that this pressure is not necessarily equal to that which, when applied to the ear drum, would just give rise to the sensation of sound. However, it is the nearest approach to the value of this pressure which can be determined at present, and as the dynamical properties of the ear become more fully known it is pointed out how the relation between the two pressures can be more accurately stated.—*Editor.*

1. *Introduction.* It has become important in the design and development of telephone apparatus and circuits to know quantitatively the various functional characteristics of the ear since the ear is an important dynamical unit in the long series of vibration transmitting apparatus constituting a telephone system. A complete analysis of this problem involves not only the properties of the physical circuit, but also the characteristics of the ear and voice and of the air passages between the mouth and transmitter and between the ear and receiver. It is the purpose of the present paper to discuss some of the characteristics of the ear and its outer air passages.

Much has been learned about the normal ear by the investigation of the characteristics of abnormal ears. This has incidentally had an application to otological diagnosis and the design and building of amplifying apparatus for the deaf.

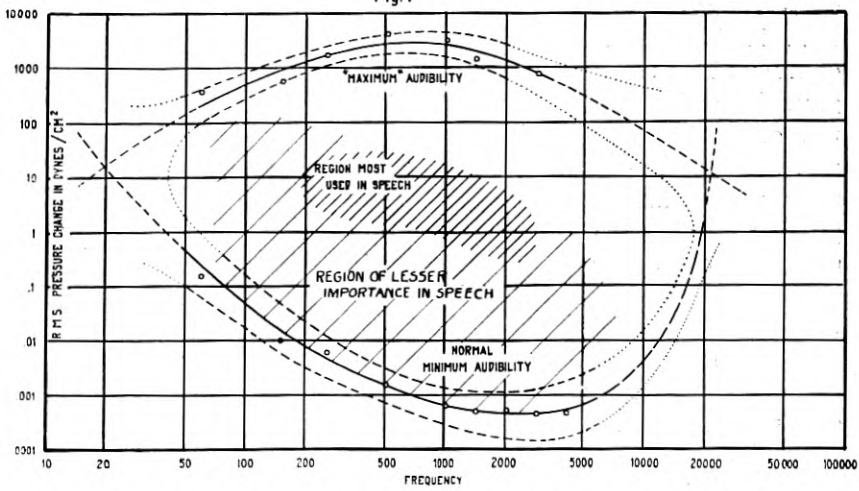
This paper is a summary of the conclusions reached to date regarding the absolute sensitivity of normal and abnormal ears, the maximum sound to which the ear can accommodate itself, the much discussed points of "upper and lower frequency limits of audition," the "quality" of audition, a brief mention of the binaural sense and the principles of rigorous dynamical analysis of the ear as a mechanism. A brief description of the apparatus used is also given.

The function of the auditory sense is to detect sounds of various kinds and wave shapes varying over a range of pressure on the ear drum of from about .001 to 1,000 dynes per cm^2 and over a considerable part of this range to differentiate with certainty between complex

sounds so nearly alike that no existing physical apparatus can separate them. The binaural feature adds a sense of orientation with respect to a source and uniform sensitivity for sounds approaching from different directions. The abnormal auditory sense may be regarded as lacking more or less in (a) range of sensation (frequency and intensity); (b) quality of sensation in various regions of the range; (c) the binaural sense. Apparatus and methods have been developed by means of which the outstanding features of these functions can be measured and to a limited extent compensated for.

2. *Minimum Audibility.* Fig. 1 shows a plot of the logarithmic average of minimum audible pressure on 72 normal ears taken through-

Fig. 1



AUDITORY SENSATION AREAS

out a range of frequency from 60 to 4,000 cycles.[1] Both the intensity and frequency scales are logarithmic. Although all skew errors in the determination of the average curve have not been eliminated, an investigation has shown that they are so small as not to affect the utility of the curve for the purpose of measuring deafness. Among the errors which obviously tend to raise this curve might be mentioned, noise in the observing room, abnormality of hearing, lack of attention, and low mentality of the observer. Care was taken to reduce these errors to a minimum without actually making separate

[1] This curve has already been published; The Frequency Sensitivity of Normal Ears, by H. Fletcher and R. L. Wegel, *Proceedings of the National Academy of Sciences*, January, 1922, and *Physical Review*, June, 1922.

quantitative measurements of each of them on a rigorous statistical basis.

The statistical deviation from the mean varies irregularly with frequency; very likely this is due mostly to the external anatomical variations in ears which cause deviations in the dynamical constants of the transmission system from sound source to the ear drum. The dotted lines following the curve of minimum audibility represent approximately the "standard deviation."

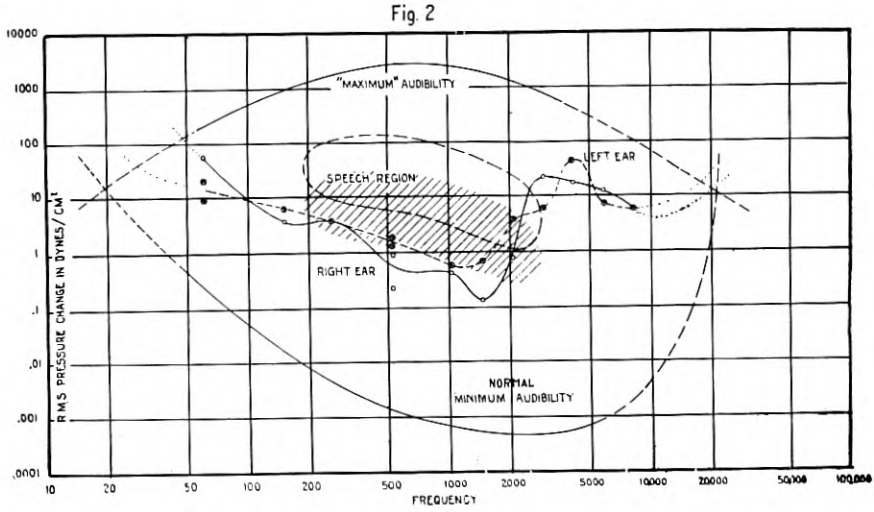
3. *Maximum Audibility.* The curve marked "Maximum audibility" represents the logarithmic average pressure on 48 normal ears required to produce the sensation of feeling. This represents the threshold of feeling in the same way that the minimum audibility curve represents the threshold of audition. A sound much louder than this is painful. The measurements were taken through a range of from 60 to 3,000 cycles. The standard deviation lines are also given from which it will be seen that this curve is quite as definite as that of minimum audibility. While this point of feeling probably has no relation to the auditory sense it does serve as a practical limit to the range of auditory sensation. A few observations indicate that people with abnormal ears have a point of feeling sound which is not greatly different from that of normal ears, but this, of course, depends on the type of abnormality. The intensity for feeling is about equal to that required to excite the tactile nerves in the finger tips.

4. *Lower and Upper Frequency Limits of Hearing.* The curves of minimum and maximum audibility in Fig. 1 will be seen to have been extrapolated to the points of intersection at high and low frequencies. The feeling sensation in the middle range of frequency is first a tickling sensation and then becomes acutely painful as the loudness is increased. As the frequency is decreased the sensation of feeling becomes milder until frequencies around 60 cycles it is sensible as a flutter, but still quite different from the sense of audition. As the frequency is still further decreased to a point where the hearing and feeling lines appear to intersect, it is difficult to distinguish between the sense of hearing and that of feeling. The low point of intersection of the two normal curves of minimum audibility and feeling sense may, therefore, be taken arbitrarily as the lower tone limit of audibility. For frequencies lower than this it is easier to feel than to hear the air vibration. The point of intersection cannot be determined by direct observation due to the difficulty in distinguishing between the two sensations. A similar intersection of the two curves occurs at some very high frequency. Sound waves of frequencies below the lower intersection and above the upper intersection are more easily sensed

by feeling. Sound waves between these limits are more easily sensed by audition.[2]

This suggests a rational way of defining and determining the two frequency limits of audibility. Measurements of these limits which have been made in the past are questionable because the intensity factor has been neglected. At the lower limit of audibility the excursions of the diaphragm and ossicles of the middle ear are probably so large that the nerves feeding these movable parts are stimulated. This observation at low frequencies as indicated in this work lends color to the hypothesis of otologists that abnormalities in the hearing

Fig. 2



SENSATION AREAS (CHIC)

of low frequencies are due to pathological conditions in the middle ear. This point is probably related to the tests on flexibility of the ear drum or ossicular chain due to the application of air pressure as observed by otologists in examination. Loss of sensitivity at low frequencies is considered an indication of obstructive deafness if there is no loss at high frequencies.

5. *Sensation Area.* From the combined standpoint of utility and logic the logarithmic relation between stimulus (pressure variation) and sensation can be assumed. The elliptical area between the two curves may then be taken to represent an area of sensation which is

[2] The extrapolation upward of the curve of minimum audibility is consistent with some recent observations of Mr. C. E. Lane at the University of Iowa, *Physical Review*, May, 1922.

characteristic of the normal ear. Any point within this area represents a definite auditory sensation in frequency and intensity. The area of sensation is analogous to the field of vision of the eye. The part of this area which is most utilized in the interpretation of speech is represented approximately by the shaded area in Figs. 1 and 2 and corresponds in a way to the center of the field of vision. A normal listener tries, by keeping at a certain distance from a speaker, to bring this part of his sensation area into play in the same way that when examining an object he directs his eyes so that it falls in the center of the field of vision.

An abnormal ear may be regarded as having an area of sensation which is smaller than the normal area but included within it. Fig. 2 is a plot of the minimum audibility of the right and left ears for a man (CHK) having a "catarrhal" deafness. The areas between these curves of minimum audibility and the curve of feeling are his areas of sensation. It will be seen that CHK retains about 50 or 60 per cent of the normal amount of sensation. He hears and interprets conversation with some difficulty.

Since the CHK curves pass through the speech region, part of it is entirely inaudible and the remainder is near minimum audibility for him. In order to make him hear well, the speech area must be raised to a higher level of intensity or loudness as indicated by the dotted curve.

In general it takes a loss of about 20 per cent of the sensation area to become noticeable and much more is disagreeable. A loss of 50 per cent requires the use of deaf apparatus. A loss of 75 per cent can be aided considerably by the use of high powered amplifying apparatus.

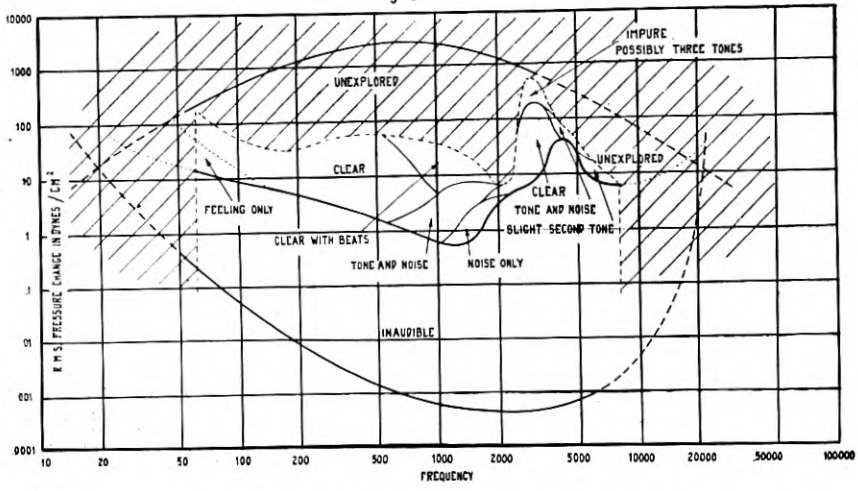
6. *Importance of Various Intensities in Speech.* It is interesting to speculate on how CHK interprets speech. It has been shown [3] that the intensity of speech may be varied over perhaps 70–80 per cent of the range of sensation without serious loss of intelligibility to the normal ear. As the sound intensity is decreased, the intelligibility drops very suddenly to zero at minimum audibility. A similar drop is to be expected at an intensity so loud as to be painful. It is evident, therefore, that the range in which speech is intelligible for CHK is very considerably limited as compared to normal. It is possible to design a deaf set which raises the intensity of the principal speech region to any desired place within the abnormal sensation area and so in a measure, compensate for this narrowed range. The region in Fig. 1 "Region of Lesser Importance in Speech," corre-

[3] H. Fletcher, *Journal of the Franklin Institute*, June, 1922.

sponds to stimuli in conversation of lesser energy content, such as the minor shadings and fainter consonant sounds. While it is physically possible to produce an amplification of speech so that this region is raised into the diminished area it is impracticable to do so because of the pain which would be caused by the louder components. A diminution in sensation area can, therefore, be only partially compensated for. In case the area is extremely narrow a deaf set furnishing optimum volume can only serve as an aid to lip reading.

7. *Quality of Hearing.* The sensation of a normal ear at any point in the auditory sense range (Fig. 1) may be described by a number of different adjectives, such for example as "clear," "musical," "even,"

Fig. 3



QUALITY AREAS - LEFT EAR (C.H.K)

"sustained," "smooth," "pure," etc. Such a description may, in fact, be taken as a reasonable indication that the quality of sensation at the point in question is normal. Abnormal ears sometimes experience a subjective degeneration of quality of pure stimuli which they describe as "rough," "harsh," "sharp," "buzzing," "vibrating," "hissing," etc. This subjective degeneration is independent of any tinnitus or head noises which the patient may have. Fig. 3 shows various regions of the sensation area which are degenerated in the case of CHK, left ear. The shaded area was not explored. The boundaries of the degenerated regions are usually more sharply marked than the outer boundaries of the sensation area. The sensations in these areas are so radically different from the sensation of a

pure tone that it is with difficulty that the patient is convinced that the stimulation is the same pure tone to which he has been listening at the other intensities. The subject of these tests is a violinist and capable of better descriptions and finer distinctions than average.

Since all speech sounds may be considered as stimuli composed of various frequency components of certain intensities, the sensation caused by such a sound may be represented on this plot by points, or by a line provided the sound has a band spectrum. If the points or line, falls within the sensation area the sound is audible. It is easy to see that if the points or the part of the line which represent those frequency components most essential to interpretation of the sound, fall within any of these abnormal areas, the sound is very likely to be misinterpreted. This adds a further source of loss in intelligibility to that already observed due to a narrowing of the sensation range. When an amplifying deaf set is designed, due care should be taken to raise the principle speech region in such a way as to cause a minimum overlapping with the abnormal areas.

Many practically normal ears have very small abnormal areas. They have always been found near minimum audibility and if this is always true would, therefore, have little influence on the hearing of the individual. They seem to be associated with "catarrhal" conditions although this cannot be stated positively.

8. *Binaural Sense.* The normal individual has learned to interpret the differential sensations of the two ears to advantage. It helps him to locate the direction from which sounds come, to have a sort of sense of orientation with respect to sounds approaching from different directions, and whether for physical or for purely psychological reasons to assist in focusing of the attention on one sound of a large number. Two ears also assist the individual in perceiving equally well sounds coming from different directions.

When one ear becomes less sensitive, even though the loss is small, the use of the binaural sense disappears and after a time is not missed, the subject depending upon other means of locating sounds. For the binaural sense to be most effectively utilized it is necessary that the ears be very nearly alike. When a binaural deaf set is made and fitted to a person with compensating sensitivity for the two ears so that both hear the sounds equally loud, the sensation is usually so novel, that if the patient is actually able to experience a binaural sensation he is very much pleased. Usually, however, he has not used his binaural sense for so long a time that it takes a considerable amount of practice before he is able to have binaural experiences. It may be noted in this connection that the same experience is en-

countered in fitting the eyes with glasses. It is found that people with two eyes which are slightly different do not see stereoscopically but if glasses are made so as to compensate and make the eyes nearly alike, it usually takes a certain amount of practice before the sense of perspective can be brought back.

APPENDIX

9. *Experimental Methods.* In order to discuss the principles of ear sensitivity measurement on a rigorous dynamical basis, it will perhaps, be clearer to describe briefly the experimental method used in producing known sound pressure in the ear canal at the various frequencies and intensities.[4]

As a source of sound, a small thermal receiver unit was used. This consisted of about twenty very small loops of Wollaston wire contained in a brass case small enough to be inserted in the external ear canal and entirely stop it up. In the average ear a volume of about 1 cm.^{3} of air is included between it and the drum membrane. A direct heating current is passed through the receiver and an alternating current of the desired frequency and intensity is superimposed to modulate its temperature. This modulation in temperature causes alternate expansion and contraction of a very thin film of air covering its surface and so produces alternations in pressure in the ear canal of the frequency of the impressed alternating current. The intensity is proportional to this alternating current if it is maintained small compared with the direct current. This arrangement permits of producing alternating or sound pressure on the ear drum with a comparatively simple dynamical relation between the source of sound and the ear drum. The thermal receiver is also dynamically one of the simplest sources of sound known.

The sound or alternating pressure was determined by calibration. This was done by inserting the thermal receiver in an air cavity of 1 cm.^{3} volume in front of a condenser transmitter diaphragm by which the alternating pressure developed by a given current in the receiver could be measured.[5] By measurement of the current for minimum audibility or "maximum audibility" or for any other intensity the pressure in the ear canal is determined.

10. *Dynamical Principles of Ear Measurements.* From a dynamical standpoint the phrase "sensitivity of the ear" as it is usually used is

[4] For further details, see "The Frequency Sensitivity of Normal Ears," H. Fletcher and R. L. Wegel, *Physical Review*, June, 1922.

[5] For the method of calibration of the condenser transmitter, see article by H. D. Arnold and I. B. Crandall, *Physical Review*, July, 1917.

rather indefinite. When a figure is given in ergs per second, the rate of flow of energy through an area equal to that of the ear opening in an unobstructed wave, is commonly meant. This has no simple relation, theoretically at any rate, to the net rate of flow of energy into the ear when the head is placed as an obstruction to the wave. The distortion of the sound field by the head varies greatly with frequency. Similarly, there is no simple relation between the energy flowing into the ear and that transmitted to and absorbed by the ear drum or by the cochlea. In the experiments recorded above, attention was paid to the experimental set-up so as to make the figures given have a more definite dynamical significance. Sensitivity is given in terms of the alternating (root mean square) pressure to produce a minimum audible sensation. The term "pressure" has so far been used in a rather loose sense. Just why this is so will be seen from the following argument.

The simplest method of describing the constants of a mechanical system is in terms of the components of its mechanical impedance and their relative dispositions in the same way that an electrical circuit is described by giving its resistance, inductance and capacity and the way in which they are connected. In a linear system having a single degree of freedom, the impedance may in general be written in the form

$$Z = r + wjm + s/jw.$$

The symbols are as follows:

$j = \sqrt{-1}$,

$w = 2\pi$ times the frequency,

r = frictional resistance to motion, with respect to a stationary body and involves dissipation of energy at a rate of $\dot{x}^2 r$ where $\dot{x}$ is the root mean square value of the relative velocity. The velocity $\dot{x}$ will be assumed simply sinusoidal in what follows,

m = mass or inertia constant involving an average storage kinetic energy of $\dot{x}^2 m$ through one cycle,

s = stiffness constant involving an average storage of potential energy through one cycle of $\dot{x}^2 s/w^2$.

If the r. m. s. alternating force acting is F, the motion at any frequency is given by

$$\dot{x} = F/Z.$$

In analyzing a system in which the constants may be considered

as "lumped," that is in which, for the purpose of practical solution, a finite number of degrees of freedom may be assumed, the method is to find the most useful way of "lumping" these constants. The motions are then represented by a series of equations, one for each degree of freedom, between the forces acting and the impedances and velocities. The determinant of the coefficients of these equations is the Lagrange determinant of the system. The only caution to be observed in lumping the constants is that the reciprocal relation, which is a property of any linear system holds also for the physical system which the assumed Lagrange determinant is supposed to represent.

The method may be illustrated by the following application to the sensitivity measurements described above.

The dynamical system used in calibration with the condenser transmitter consists of three parts:

(a) The very thin pulsating air film over the thermal receiver filaments. The expansion of air around the wires is represented by the "diffusion" equation, the solution of which in such a case of cylindrical symmetry is given as a Bessel's function of the distance from the wire.[6] This wave is so quickly damped in travelling away from the wire as to be negligible beyond the first zero point of the Bessel's function. The vibrating system of this receiver may then be considered as a cushion of air next to the wire of a thickness a little less than the first half wave length of the heat wave. The thickness of this cushion is an inverse function of the frequency.

(b) The air chamber between the thermal receiver and condenser transmitter diaphragm having a volume of 1 cm.^{3} and enclosed by practically unyielding walls with no openings.

(c) The condenser transmitter diaphragm, being stretched very tightly and air damped. It may also be regarded as unyielding, or as having an impedance very high compared to that of the connecting air chamber.

If for simplicity the mass reaction and internal losses in the air chamber may be neglected, it may be seen that the moving system of the receiver may be regarded as a weightless and frictionless "diaphragm" surrounding the wires at a distance equal to the effective thickness of the active air film and may be shown to have an intrinsic stiffness reactance of:

$$Z_1 = \frac{s_1}{jw} = \frac{\gamma p_0 a_1^2}{jwv_1}.$$

[6] See Wente, *Physical Review*, April, 1922.

In this expression, γ is the adiabatic constant of air, p_0 the atmospheric pressure, a_1 the area of the fictitious diaphragm, and v_1 the volume of air in the film. This diaphragm is loaded externally by the air chamber, when the transmitter diaphragm is prevented from moving, by a stiffness reactance of

$$M_1' = \frac{S_1}{jw} = \frac{\gamma p_0 a_1^2}{jwv},$$

in which v is the volume of the air chamber. Similarly, the load of the air chamber on the transmitter diaphragm, whose area is a_2, is

$$M_2' = \frac{S_2}{jw} = \frac{\gamma p_0 a_2^2}{jwv}.$$

The air chamber also acts as a mutual impedance between the thermal unit and the transmitter diaphragm equal to

$$M_{12}' = \frac{S_{12}}{jw} = \frac{\gamma p_0 a_1 a_2}{jwv}.$$

If, further, the intrinsic impedance of the transmitter diaphragm, which may be any function of frequency, be denoted by Z_2, the equations of motion of the system may be written

$$F = (Z_1 + M_1')\,\dot{x}_1 - M_{12}'\,\dot{x}_2,$$
$$0 = -M_{12}'\,\dot{x}_1 + (Z_2 + M_2')\,\dot{x}_2.$$

In these equations, F is the force acting on the thermal receiver "diaphragm" due to alternating current, $\dot{x}_1$ the velocity of its motion and $\dot{x}_2$, the velocity of motion of the condenser transmitter diaphragm.

A rough calculation shows that v_1 is very small compared with v, so that S_1 may be neglected compared to s_1 and that the reaction $M_{12}'\dot{x}_2$ may be neglected. The analysis of the condenser transmitter shows Z_2 to be very large compared to M_2'. These equations may then be rewritten

$$F = Z_1\,\dot{x}_1,$$
$$M_{12}'\,\dot{x}_1 = Z_2\,\dot{x}_2. \qquad (1)$$

The equations of motion, when the receiver is inserted in the ear, may be derived in a similar way. In this case, although the volume of air between the receiver and ear drum is the same as before, the

walls may yield appreciably, particularly in some frequency ranges. The mutual impedance between the receiver and ear drum, is, therefore, not necessarily a simple stiffness reactance. Also the loads due to it on the thermal receiver and ear drum, which in this case takes the place of the transmitter diaphragm, are not simple stiffness reactances. The constants in the case of the ear system will be denoted by the same letters as those used in the calibration but with the primes dropped, with the exception that the intrinsic impedance of the ear drum is denoted by D. D includes the reactions of the ossicles of the middle ear and the cochlea and is probably a complicated function of frequency. If, as may be expected, nature's design is efficient, then D must be of the same general order of magnitude as the load on the ear drum, M_{12}, of the ear canal. This probably constitutes the largest difference between the calibration and the observational systems. Strictly, of course, the condition for maximum power absorption by the ear drum from the air is that D be the conjugate of the impedance of the load on it due to the unobstructed ear canal. This condition is not obtained in nature because of such requirements placed on the design as protection from injury, etc.

In the case of the ear, M_1 may again be neglected, compared to Z_1, and the reactance, $M_{12}\dot{x}_2$ may be neglected. Then

$$\begin{aligned} F &= Z_1 \dot{x}_1, \\ 0 &= -M_{12}\dot{x}_1 + (D + M_2)\dot{x}_2, \end{aligned} \tag{2}$$

where $\dot{x}_2$ represents the velocity of motion of the ear drum. Suitable variations with frequency are implied in each of the "constants" of this system.

We are now in a position to see just what has been measured and called, for the sake of brevity or want of a better name, "minimum audible pressure" in the first part of this paper.

Let $\dot{x}_1$ now represent the velocity of the receiver diaphragm in both systems corresponding to that necessary to obtain a minimum audible sensation in the ear, and F the corresponding force. Then $\dot{x}_2$ will be the velocity of the ear drum corresponding to minimum audibility in equation (2). In the calibration, the pressure p' on the condenser transmitter diaphragm corresponds to $\dot{x}_1$. The total force acting on this diaphragm is $p'a'$ where now a' designates its area. Since this force is relieved by the motion of the diaphragm, it is seen from equation (1) to be equal to

$$p'\,a' = M'_{12}\,\dot{x}_1. \tag{3}$$

Similarly if the actual pressure on the ear drum is p, and its effective area, a, the total force on the ear drum $pa = D\,\dot{x}_2$. Combining equations (2) and (3) gives

$$p' = \frac{a}{a'}\frac{M'}{M}\left(\frac{M_2\,\dot{x}_2}{a} + p\right), \tag{4}$$

or

$$p = p'\frac{M}{M'}\frac{a'}{a} - \frac{M_2\,\dot{x}_2}{a}. \tag{5}$$

The pressure p is the actual pressure on the ear drum. The pressure p' is that measured and plotted in the diagram. If the walls of the ear canal and the ear drum were unyielding, p and p' would be identical for then $M = M'$ and $M_2\,\dot{x}_2/a$ would vanish. If the yield of the ear canal walls were such as to relieve half the pressure in the canal and that of the ear drum about the same, the difference would be considerably less than one of the divisions, in the diagrams, on the intensity scale. If the drum impedance D should be found to be negligible compared to its load M_2 the difference would be considerable. This, however, is hardly to be expected even through narrow ranges of frequency. If the impedances in the formulas were measured the energy flow into the ear drum could be computed.

In conclusion, the present status of the ear problem may be summarized. The philosophy of external ear dynamics has been touched on but there still remain difficult problems both theoretical and experimental. A start has been made on a sound basis in the explanation of the action of the cochlea by Roaf, " Analysis of Sound Waves by the Cochlea," *Philosophical Magazine*, February 1922. Nothing dependable has as yet been published on the action of the middle ear for audio frequencies. It is usually assumed that the various parts undergo relative displacements at audio frequencies in the same way as they react to static forces but this is very likely far from the truth.

The Theory of Probabilities Applied to Telephone Trunking Problems

By EDWARD C. MOLINA

THE Theory of Probabilities lends itself to the solution of many important telephone problems. These problems arise not only in connection with the trunking of calls but also in statistical studies which underlie the making of fundamental plans, in studies carried on in physical research and in the manufacturing of telephone apparatus.

The purpose of the present paper is to discuss certain simple types of trunking problems which can readily be handled to a sufficient degree of approximation by well-known probability methods. It would be quite impossible, within the scope of a single paper to give a complete discussion of trunking problems in general. For years [1] it has been known that light could be shed on these problems by the application of probabilities and many articles [2] have appeared on this subject; however the treatment to be found in the literature is, as yet, by no means comprehensive.

About 1905, the development of machine switching systems arrived at a stage where the relative efficiencies of different sizes of trunk groups became of prime importance.

In designing and engineering machine switching systems, it is necessary to compare the costs of various plans using trunk groups of widely different sizes, in order to choose the cheapest arrangement. Some plans use trunk groups as small as 5 and others groups as large as 90.

Machine switching development, therefore, gave a great impetus to the application of the Theory of Probabilities to telephone engineering and in the Bell System work along this line has been in progress, systematically, for many years. This work has included not only the theoretical solutions of various trunking problems, but has also involved the computation of special probability tables and collection of data by means of which theoretical results have been closely checked.

In the articles which have hitherto appeared, little or no effort has been made to present the mathematical theory of trunking in a

[1] G. T. Blood of the A. T. & T. Co. in 1898 found a close agreement between the terms of a binomial expansion and the results of observations on the distribution of busy calls. The first comprehensive paper was one written by M. C. Rorty in October 1903 and was quite widely circulated within the Bell System.

[2] An excellent bibliography is given by G. F. O'Dell in the P. O. E. E. J. for October 1920.

manner that can be understood by those who are not experts on the subject. It is hoped that this article will assist the reader in understanding both what has been and will be written on the subject. As Poisson [3] has said "a problem relative to games of chance and proposed to an austere Jansenist by a man of the world, was the origin of the calculus of probabilities," and today the reader will find that in the majority of text books the subject is introduced by the solution of games of chance and particularly of dice problems. This established custom will be followed by the present writer who, in the course of this article, will show how various fundamental trunking problems can be transformed into equivalent dice problems. This being done, solutions will be found to be at hand.

Three trunking problems, each one step more complicated than the preceding, will be dealt with. In order to facilitate the transformation to the three equivalent dice problems it is desirable that the basic assumptions made be as simple as possible. The asumptions made in all three problems are:

A—*During the period of time under consideration, the busy hour of the day, each subscriber's line makes one call which is as likely to fall at any one instant as at any other instant during the period.*

Conditions substantially approximating this assumption frequently occur in practice.

B—*If a call when initiated obtains a trunk immediately it retains possession of that trunk for exactly two minutes. In other words, a constant holding time of two minutes duration will be assumed.*

In practice, holding times, of course, vary from a few seconds to many minutes and it may at first sight seem that the assumption of a constant holding time might lead to results deviating too much from practice to be of value. On this point, the theory of probabilities itself sheds some interesting light. As will be pointed out in the following problems, the assumption of a constant holding time is the equivalent of a dice problem in which a single die, or several identical dice are considered. The telephone problem with variable holding times may be reduced to the consideration of many dice, each with a different number of faces. Suppose 600 throws are made with a die having 6 faces so that on the average $\frac{1}{6}$ of 600 or 100 aces would be expected. With Bernoulli's formula it is easy to find the probability that the number of aces which turn up shall lie between 75 and 125, that is to say, within 25 on each side of the average. Now suppose 200 throws are made with a die having 20 faces, 200 with a

[3] Poisson, Recherches Sur La Probabilite Des Jugements, 1837.

10 face die, 100 with a 5 face die and finally 100 with a 2 face die. These 600 throws would also give on the average 100 aces. Using Poisson's generalization of the Bernoulli formula [3] we can calculate the probability that these 600 throws with various kinds of dice shall give a number of aces lying between 75 and 125. This probability will be *greater* than in the case of the 600 throws with the die with a constant number of faces, *i.e.*, the chance that the result will come outside the range 75 to 125 is *less*.

The thought is at once suggested that for the same total volume of traffic and average holding time, fewer calls would be lost when the holding time is *not* constant.[4] The above theory was tested in practice a few years ago by the engineers of the American Telephone and Telegraph Company, who made pen register records of hundreds of thousands of actual calls as handled by groups of machine switching trunks at Newark, New Jersey. A pen register was made which operated as follows: Each trunk in the group was represented by a pen. These pens were mounted side by side and each was controlled by a magnet in such a manner that when the trunk was busy the pen made a mark on a wide strip of paper driven at constant speed under the pens. There was thus obtained a record showing when each call originated and when it was concluded. An artificial record was now made showing what would have happened if each call had lasted for the average holding time as determined from the original record. Some 100,000 calls were analyzed in this manner and it was found that with a group of trunks of a size to carry the calls of the original record with only a small loss, 30 per cent. more calls would have been lost if the traffic had been as shown by the *artificial* record. It should be borne in mind, however, that a 30 per cent. change in a probability of the order of one in one hundred, considering the values we are dealing with, is practically negligible.

C—*If a trunk is not obtained immediately the calling subscriber waits for two minutes and then withdraws his call. If while waiting a trunk becomes idle he takes it and converses for the interval of time remaining before his two minutes are up.*

This assumption, although artificial, simplifies materially the analysis of the problems. Just what happens in practice to every call

[4] This result is here reached by assuming that each subscriber originates one call per hour. The conclusions are the same, however, even when this is not true, provided the term "holding time" is understood to mean the aggregate of all the talking times of the subscriber in an hour.

It may also be mentioned in passing that for a fixed volume of traffic, the discrepancy decreases as the number of subscribers who originate that traffic increases; that is, it is less when the group is composed of a large number of relatively idle lines than when it is composed of a small number of very busy ones.

which fails to get a trunk immediately is unknown. It is obvious, however, that when the number of trunks is such that the liability of the call failing to get a trunk immediately is very small—for example: of the order of one in one hundred—the reaction of these calls on other calls must be negligible independently of whatever assumption [5] is made in place of C.

PROBLEM I

Referring to Fig. 1 consider a group of 269 subscribers' lines each equipped with a 20-point line switch. When a subscriber removes his receiver his line switch revolves and picks up the first idle trunk which

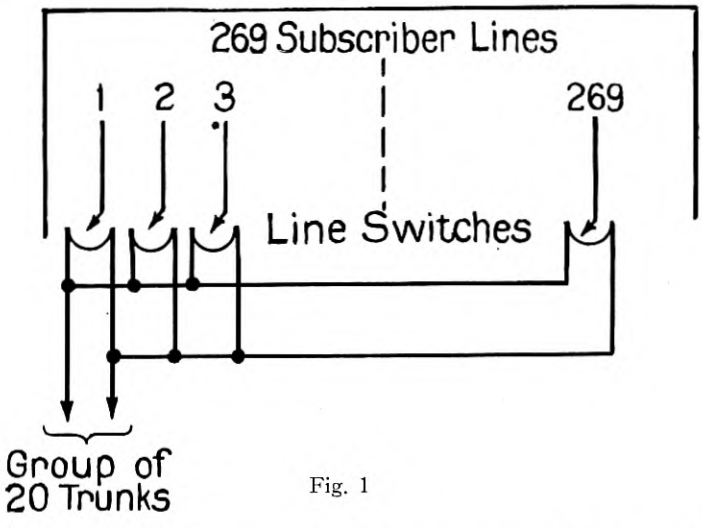


Fig. 1

it comes to. The 20 points of all switches are multipled together so that a single group of 20 trunks must handle the calls originating from these 269 lines.

What is the probability that when a particular subscriber X calls he fails to obtain a trunk immediately?

Referring to Fig. 2 let point P represent the unknown instant within the hour at which X calls. Consider the two minutes immediately preceding the instant P. Evidently, by assumption C, calls falling outside of this particular two-minute interval can not prevent X from obtaining a trunk.

[5] It is well known that the Erlang formula which is based on an assumption diametrically opposed to assumption C, namely that calls which find all trunks busy do not wait for a trunk to become idle, gives essentially the same results (for small probabilities, which are the only ones of interest in practice) as the Poisson furmula which assumes C.

If, however, at least 20 of the remaining 268 subscribers initiate their calls within the particular two minutes under consideration, there will be no trunk immediately available for X. This follows from assumptions B and C.

Consider some one of these 268 other subscribers, for example Y. The probability that Y calls in the two minutes under consideration is by assumption A, the ratio of 2 minutes to 60 minutes, or 1/30, which is exactly the same as the probability that he would throw an ace if he were to make a single throw with a 30-face die. Likewise the probability that still another subscriber calls in the two minutes under consideration is exactly the same as the probability that this

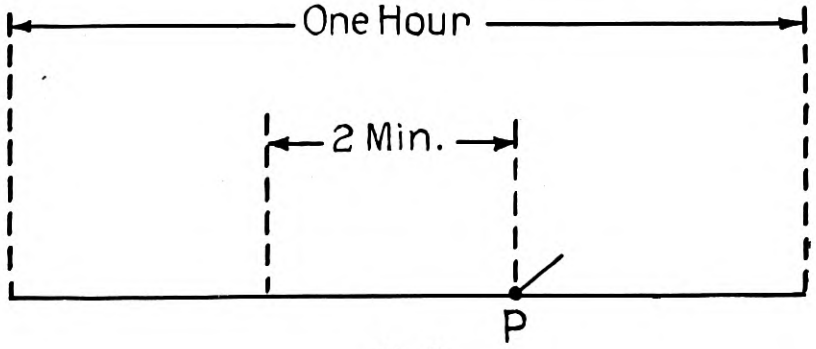


Fig. 2

other subscriber should throw the ace in a single throw with a 30-face die.

It is evident then, that the probability that X fails to get a trunk immediately is the same as the probability of throwing *at least* 20 aces if 268 throws are made with a 30-face die. To facilitate the determination of this probability and the solution of similar problems, probability tables of a type shown in Table I have been computed.[6] In the table, the average number of times an event may be expected is represented by a. The probability that the event occurs at least a greater number of times $c = a + d$ is represented by P. In the problem under consideration, the average number of aces expected is $8.96 = \frac{269}{30}$. Likewise in the present problem $c = 20$. Turning to the table, we find that corresponding to $c = 20$ and $a = 8.96$, the value of the probability P is .001. In the particular telephone problem under consideration this means that once in a thousand times

[6] Table I is to be found in the Appendix and its origin is there explained.

when X calls, at least 20 of the other subscribers will have called in the two minutes immediately preceding, and therefore X fails to get a trunk immediately. In other words, we may consider that on the average one in every thousand calls is lost.

In the problem just considered, a known number of subscribers' lines have had a known number of trunks assigned to them and we have inquired the probability that any subscriber would fail to find an idle trunk. It is frequently desirable to change the statement of

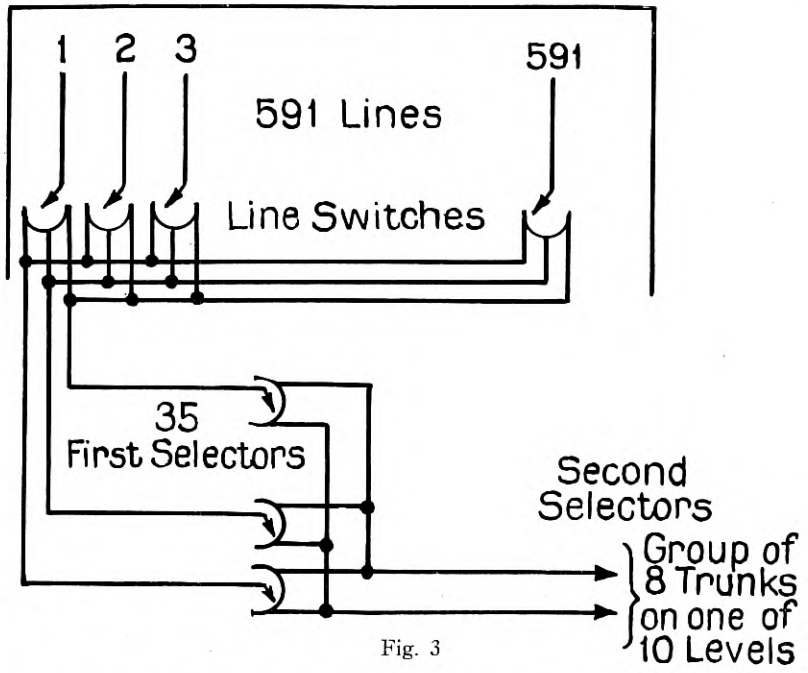


Fig. 3

the problem slightly. For instance: given a known number of subscribers' lines and having decided upon a desirable value of the probability P, we may inquire the number of trunks which must be assigned. It is evident that in this problem we would enter the table knowing the value 8.96 and .001, and would find corresponding to these, the number 20 as representing the size of the required group of trunks.

Problem II

Referring to Fig. 3 consider a group of 591 subscribers' lines each equipped with a 35 point line switch giving access to first selectors. We will suppose for illustration that each first selector has 10 levels

or choices. The reader unfamiliar with automatic systems may consider a 10 level selector as one from which calls may be sent in 10 different directions. Assume that each level is equipped with 8 trunks to second selectors. The 591 line switches are multipled together so that *one* group of 35 first selectors must handle the calls originating from these 591 lines. The 35 first selectors are multipled together so that *one* group of 8 second selectors must handle the calls originating from the 591 lines for a particular level. It is assumed that the 591 calls are distributed at random with reference to the 10 levels of the first selectors.

The probability that X should fail to obtain immediately a first selector can be determined as in the first problem, but now let us determine what is the probability that subscriber X (having obtained immediately a first selector) fails to obtain immediately one of the 8 trunks of a particular one of the 10 levels on the first selectors.

For subscriber Y to interfere with X it is necessary that Y originate his call in the two minutes preceding the instant at which X calls and also that Y call for the particular one of the 10 levels in which X is interested.

The probability of Y fulfilling the first condition is equal to the probability of throwing the ace with a 30 face die. The probability of Y fulfilling the second condition is equal to the probability of throwing the ace with a 10 face die.

The question may then be stated in the form of a dice problem as follows: 591 throws are made with a 30 face die giving C aces. C throws are made with a 10 face die giving D aces, and the question is the probability that D is not less than 8. Assuming no restriction [7] on the value of C this probability is the same as that of throwing at least 8 aces in 591 throws with a die having (30) (10) = 300 faces.

The average number of aces to be expected is (591/300) = 1.97 and with this average the tables tell us that once in a thousand times we may expect at least 8 aces.

[7] Since it is assumed that X obtained a first selector it follows that in the 2 minutes preceding the instant when X called the number of calls must have been less than the number of first selectors and we should, therefore, not count the throws giving values of C which are not less than the total number of first selectors. This restriction becomes of practical importance only where a large proportion of the calls from the first selectors go to one level. To take an extreme case, assume that all the calls went to one level, and that therefore each 10 first selectors would require 10 second selectors to handle the traffic. Placing no restriction on the value of C, since C exceeds the number of first selectors occasionally, we would get the result that 10 second selectors were not enough to handle all the calls from 10 first selectors, which is of course absurd. Where, however, the values of C exceeding the number of first selectors are assumed to be distributed over all 10 levels of the first selectors their effect on the number of second selectors is negligible.

Problem III

In practice, a modification of Problem II frequently arises. Assume an arrangement similar to that of Problem II except that the number of lines is multiplied by a factor of perhaps 3 or more, each line switch, however, still having access to all first selectors. The required number of first selectors will also be larger but not in exactly the same ratio because the margin of idle selectors need not be relatively as great in the large system as in the small. An enlarged group of trunks running from the first selectors to the second selectors will now be required, and it will be assumed that there are four times

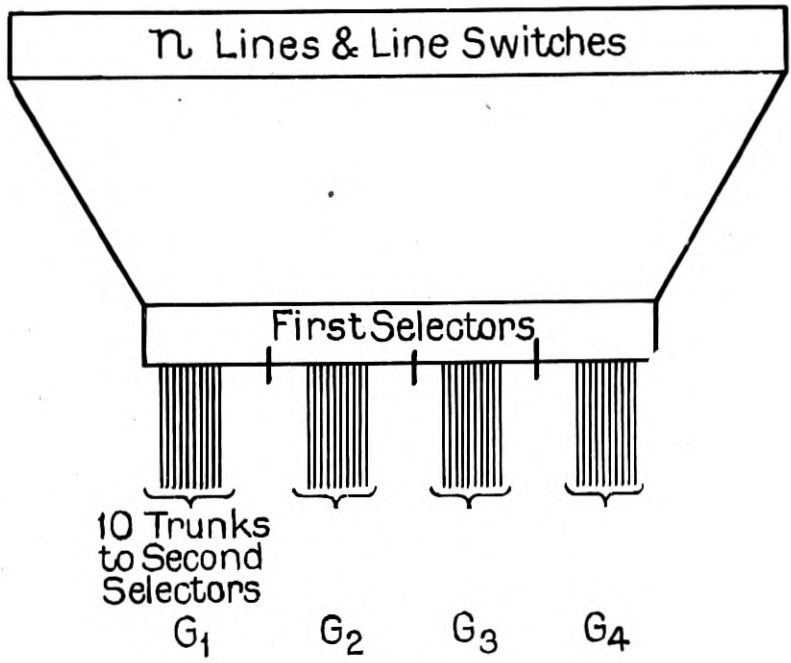


Fig. 4

as many trunks coming from each level of the first selectors as there are points of contact on each level. To meet this situation, the first selectors and their outgoing trunks are divided into four sub-groups as shown in Fig. 4. The corresponding sub-groups of second selectors are designated by G_1, G_2, G_3, G_4, the number of trunks to each sub-group being 10. The solution of this problem depends primarily on the manner in which the line switches distribute calls to the first selectors. Three cases will be considered.

Case 1

Referring to Fig. 4 consider a group of $n = 1486$ lines, and let the traffic be divided among the ten levels or directions available in such a way that on the average $\frac{1}{3}$ of the calls are made for a particular direction or level. Let us suppose the circuit connections between the line switches and first selectors to be such that the calls are distributed individually at random. By this is meant that the first selector seized by a calling line is as likely to be one having access to sub-group G_1 as to sub-group G_2, G_3 or G_4. Note carefully that this distribution is assumed whether or not the calling line wants the particular level under consideration. One way of securing this random distribution by sub-groups would be to allow the line switches first to choose by chance one of the four sub-groups of first selectors and then to choose an idle first selector in the sub-group.

Question—What is the probability that when subscriber X calls he fails to obtain immediately a trunk to a second selector? It is assumed that X obtained a first selector and that his call is for the level under consideration.

As before we are interested in the calls made during the two minutes preceding the instant at which X calls. Let the number of these calls be C. Of these C calls a certain number D want the level for which X has called. If at least 10 of these D calls were distributed by the line switches to first selectors having access to the same sub-group as the one to which the first selector seized by X has access, then there will be no idle trunk in the sub-group for X. Our telephone trunking problem evidently transforms to the following series of dice problems.

1st. 1486 throws are made with a 30 face die giving C aces.
2nd. C throws are made with a 3 face die giving D aces.
3rd. D throws are made with a 4 face die giving x aces, and the question is the probability that x is not less than 10.

By the theory of dice (assuming no restriction [6] on the value of C) the probability is the same as that of throwing at least 10 aces in 1486 throws with a die having $30 \times 3 \times 4 = 360$ faces. The average number of aces to be expected being $1486/360 = 4.13$ the probability tables give .01 as the answer.

Case 2

As in Case 1 assume that on the average $\frac{1}{3}$ of the calls are for the level under consideration, but take $n = 1725$ for the number of lines. Now suppose the circuit connections between line switches

and first selectors to be such that the calls are distributed uniformly to the first selectors, meaning that if at any instant C calls exist, $C/4$ of them are on first selectors having access to the 10 trunks of sub-group G_1, $C/4$ are on first selectors having access to the 10 trunks of sub-group G_2 and so on. With a constant holding time such as assumed this result could be secured by a device common to all line switches which would route the first call to the first sub-group, the second call to the second sub-group, etc.

X will, as before, be interested in the calls falling in the two minutes preceding him. By hypothesis $\frac{1}{4}$ of these will have been distributed to first selectors having access to the same sub-group of second selectors as the first selector seized by X. Finally, the probability is $\frac{1}{3}$ that one of these calls wants the level in which X is interested. The equivalent dice problem is therefore:

1st. 1725 throws are made with a 30 face die and the number of aces which turn up are noted. Let this number be C.
2nd. $C/4$ throws are made with a 3 face die.

What is the probability that this sequence of throws results in at least 10 aces? This probability is not that of getting at least 10 aces if 1725 throws are made with a die having $30 \times 3 = 90$ faces. We must write separately the formula for each of the two steps of the problem, then multiply them together and finally sum the product for all values of $C/4$ from 10 up. If this is done, again ignoring the restriction on the upper limit of C, the answer will come out 0.01. Note that whereas in Case 1 the average volume of traffic carried by a sub-group of 10 trunks was 4.13, in this case, with the same probability of failure, it is 1725 (1/30) (1/4) (1/3) = 4.79.

Case 3

In conclusion, a third and very interesting case will be mentioned. A distribution of calls *collectively at random* would be an appropriate name, and its nature may be described as follows:

Number each first selector and a corresponding card; shuffle the cards and deal out, for example, 37 of them. The distribution under consideration is such that when 37 calls exist the probability that they occupy a specified set of 37 selectors is equal to the probability that the cards dealt have the corresponding numbers. This distribution of calls would be measurably secured by arranging the line switch multiple so that the trunks to the first selectors appear so far as possible in a different order before every line switch. This case of distribution differs from that of Case 1. In Case 1, if the first call

falls on a first selector having access to sub-group G_1, for example, the second call still has the same chance of falling on a first selector having access to sub-group G_1 as on one having access to any one of the other three sub-groups. In Case 3, however, the busy first selectors tend to be distributed uniformly between the 4 sub-groups, so that if any sub-group should have a preponderance of busy first selectors the probability of its receiving another call is less than the probability that one of the other sub-groups, with more idle first selectors, should receive it. The full discussion of this case is reserved for the future.

APPENDIX

Introduction to the Mathematical Theory of Probabilities

If it is known that one of two events must occur in any trial or instance, and that the first can occur in u ways and the second in v ways, all of which are equally likely to happen, then the probability that the first will happen is mathematically expressed by the fraction

$$\frac{u}{u+v},$$

while the probability that the second will happen is

$$\frac{v}{u+v}.$$

Denote these probabilities by p and q respectively; then we have:

$$p = \frac{u}{u+v}, \qquad q = \frac{v}{u+v}, \qquad p+q = 1,$$

the last equation following from the first two, and being the mathematical expression for the certainty that one of the two events must happen.

If the probabilities of two independent events are p_1 and p_2 respectively, the probability of their concurrence in any single instance is p_1p_2, and in general if $p_1, p_2, p_3, \ldots\ldots p_n$ denote the probabilities of several independent events, and P the probability of their concurrence, then

$$P = p_1p_2p_3 \ldots\ldots\ldots p_n.$$

Consider, now, what may happen in n trials of an event, for which the probability is p and against which the probability is q. The

probability that the event will happen every time is $pppp \ldots\ldots p$, where the factor p appears n times; that is the probability is p^n. The probability that the event will occur $(n-1)$ times in succession and then fail is $p^{n-1}\,q$.

But if the order of occurrence is disregarded, this last combination may arrive in n different ways; so that the probability that the event will occur $(n-1)$ times and fail once is $np^{n-1}\,q$. Similarly, the probability that the event will happen $(n-2)$ times and fail twice is $p^{n-2}\,q^2$ multiplied by $n(n-1)/2$, etc. That is, the probabilities of the several possible occurrences are given by the corresponding terms of the binominal expansion of $(p+q)^n$. Let

$$P = p^n + \binom{n}{1}p^{n-1}q + \binom{n}{2}p^{n-2}q^2 + \ldots\ldots + \binom{n}{c+1}p^{c+1}q^{n-c-1} + \binom{n}{c}p^c q^{n-c}, \qquad (1)$$

where $\binom{n}{x}$ means $n\,(n-1)\,(n-2)\ldots\ldots(n-x+1)/(1)\,(2)\ldots\ldots(x)$.

Then P = probability that the event happens exactly n times, plus the probability that it happens exactly $(n-1)$ times . . . plus the probability that it happen exactly c times: in other words, the probability that the event happens at least c times in n trials.

If the series for P contains few terms it may be computed easily. In general, however, it is impracticable to compute P by means of the above binomial expansion. Other forms for the value of P must, therefore, be developed.

One of the most convenient approximations for P when p is small has been developed by Poisson. It is known as Poisson's Exponential Binomial Limit and gives the value of P by the following expansion

$$P = e^{-a}a^c/(c)! + e^{-a}a^{c+1}/(c+1)! + e^{-a}a^{c+2}/(c+2)! \ldots\ldots\ldots\ldots ad\ inf. \qquad (2)$$

where e = base of natural logarithms = 2.718 . . . . . , $a = (np)$ and $(c)! = c\,(c-1)\,(c-2)\,(c-3)\ldots\ldots(3)\,(2)\,(1)$.

The following Table gives corresponding values of P, a, c satisfying equation (2).

TABLE I.

Averages (a) Corresponding to Deviation (d) plus Average (a) to be Expected with Different Probabilities

Deviation Plus Average, $c = a + d$	Probabilities						Deviation Plus Average, $c = a + d$
	.001	.002	.004	.006	.008	.010	
	Average = a						
1	.001	.002	.004	.006	.008	.010	1
2	.045	.065	.092	.114	.133	.149	2
3	.191	.243	.312	.361	.402	.436	3
4	.429	.518	.630	.709	.771	.823	4
5	.739	.867	1.02	1.13	1.21	1.28	5
6	1.11	1.27	1.47	1.60	1.70	1.79	6
7	1.52	1.72	1.95	2.11	2.23	2.33	7
8	1.97	2.20	2.47	2.65	2.79	2.91	8
9	2.45	2.72	3.02	3.22	3.38	3.51	9
10	2.96	3.26	3.60	3.82	3.99	4.13	10
11	3.49	3.82	4.19	4.43	4.62	4.77	11
12	4.04	4.40	4.80	5.06	5.26	5.43	12
13	4.61	5.00	5.43	5.71	5.92	6.10	13
14	5.20	5.61	6.07	6.37	6.60	6.78	14
15	5.79	6.23	6.72	7.04	7.28	7.48	15
16	6.41	6.87	7.39	7.72	7.97	8.18	16
17	7.03	7.52	8.06	8.41	8.68	8.90	17
18	7.66	8.17	8.75	9.11	9.39	9.62	18
19	8.31	8.84	9.44	9.82	10.11	10.35	19
20	8.96	9.52	10.14	10.54	10.84	11.08	20
21	9.62	10.20	10.84	11.26	11.57	11.83	21
22	10.29	10.89	11.56	11.99	12.31	12.57	22
23	10.97	11.59	12.28	12.73	13.06	13.33	23
24	11.65	12.29	13.01	13.47	13.81	14.09	24
25	12.34	13.00	13.74	14.21	14.57	14.85	25

The Relation Between Rents and Incomes, and the Distribution of Rental Values

By W. C. HELMLE

SYNOPSIS: Many parts of telephone plant, such as central office buildings and equipment, conduits, underground and aerial cable at the time of installation must have the capacity to handle not only the immediate demand for telephone service, but also to take care of growth for a number of years to come. In order to engineer such items of telephone plant economically it is necessary to know in advance as accurately as possible what the demand for telephone service will be five, ten, or twenty years in the future. Forecasts of the future market are very necessary for plant engineering, operating plans, rate treatment, and other purposes, in multi-central office cities. In such cities detailed estimates are made of the market some twenty years ahead and of its telephone development under stated rate conditions. Such estimates are called *commercial surveys*, and they involve a study of the various factors which, in the course of events, will be likely to control the industrial, commercial and residential development of the city concerned.

In the course of such a survey, a rental classification of all families is obtained and at the same time a record is made of existing telephone service in each rental class. The rent data of this article have been gathered in representative large cities throughout the country and the results as here set forth are being used together with many other kinds of data to guide the engineering of future additions to the plant of the Bell Telephone System.

In general the income of a family is an index of the market it creates for various commodities including telephone service. Rental values may also be considered as such an index and the present study seeks to correlate rents with incomes. Rents can be readily recorded and classified, whereas it is not feasible to determine the money incomes of large numbers of families. While it may be ideally possible, by a study of rent data, to compare the inherent markets for telephone service and also the strength of the telephone habit in various cities, there are many practical limitations to such a procedure. Comparison of the residence market for telephone service in different cities, as determined by rent values, is made difficult by the fact that the variation between cities in rentals paid for substantially similar dwellings is considerably greater than the variation in prices for food or clothing. Further, there is considerable variation in rent levels even in different sections of any one city. Attempts to compare rent distributions by application of the usual statistical measures of dispersion and skewness have proved unsatisfactory. However, a method of charting has been found by which rent distributions may be readily compared with one another and an index of spread or dispersion determined. It has been found that cumulative curves of rent distribution may be plotted on logarithmic probability paper to yield straight lines for a large number of cities. These are called logarithmic skew distributions. Although it has not been found possible to assign any special significance to the particular value of the index of rent dispersion in any city, this index appears to remain practically constant for that city regardless of changes in the level of prices. In the appendix the mathematical features of the logarithmic skew curve are discussed.—*Editor.*

IT is a well recognized fact that the better class families, *i.e.*, those with higher incomes, are a better market for telephone service than the poorer families. For purposes of market analysis in commercial surveys it is not feasible to determine the money incomes received by families but the rental values of dwellings, which, as

will be shown, are a measure of the incomes of their occupants, are comparatively readily collected and classified. Rent data obtained in the course of a commercial survey show the "character" of a city and are used as a basis for estimates of the future residence telephone market.

In view of the importance of these rent data, it seems desirable to study them in some detail to find out just what their limitations are.

There may be set down in advance certain things which it is desirable to know, as, for instance, the relationship between money incomes and house rents and methods and limitations of comparison of different cities on the basis of rental values. On the first point, as applied to any particular city, a knowledge of the relation between incomes and rents is desirable in a general way, although there is no necessity to translate the telephone market expressed in terms of rent types into a scale of incomes. On the second point, the comparison of different cities, it should be ideally possible by a study of rent data to compare the inherent economic markets for telephone service, and also to measure differences in the strength of the telephone habit, but in practice only rough approximations may be made.

Certain limitations to work of this kind are fairly evident. The most obvious difficulty is the fact that rent levels have changed along with the general price level. Rent levels in various cities differ according to the varying degrees of housing congestion and the varying social standards of the population. Furthermore, the variation in rent levels extends to different sections of any one city. The mere fact that a given family paid say $30 rent is not an indication of that family's economic condition or its value as a telephone prospect, unless there is also known the city and the part of the city in which that family lived, and the time when the given rent was paid. Therefore rent data from different cities and of various dates are not directly comparable at their face value. To adjust the money values of house rents for an accurate comparison of the telephone markets in different cities would require a knowledge of the relative proportions of income spent for rent in the different cities, of the relative levels of incomes and rents at the time of the surveys as compared with their levels in some base year, and perhaps of other factors equally difficult to estimate.

Various rent tabulations can not be compared one with another without knowing something of the way in which rents are distributed about their average. The nature of the distribution is determined by the house count data, but from those data in their usual form it

is not easy, when making comparisons, to make proper allowances for differences in rent levels and in the schedule of rent classes. In the following pages there is discussed a method of charting by which rent distributions may be readily compared, and their spread or dispersion determined.

The Relation Between Rents and Incomes

Rents as a Market Index. The relation between rents and incomes is concerned with the use of rental values both as an index of telephone market in a given city, and in comparing the markets in different cities. In what follows it is not always possible to separate these two views, but the distinction should be borne in mind by the reader.

The goal of an analysis of residence telephone market is to determine the future sales possibilities. In theory either incomes or rents may be considered as an index of the telephone market. The market index adopted in commercial surveys is the rental of dwellings. This may be considered either as a direct measure of the ability and desire of families to subscribe to telephone service, or as an indirect index, if incomes are considered the real measure of the market. If the first viewpoint is accepted, it may be logically concluded, although not proved, that rents are a better index of telephone market than are incomes. Incomes, as measured in money, are the nearest approach which may be made to a measure of the position of families on an imaginary scale of economic welfare. An attempt to translate rent data to an income basis, as a working method in commercial surveys would introduce errors with no compensating advantage, but a translation of this kind is more or less unconsciously made in making comparisons.

Sources of Information. Much of the literature on the question of house rents versus incomes is generalization based on limited or antiquated data. Such careless statements as "rent approximates about one-third of the average worker's income," may be found in the literature of the subject. Adam Smith, the father of Political Economy, "made the assertion, surprising to us in these days, that the proportion of income spent in house rent is highest among the rich." Frederick Engels concluded in 1857 that rent was 12 per cent and heat and light 5 per cent of the workingman's expenditure, regardless of the amount of his income.

Investigation of budgets in recent years has been confined almost entirely to the field of the wage earning class. The first really comprehensive study was made by the United States Bureau of Labor

Statistics in several states in 1901 and 1902, and is detailed in the 1903 annual report of that organization. It consisted chiefly of a study of 11,156 so-called "normal" families, each including a husband at work, a wife, not more than 5 children all under 14 years and no lodgers or servants. The average income of these families was $651. Original work on a smaller scale has been done by R. C. Chapin (1908) and by the Philadelphia Bureau of Municipal Research (1918). The Bureau of Labor Statistics collected a large amount of data in 1918 and 1919[1] concerning the incomes and expenses of 12,837 families in 92 towns having an average income of $1491. This investigation included families of wage earners and low salaried men, but none of the slum or recent immigrant classes. Families of the lowest type are automatically excluded from such studies as this by their inability to supply the desired information from accounts or from intelligent estimates.

Distribution of Family Expenses. Representative distributions of family expenditures are given in Table I. The National Industrial Conference Board has adopted for use in computing their cost of living index and representative budgets a list of standard weights made by combining the results of a number of studies made from 1901 to 1917. Most importance was assigned to the first Bureau of Labor Statistics study, the results of which it closely resembles. The standard weights used by the Bureau of Labor Statisics are the result of surveys made in 22 cities from July 31 to November 30, 1918, covering families whose average income was $1,434.

TABLE I.

Distribution of Total Family Expenditure

	Averages Found in		Standard Weights Used by	
	Bur. Lab. Stat. First Study 1901–1902	Bur. Lab. Stat. Second Study 1918–1919	Bur. Lab. Stat. Weights in Cost of Living Index	Nat'l Ind. Conf. Board Budgets and Index of Cost of Living
Food	43.13%	38.5%	38.2%	43.13%
Rent	18.12	13.3	13.4	17.65
Clothing	12.95	16.5	16.6	13.21
Fuel and Light	5.69	5.3	5.3	5.63
Sundries	20.11	26.4	26.4	20.38

From Table I it may be inferred that the per cent spent for rent is reduced in a period of inflated prices, at least during the first part

[1] *Monthly Labor Review*, May–December, incl., 1919.

of that period. This is reasonable, since rents respond less rapidly than most other prices to fluctuations in the general price level. An extreme example of this type is found in Germany where rents, which are to some extent under government regulation, "at the present time absorb not more than 3½ per cent of total expenditure as against 20 per cent before the war." [2]

The percentage distribution of total expenses depends on the size of the family, the income received and the city lived in. Of course, it must be understood that any particular family may differ widely from general averages. Other things being equal, large families spend more for food and clothing and less for rent and sundries than do small families. Large families of the lower middle class accommodate themselves to whatever housing accommodations they can afford after the more inflexible demands for other things have been provided for. Less than one room per person is considered over-crowding and the recent Bureau of Labor Statistics investigation found this condition to exist rarely, except in families having more than three children. Families with one to three children were found to have 1.0 to 1.3 rooms per person in almost all cities.

Amount of Income vs. Per Cent Spent for Rent. The extent to which the distribution of expenses is modified by the amount of income received is known only within the very limited range for which data are available. The best recent figures are those of the 1918-1919 study of the Bureau of Labor Statistics. These are given here for 12,096 white families in 92 cities and towns:

TABLE II.

Income	Per Cent of Total Expenses Spent for					
	Food	Clothing	Rent	Fuel and Light	Furniture and Furnishings	Misc.
Under $900	44.1	13.2	14.5	6.8	3.6	17.8
$900–$1200	42.4	14.5	13.9	6.0	4.4	18.7
$1200–$1500	39.6	15.9	13.8	5.6	4.8	20.2
$1500–$1800	37.2	16.7	13.5	5.2	5.5	21.8
$1800–$2100	35.7	17.5	13.2	5.0	5.5	23.0
$2100–$2500	34.6	18.7	12.1	4.5	5.7	24.3
$2500–up	34.9	20.4	10.6	4.1	5.4	24.7

When the original data are examined in detail, it appears that in almost every city as incomes increase the per cent spent for rent

[2] M. Elsas, *Economic Journal*, September, 1921, p. 332.

and food decreases and the per cent for clothing increases. The decrease in the per cent for food as incomes increase is slight and the increase in the per cent for clothing is especially marked in the higher incomes within the range covered. Thus, it appears that among families of moderate incomes as incomes rise the increase is spent by preference for clothing rather than for food or rent. The relative decrease in expenditure for rent as incomes increase is significant in rental analysis. This means that while a 10 per cent difference in rents among the lower rents in a city indicates an average difference in income of about 10 per cent, a similar difference among the higher rents indicates a difference in income of much more than 10 per cent.

Rent Levels in Various Cities. As nearly as may be determined from the Bureau of Labor Statistics data, there is no regular tendency for Eastern, Western or Southern cities to differ from the average of all cities, either in the amount of wage-earners' incomes or in amounts spent for food or clothing. In Southern cities somewhat less is paid for rent than in other cities. This refers only to white families. Negro families have smaller average incomes than white families and at any given income they spend less for rent, more for food and, to a less degree, more for clothing than white families. The size of a city, so far as may be told from these data, does not determine either total incomes or expense for food, rent or clothing.

In different cities the difference in rent levels, that is, variation in rentals paid for substantially similar dwellings, is considerably greater than the difference in levels of prices for food or clothing. The variation in price levels is about twice as great for rents as for food or clothing, reckoned as percentages of the amounts spent for each class. The expenditure for food by the lower middle class families included in this investigation is more nearly the same in different cities than is the expenditure for rent or for clothing. Food expense is the only one of these classes in which all cities are as closely grouped as in total expenses, considering deviations from the averages on a percentage basis. The amounts spent for rent show relatively wide variation between cities. It appears that if a workman moves from one city to another to secure increased wages a large proportion of the increase in income goes for increased rent. This is to be expected since land rents and, to a less extent, construction costs are peculiar to each individual city, much more than food or clothing costs.

A comparison of rent data from the 1918-1919 investigation of the Bureau of Labor Statistics and data from Commercial Surveys leads to the conclusion that differences in average rents in various cities are due at least as much to differences in the level of prices for rents

as to differences in the grade of the population. Wage-earners and low salaried people of the types studied by the Bureau of Labor Statistics occupy about the same position in the community in a large number of cities. As a rule they pay about 80 to 90 per cent of the median[3] rent in any city. Exception must be made in the case of cities having an unusually large proportion of negro or very low grade white population. It is interesting in this connection to compare wage rates for different classes of labor in various cities. The variation between cities in wage rates for common labor is proportionately much greater than the variation in wages for work requiring some skill, such as bricklaying and structural iron work.

As examples of the impossibility of accurately rating the grade of a city's population by its median rent alone, we may take four cities where surveys were made in 1921. Spokane and Houston had practically identical median rents of $23.00 and $23.40 respectively, but Houston is not as good a telephone market as Spokane. In Cleveland and Minneapolis the median rents were found to be $35.50 and $31.00 respectively, but this is no measure of the grades of the two cities.

Rent Data from Various Sources, Including England. Some additional rent data is presented here without extended comment. The two following tables show the proportion which rent bears to total expense in different communities.

TABLE III.

Pre-War Expenditures for Rent with a "Normal" Standard of Living

(Senate Report on "Woman and Child Wage Earners")

Manhattan	20.7%
Fall River	17.6%
Georgia and North Carolina	6.3%
Homestead, Pa	15.5%

TABLE IV.

Allowances for Rent in Post-War Standard Workingman's Budgets

	Date	Per Cent for Rent	Rent
No. Hudson Co., N. J.	Jan., 1920	13.5–14.1	$18.00–$19.00
Cincinnati	May, 1920	15.6	22.00
Lawrence	Nov., 1919	13.1–14.1	15.00– 19.50
Fall River	Oct., 1919	9.2–11.6	9.75– 15.20
Philadelphia	Nov., 1919	16.6	.....
United Mine Workers	Dec., 1919	10.2	.....
Washington, D. C	Aug., 1919	13.3	.....

[3] See Appendix.

The first four of these "standard" budgets are by the National Industrial Conference Board, and the others in order by the Bureau of Municipal Research (Philadelphia), Professor W. F. Ogburn, and the U. S. Bureau of Labor Statistics.

It has already been mentioned that the per cent spent for rent shows a tendency to decrease with increasing incomes. This trend is confined by data from other sources, as follows:

TABLE V

Per Cent of Income Spent for Rent at Different Income Levels

Income	Philadelphia Bur. Mun. Res. 1918 260 Families	Chapin's N. Y. Study 1907–1908 391 Families	U. S. Bur. Lab. Stat. 1901–1902 11,156 Families
$400– $500	.....	26.8%	18.6%
500– 600	.....	25.9	18.4
600– 700	20.5%	23.6	18.5
700– 800	17.6	21.9	18.3
800– 900	18.1	20.7	17.1
900– 1000	15.8	19.0	17.6
1000– 1100	16.4	18.1	17.5
1100– 1200	14.3	16.2	.....
1200– 1300	14.7	19.8	.....
1500– 1600	12.3	16.3	.....
1900 up	10.2	.....	.....

Although no data are available for families above the lower middle class, the relationship may be extended by conjecture into the higher income levels. That this is reasonable is brought out in subsequent pages in comparing distribution curves of rental values and incomes.

Some interesting conclusions from English experience are given by Sir J. C. Stamp.[4] The rent corresponding to an income of £160 averages at least £5 greater in London than outside that city. Among the lower incomes, say up to £1000, the variation or dispersion of the percentages paid for rent becomes less as the amount of the income rises. Owner-occupants live in larger houses than tenants with the same total income. It was not found, as is generally supposed, that professional men pay relatively more for rent than business men. The following table is from the work above mentioned;

[4] J. C. Stamp, "British Incomes and Property," 1916.

TABLE VI

Income	*Relative Amount Paid for Rent*
£200–£250	1.0
300– 400	.8
500– 750	.7
1000–1500	.5

Income	*Rent*	*Per Cent Rent*
£160	£28	17.5%
400–£500	40–£50	10
4000	200	5

The second of these two tables represents average conditions for Great Britain.

Rent Distributions

In making comparisons of survey rent data it is desirable to distinguish differences in price levels as they affect rents, differences in economic grade of the population, and differences in the distribution of families about their average grade. Failure to take account of these three factors will result in misleading impressions, which may be illustrated by summaries from successive surveys in Atlanta. The following table shows composites of private residences, flats and apartments:

TABLE VII

Rent Classes	No. of Families		Per Cent Increase
	1913	1920	
$75 up	793	3199	303
55–$75	1129	3582	217
40– 55	2045	4196	105
25– 40	4802	9713	102
20– 25	4197	4725	13
15– 20	5688	4757	—16
10– 15	6881	7223	5
Under $10	21064	15023	—29
Total	46599	52418	12.5

It might be inferred from this set-up that the condition of the poorer families had been very much improved or that the average family had attained a higher condition of well-being. It will be shown later that there was no material change in the distribution of rental values about an average rent when rents are considered as percentages of that average, and it is probable that the principal,

if not the sole cause of the changes shown in the table above, is the general rise in the level of prices.

Methods of Study—Graphic Representation. The most convenient and practical method of studying rent distributions is by the use of graphs and charts. The distribution of values of rents or of other variables may be charted in either a detail or a cumulative form. A detail curve shows at any value of the independent variable the frequency of occurrence of items of that value. A cumulative curve shows at any value of the variable the number (or better, the per cent) of all cases which have values below (or above) that value. Cumulative curves are better than detail curves for presenting rent data since the number of classes into which the data are divided is small and the class widths are non-uniform, resulting in uncertain

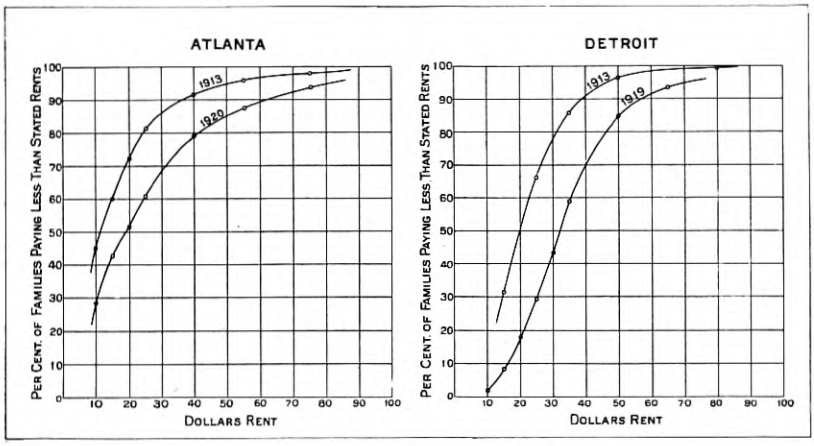


Fig. 1

curves of the detail type.[5] Attempts to compare rent distributions by application of the usual statistical measures of dispersion and skewness have proved unsatisfactory.

Typical rent distributions plotted in the cumulative manner on ordinary coördinate paper are shown in Fig. 1. Diagrams of this type may be used to determine the rent paid by families of corresponding position in the rent scale at the dates of successive surveys, but they do not give a very clear picture of changes in the distribution of rents and from them it is not readily apparent whether rents are closely concentrated or widely distributed in any given case.

[5] Detail curves for rent and income distributions are most easily drawn on paper with a logarithmic scale both ways.

Logarithmic Probability Charts. Cumulative curves for rent distributions may be plotted on logarithmic probability paper[6] in which case the resulting graph is a straight line for a large number of cities. Such a graph will be said to represent a *logarithmic skew distribution.* In the appendix there is given a discussion of frequency curves, with special reference to curves of this type. The essential point in reading charts on logarithmic probability paper is that the slope of the line determines both the spread or dispersion of the data and the skewness or lack of symmetry of distribution. Since the horizontal scale is logarithmic it follows that the dispersion is represented on a percentage and not a linear basis. A steep slope indicates a close concentration of the data, a less steep slope indicates a wider distribution, and parallel lines indicate distributions which are identical on a percentage basis. As explained in the appendix, the most convenient index or coefficient for expressing the spread or dispersion of a distribution is the ratio of the upper quartile[7] to the median rent. If the curves for a given city are closely parallel for successive surveys it follows that there has been no material change in the character of the distribution. In other words, rents have increased approximately proportionately at all points of the scale.

Examples of charts of this kind (Figs. 2-4) are shown for twelve cities for which successive surveys are available. Curves for successive surveys are nearly parallel in eight of the twelve cities. For Cleveland, Dallas and Houston there are distinct differences in the curves for the two dates, indicating changes in the distribution of rents, which changes may be measured since horizontal distances between points on the curves for two dates represent the percentage increases in rents.

When a rent distribution is plotted on logarithmic probability paper the points do not always lie on a straight line, but a straight line of best fit may be chosen by eye, giving greatest weight to points near the middle of the scale of ordinates. Of the rent distributions for large cities which have been plotted on this paper, nearly one-third are very closely represented by straight lines, an equal number are slightly concave upward, and the remainder are more or less concave downward. Most of the deviations from straight lines are slight. The examples submitted herewith (cities in which successive surveys have been made) are rather poorer than the average in this

[6] See an article by G. C. Whipple in the *Journal of the Franklin Institute* for July and August, 1916, for a description of this paper and some examples of its use in the field of sanitation.

[7] See Appendix.

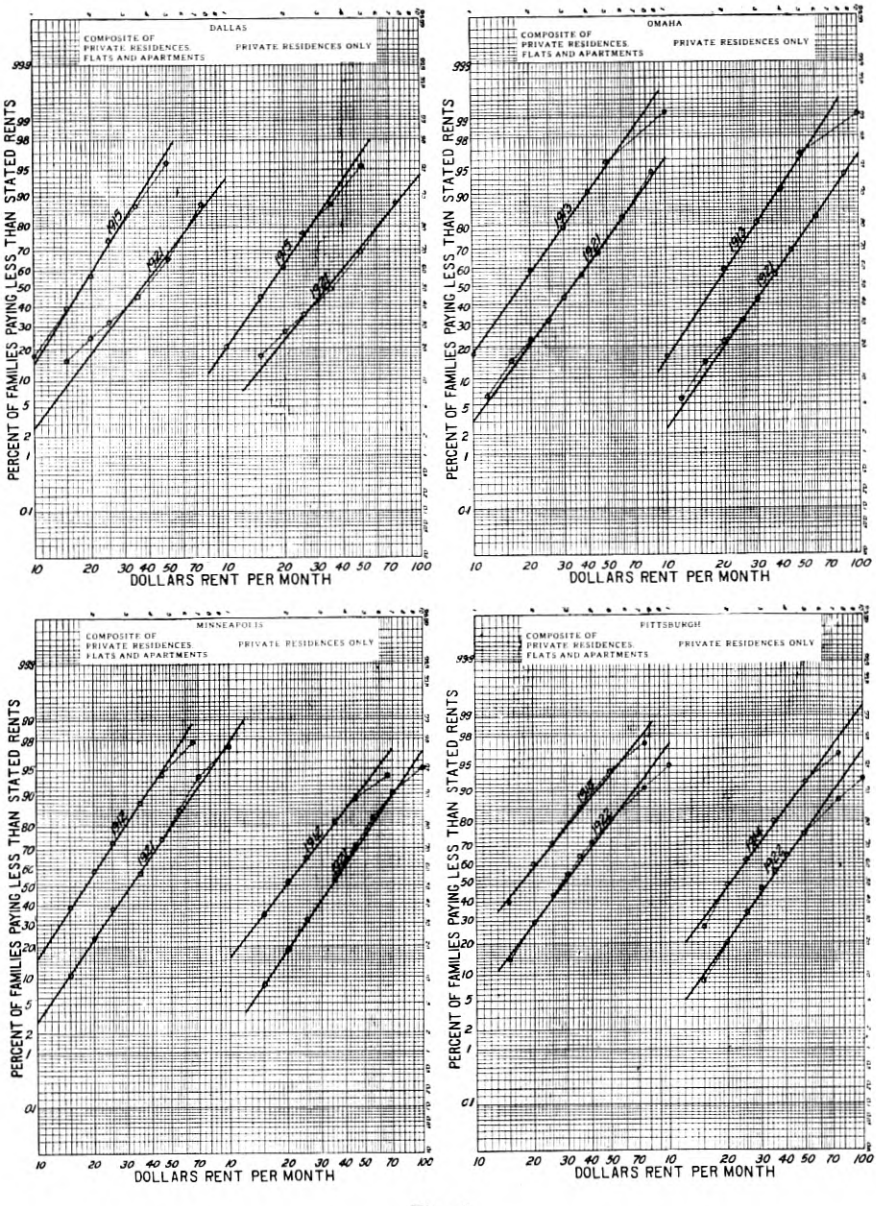

DALLAS
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
OMAHA
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
MINNEAPOLIS
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
PITTSBURGH
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
PERCENT OF FAMILIES PAYING LESS THAN STATED RENTS
DOLLARS RENT PER MONTH

Fig. 2

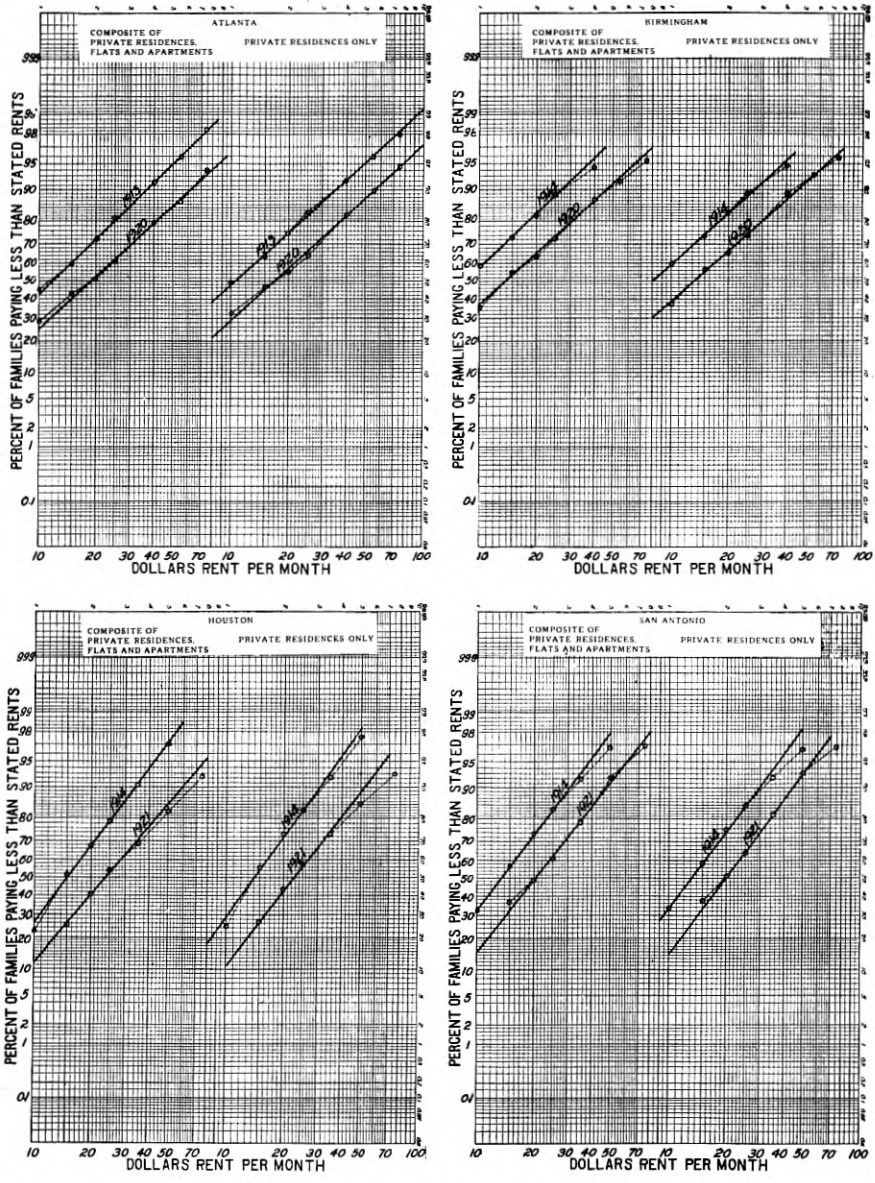

ATLANTA
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1920
1930
1929
1920
PERCENT OF FAMILIES PAYING LESS THAN STATED RENTS
DOLLARS RENT PER MONTH
BIRMINGHAM
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1930
1920
1929
1920
PERCENT OF FAMILIES PAYING LESS THAN STATED RENTS
DOLLARS RENT PER MONTH
HOUSTON
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1930
1920
1929
1920
PERCENT OF FAMILIES PAYING LESS THAN STATED RENTS
DOLLARS RENT PER MONTH
SAN ANTONIO
COMPOSITE OF PRIVATE RESIDENCES, FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1930
1920
1929
1920
PERCENT OF FAMILIES PAYING LESS THAN STATED RENTS
DOLLARS RENT PER MONTH

Fig. 3

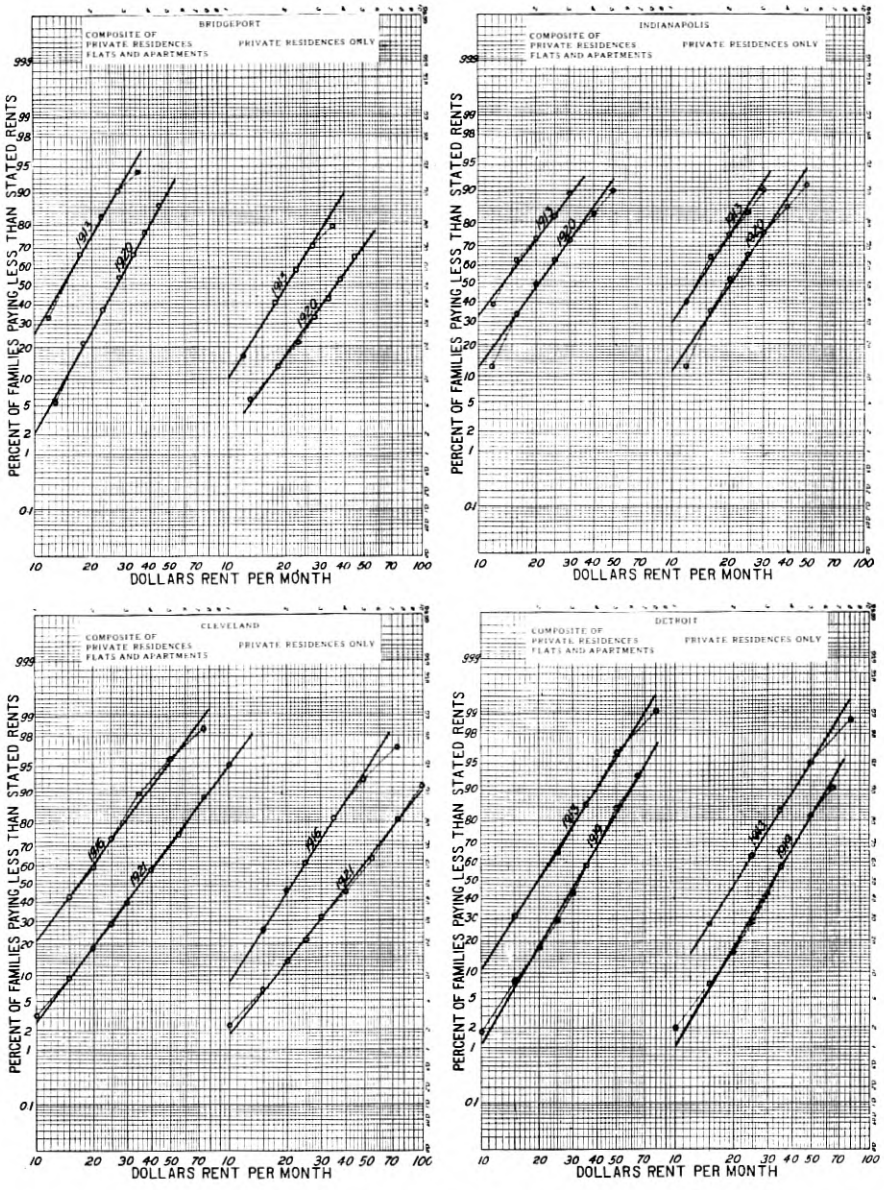

BRIDGEPORT
COMPOSITE OF PRIVATE RESIDENCES FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1913
1930
1913
1930
INDIANAPOLIS
COMPOSITE OF PRIVATE RESIDENCES FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1913
1930
1913
1930
CLEVELAND
COMPOSITE OF PRIVATE RESIDENCES FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1916
1923
1916
1941
DETROIT
COMPOSITE OF PRIVATE RESIDENCES FLATS AND APARTMENTS
PRIVATE RESIDENCES ONLY
1914
1929
1914
1939
PERCENT OF FAMILIES PAYING LESS THAN STATED RENTS
DOLLARS RENT PER MONTH

Fig. 4

respect. Concavity upwards represents a distribution which is less skew than the theoretical logarithmic skew curve, and concavity downwards a distribution of greater skewness. No great importance can be assigned to small differences of this sort, as they are not permanent between successive surveys, whereas the general type of the distribution is quite constant for a given city.

The justification for assuming that rents follow the logarithmic skew curve is made stronger by certain data from Volume 19 of the report of U. S. Immigration Commission made in 1912. This commission collected a large mass of data concerning the living conditions of families of the immigrant type. The data are classified by nationality of head of family, by income, etc. Distributions of amounts paid for house rent per apartment, per room and per person for certain nationalities are shown in Fig. 5. The data shown were chosen from those classes which were made up of the largest numbers, and the deviations from straight lines shown by data for other groups are in both directions, so the straight line relation may be considered fairly representative. It may be noted that the rents per month per person show a greater dispersion than the rents per room or per apartment. These latter moreover show as small a spread as do rents for any of the cities studied as a whole.

Fig. 5 also shows the distribution of British house rents at intervals during the period 1890-1913. There has been a gradual but steady decrease in the dispersion of rents during the period covered. Unless the relation between rents and incomes has radically changed, this means that the inequality of distribution of wealth has been decreased, and that the condition of the poor has been improved as compared to that of the rich. Data for 1830 indicate that the inequality of distribution was distinctly greater at that date than in 1890. Changes in the relative condition of the rich and poor may be readily demonstrated by charts of this kind, but of course conclusions regarding absolute degrees of well-being must be reached by other means.

Distribution of Rents Compared with that of Incomes. Significant conclusions regarding the relation between rents and incomes may be drawn from a comparison of their respective distributions. Fig. 6 shows a detail curve for income distribution in the United States based on preliminary data of the National Bureau of Economic Research. These data are subject to revision but are the best available and are sufficiently accurate for comparative purposes. The usual way to chart income distribution assumes conformity with Pareto's law which says that the frequency curve of incomes may be plotted as a straight line on double logarithmic paper, either on a

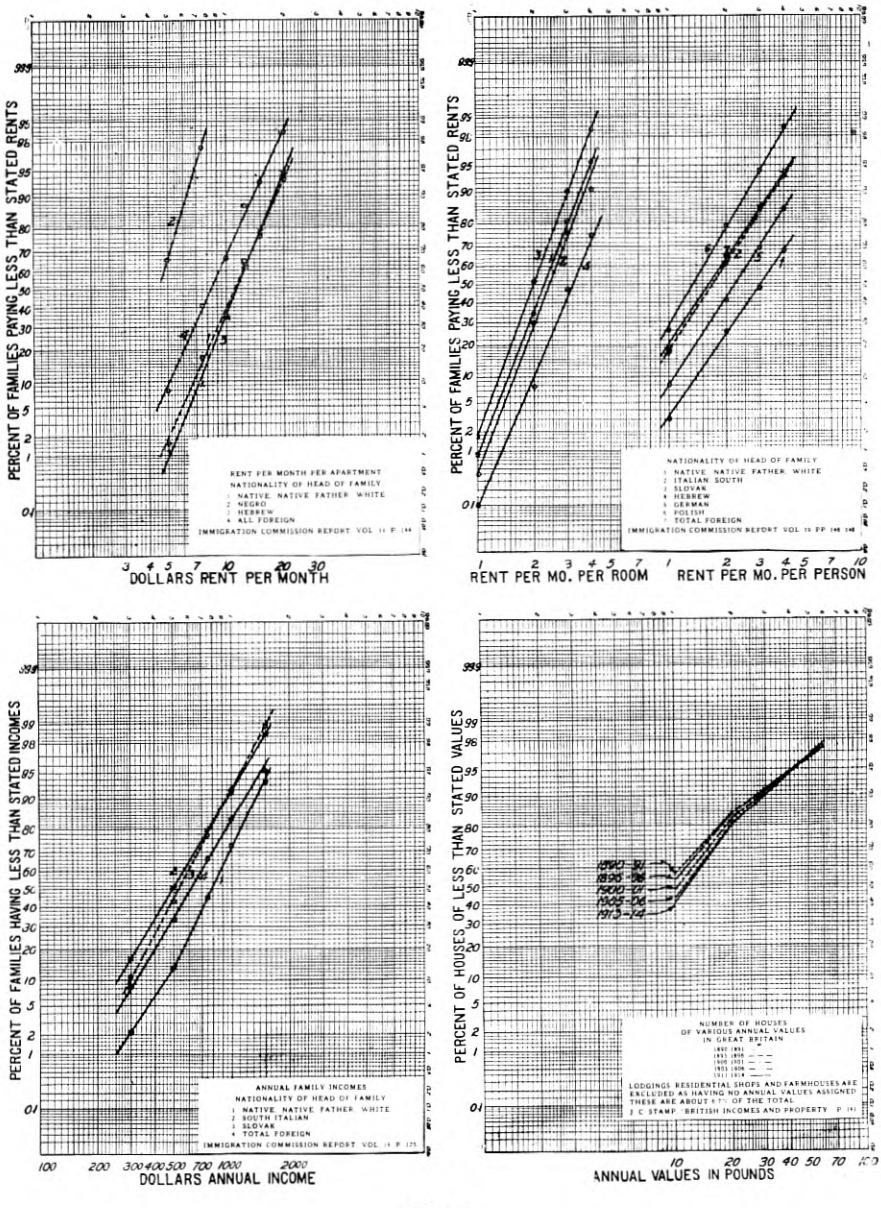

PERCENT OF FAMILIES PAYING LESS THAN STATED RENTS
DOLLARS RENT PER MONTH
RENT PER MO. PER ROOM
RENT PER MO. PER PERSON
PERCENT OF FAMILIES HAVING LESS THAN STATED INCOMES
DOLLARS ANNUAL INCOME
PERCENT OF HOUSES OF LESS THAN STATED VALUES
ANNUAL VALUES IN POUNDS

Fig. 5

detail or cumulative basis. This law does not hold for the lower income levels which may be best represented by a curve of approximately hyperbolic form, as shown in Fig. 6. The same income data are shown plotted on logarithmic probability paper in an insert on

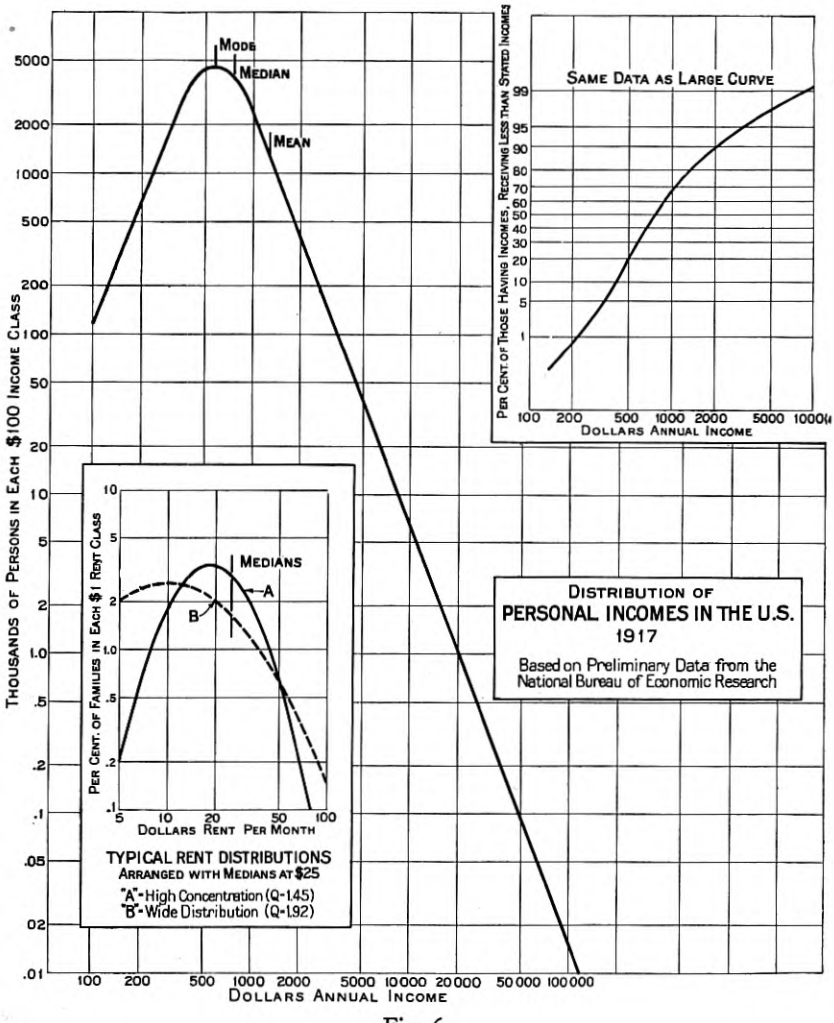


Fig. 6

Fig. 6. From the form of this curve it may be concluded that the distribution of the lower two-thirds of both incomes and rents is similar but that the spread of the higher incomes is much greater than the spread of the corresponding rents.

Another comparison of incomes and rents may be made from a second insert on the same chart. It may be readily demonstrated that the normal curve of error plots as a parabola on semi-logarithmic paper and the logarithmic skew curve as a parabola on double logarithmic paper. Two parabolas which represent extreme conditions of spread and of concentration of rents in large cities are shown. If the degree of dispersion remains fixed a change in the rent level merely shifts the parabola on the chart without changing its shape. The parabolic shape of rent curves and the hyperbolic shape of the income curve indicate that rents are somewhat less concentrated locally about their mode,[8] but are more concentrated as an entire group than are incomes. These curves can not well be superposed for comparison since areas are not equivalent on different parts of the chart. Those incomes which are closely grouped around the mode represent wage-earners of such a type that several may come from a single family. Conclusions regarding comparison of incomes and rents must be made with caution since rents are on a family basis and incomes on an individual basis. No satisfactory data are available to show the variation in income distribution between small subdivisions of the United States, as cities, but it is reasonable to assume that there is some such variation for incomes as well as for rents, although perhaps not of so great a range.

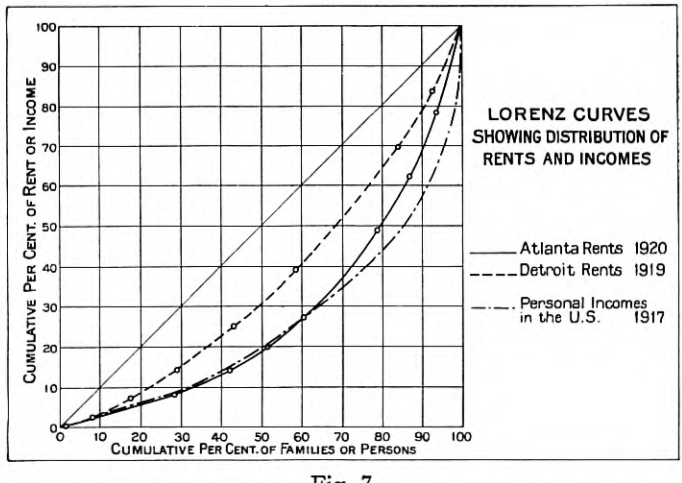


Fig. 7

A third comparison of incomes and rents is made possible by the use of Lorenz curves illustrated in Fig. 7. On this form of chart a

[8] See Appendix.

diagonal line at 45° represents a uniform distribution and the further a given curve falls to the right of and below that line the more unequal the distribution represented. If incomes were plotted on a family basis, the resulting curve would lie somewhat closer to the diagonal line than the one shown, but it is fairly evident that incomes are more unequally distributed than rents. For instance, the top 10 per cent of incomes are on the average about 42 per cent of the total income, while the top 10 per cent of rents are from 22 to 32 per cent of the aggregate rent in most cities. These three comparisons confirm the idea discussed in the first part of this paper, that the proportion of income spent for rent is less among the larger incomes.

Of the extensive data on income distributions few can well be used for comparison with rent data. In order that a cumulative curve really mean anything, it must represent an entire group, not merely items from one end or the other of the complete scale. Therefore, the various tables of earnings of working class families and individuals are of doubtful use here, although they do show, plotted on double logarithmic or logarithmic probability paper, that the type of distribution of earnings about an average value is practically identical for various nationalities in similar industries, or for men, women and children in all industries. However, the average earnings of the various classes are widely different. A few examples are shown in Fig. 5.

Income tax returns are of some interest although they are defective in several respects: they only include the upper part of society, a large number of persons fail to make returns and large amounts of income are tax exempt. Federal income tax data, which are available on a uniform basis for the years 1917–1920, may best be studied by plotting on double logarithmic paper, preferably after reducing the figures for the various states to a basis of returns per 1000 population. There are small changes from year to year in the position of the curve for any given state, which are not significant, since they may be due either to changes in the average income, or to increased efficiency of tax collection. Changes from year to year in the slope of the curve for any one state are small, indicating that there exists in each state a definite type of distribution of wealth and earning power. Differences in the position and slope of the curves for different states are conspicuous, indicating that both the per capita income and the distribution of the total income among individuals are different in different states. New York, for instance, shows a wide spread, *i.e.*, a relatively large number of very high incomes, and Iowa shows a narrow spread, *i.e.*, a large number of incomes around

$2000–5000, and comparatively few incomes over $20,000. New Jersey occupies an intermediate position. Alabama, more or less typical of Southern states, shows a much smaller number of returns in proportion to population than any of these states, and a distribution nearly, but not quite, as closely grouped as Iowa. The fact that a particular shape of curve is typical of a given state, and that the curves are different for different states, corresponds to similar characteristics of rent curves for cities. British income statistics show about the same degree of dispersion as do returns for the United States as a whole.

Distributions of the logarithmic skew type may be found in other fields than those of incomes and house rents. The theory has been advanced by some statisticians that while the normal curve of error is characteristic of observational errors, errors of estimate agree with that law if logarithms rather than actual estimates be considered. Price fluctuations, corporation earnings, and the profits of farmers are distributed in a similar manner. The lengths of life of telephone contracts agree quite closely with this type of distribution, if we allow for the fact that very long lives are relatively few in number because they started when the telephone business was comparatively small. A peculiarity of rent distribution is that if we choose only families having telephone service, or families having any one class of service, we obtain a logarithmic skew distribution about as closely as though we plotted all families in a city.

Application to Survey Data. The charting method described above was applied to rent data for 57 cities, both for composites of private residences, flats and apartments and for private residences alone. Table VIII gives the median rents and values of the rent dispersion index Q[9] for the composite data. Results for private residences differ in most cases only very slightly from the results given; there is no dominant tendency for the spread of private residence rents to be greater or less than that of all rents in a city, but the median rent in private residences is usually somewhat greater than that for the composite.

An effort was made to determine the significance of the various values of the index Q, but the results are chiefly negative. There is some tendency for the smaller cities to have a wide spread of rent values; *i.e.*, a high value for Q, but there is considerable scattering of the data. This tendency is most apparent in the South, where the smaller cities have extremely high values for Q. The relationship between the index Q and the per cent of families with telephone

[9] See Appendix for a quantitative definition of Q.

service is not very well defined. Cities with a very high residence development have low values for the index and Southern cities with poor residence development have high values for Q, but the intermediate scattering of data is quite wide. It might be supposed that cities with high values for Q, which indicate a wide spread of social strata, would have a relatively large number of business firms, either total or retail, to meet the widely divergent needs of the population. As a matter of fact, no such relationship is apparent. There is, however, positive correlation between Q and the proportion of institutions to population. This may be due in part to the fact that high values of Q are found in Southern cities which have separate churches and schools for whites and negroes.

Although no special significance has been found for the particular degree of rent dispersion found in any city, some interest attaches to the fact that this index remains practically constant in a given city, regardless of changes in the level of prices. The diagrams illustrating this point have already been discussed. If the type of distribution is not found constant in a particular city, it would seem probable that a change in character of the population is taking place, but a change in the average economic grade might occur without any change in the type of distribution. When two distributions, each of which agrees with a logarithmic skew curve, are added together, the new combined distribution may be represented by another logarithmic skew curve only in case both the medians and coefficients of dispersion for the two original curves are identical. It follows that if the index of rent dispersion in a city is found to be the same in successive surveys and if it may be assumed to have remained constant during the interval, then the new families which have come into a city at any time comprise a group having substantially the same coefficient of dispersion and median rent as the families which made up the original population. The apparent permanence of the type of rent distribution in a city may be considered, along with the telephone habit, as a reasonable explanation of the rather high degree of stability of station distribution by classes of service among residence subscribers.

In commercial survey work a city is divided into *market areas*, known also as *homogeneous sections*, which are so laid out that in any one section the families at any stated rent are similar telephone prospects. A study of rent distributions in market areas was carried out in a number of cities, considering only those market areas in each city which had fairly large populations. There appears to be no relationship between the index of rent dispersion and either the median rent, the per cent of families in private residences, or the per cent

of families with telephone service, in the various areas in any one city. Whether a particular area is suburban or downtown, likewise has no apparent effect on the value of Q. It was found in Atlanta, where the division of the city into market areas was substantially the same in successive surveys, that the distribution index which had previously been found to be stable for cities as a whole, behaved in the same way in separate sections of the city. It was found that the rent distribution index for any single market area is smaller, usually much smaller, than the index for the entire city in which the area is located. One section in Atlanta is the only exception found to this rule. In market areas it was noted that a considerable number of the graphs on logarithmic probability paper were formed of two intersecting straight lines. This indicates that the sections are not really homogeneous, but contain elements of population radically different in character. This condition can not be obviated by the most careful laying out of section boundaries in case there exists a mixture of families of essentially different types, as when negro residences are scattered among a predominantly white population.

TABLE VIII

Indices of Rent Distribution in Large Cities

Composites of Private Residences, Flats and Apartments

	Year	Per Cent Families in Private Residences	Per Cent Families with Service	Median Rent	Q = Upper Quartile ÷ Median
New England and Eastern					
Washington	1922	66.9	43.0	$35.00	1.71
Pittsburgh	1922	61.1	37.4	28.50	1.54
Baltimore	1914	68.6	16.4	13.50	1.51
New Haven	1919	24.6	24.5	21.00	1.44
Portland, Me.	1921	36.8	49.5	23.80	1.40
Hartford	1915	20.1	25.8	19.00	1.37
Providence	1916	26.6	26.5	14.60	1.34
Springfield, Mass.	1921	29.2	45.8	30.60	1.34
Bridgeport	1920	24.2	21.4	26.50	1.32
Philadelphia	1917	81.6	18.4	17.00	1.32
Altoona	1922	90.2	45.8	24.00	1.27
AVERAGE		48.3			1.41
Central					
Chicago	1920	22.4	50.0	27.00	1.55
Cleveland	1921	45.3	32.4	35.50	1.55
Evansville	1916	89.0	34.6	12.00	1.54
Grand Rapids	1915	68.4	33.4	12.80	1.52
Milwaukee	1921	40.0	39.6	25.00	1.52
Indianapolis	1920	84.4	53.8	21.00	1.51
Akron	1920	73.0	19.3	34.00	1.41
Detroit	1919	49.9	30.6	32.00	1.40
Youngstown	1919	83.0	40.8	27.00	1.37
Toledo	1920	76.6	42.0	26.00	1.35
AVERAGE		63.2			1.47

TABLE VIII—*Continued*

	Year	Per Cent Families in Private Residences	Per Cent Families with Service	Median Rent	Q = Upper Quartile ÷ Median
Southern					
Montgomery	1913	93.6	23.7	5.50	2.82
Macon	1913	94.6	21.8	6.00	2.41
Charlotte	1914	95.8	24.6	8.00	2.30
Savannah	1916	63.5	16.2	7.80	1.98
Birmingham	1920	94.6	17.8	13.50	1.96
Memphis	1915	86.4	18.5	10.50	1.90
Atlanta	1920	83.3	28.6	19.00	1.89
Mobile	1918	92.4	19.7	7.50	1.87
Chattanooga	1915	89.2	23.7	9.00	1.83
Richmond	1922	51.6	36.5	20.00	1.75
Jacksonville	1919	75.2	24.9	13.00	1.75
Louisville	1920	71.0	28.6	14.50	1.65
New Orleans	1916	85.3	11.6	12.00	1.46
AVERAGE		83.3			1.97
Southwestern					
Tulsa	1919	88.5	40.8	$33.00	1.85
Fort Worth	1921	92.0	37.3	24.00	1.82
Little Rock	1920	93.5	40.1	19.00	1.74
Houston	1921	87.8	44.3	23.40	1.67
San Antonio	1921	91.3	31.5	20.50	1.62
Dallas	1921	88.1	54.4	38.00	1.60
Kansas City	1916	71.4	32.8	16.80	1.52
St. Louis	1917	35.3	25.0	15.00	1.50
St. Joseph	1916	90.0	41.0	13.80	1.49
Oklahoma City	1918	85.2	46.4	23.00	1.43
AVERAGE		82.3			1.64
Northwestern					
Omaha	1921	82.8	69.4	33.40	1.54
Sioux City	1918	87.6	51.0	21.50	1.53
Des Moines	1916	85.8	48.8	18.00	1.50
Duluth	1915	62.0	52.0	18.00	1.48
Lincoln	1919	87.8	60.0	23.80	1.48
Minneapolis	1921	51.6	57.2	31.00	1.45
AVERAGE		76.3			1.49
Pacific and Mountain States					
Butte	1914	75.0	35.2	17.00	1.50
Portland	1916	80.8	49.0	13.80	1.48
Denver	1917	83.7	42.4	16.00	1.47
Seattle	1918	75.7	46.8	22.00	1.46
San Diego	1918	81.8	42.4	16.00	1.44
Salt Lake City	1917	86.0	51.2	18.50	1.43
Los Angeles	1917	78.4	47.3	18.50	1.41
San Francisco	1918	53.1	47.4	21.50	1.40
Spokane	1921	86.0	57.0	23.00	1.39
Sacramento	1918	62.6	46.7	19.00	1.39
Tacoma	1921	87.8	46.3	24.00	1.35
AVERAGE		77.4			1.43

APPENDIX

Mathematics of the Logarithmic Skew Curve

Frequency curves may be symmetrical or skew. The particular symmetrical distribution known as the normal curve of error is typical of distributions of observational errors and in general of all phenomena obeying the laws of chance. It is approximated by a number of other distributions which have not obviously originated in the same way, which implies that "the variable is the sum of a large number of elements each of which can take the values 0 and 1, these values occurring independently and with equal frequency." Skew distributions may take a variety of forms but the type shown in the diagram is closely approached by a large number of rent distributions. The essential characteristic of this curve, which may be called the *logarithmic skew curve*, is that logarithms of the values of the variable are distributed according to the normal curve of error. This skew curve is of course not the only one which might be selected to represent rent data, but it presents the fewest mathematical difficulties and gives a sufficiently close approximation for all practical purposes.

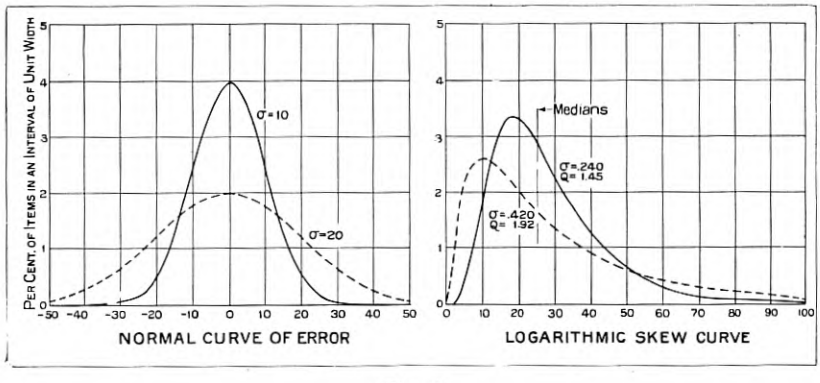


Fig. 8

The normal curve of error has the equation:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, \qquad (1)$$

where e is 2.7183, the base of natural logarithms and σ is a measure of dispersion, known as the standard deviation. An ordinate of the curve is called the frequency, and expresses the fraction of the whole number of items which occurs per unit interval of the variable x.

Substituting $\log X$ for x, a new equation may be obtained, in which y is the frequency per unit of x (or $\log X$). An expression for the frequency per unit of X is desired, which may be called Y. It may be shown that the desired equation is

$$Y = \frac{1}{X\sigma\sqrt{2\pi}} e^{-\frac{(\log X)^2}{2\sigma^2}}. \tag{2}$$

This is the equation of what we have called above the *logarithmic skew curve*, which is really not a curve of error in the same sense as equation (1) is.

In the course of this discussion it will be convenient to refer to certain features of the frequency curves by the accustomed terminology of statistics. The median item of a group is such that one-half of all the items are larger, and one-half are smaller, and is the central item when they are arrayed in order of size. The quartiles, upper and lower, together with the median, divide the array into four parts, each containing one-fourth of the items. The percentiles divide the array into 100 equal parts. The mode is that value of the variable which is of most frequent occurrence.

In the normal curve of error σ, the standard deviation, is technically defined as the square root of the mean of the squares of the deviations of the items from their mean. For present purposes it may be regarded as a measure of dispersion approximately equal to the difference between the values of x at the 84th percentile and the median. In the logarithmic skew curve σ is the difference between the corresponding logarithms.

The origin of x in the normal curve of error is the arithmetic mean, median, or mode, which are coincident. When a logarithmic scale of abscissas is introduced, the median value of x (or $\log X$) corresponds to the median value of X, which is smaller than the mean value of X, and larger than the mode. In a logarithmic skew curve the median may be considered the origin, and at this point x (or $\log X$) is equal to zero, and X is equal to unity. When this curve is applied to house rents the median rent occurs at this point. The relation between rents and values of X is a simple one. If rents be denoted by R, and if M be the median rent, then

$$R = MX. \tag{3}$$

The relationship of the various scales is presented in Fig. 9. The scales for X and $\log X$ may be considered fixed, and the scale for

R a movable one, as on a slide rule, corresponding values always being opposite each other.

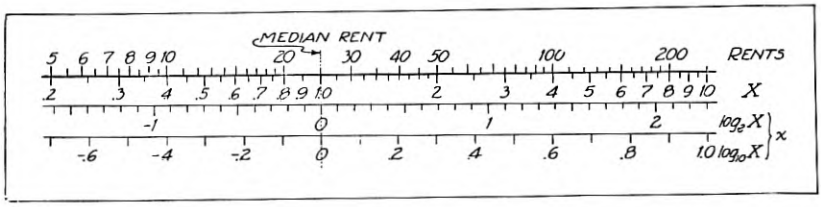


Fig. 9

Equation (2) above for the logarithmic skew curve gives the frequency per unit of X. The frequency, per unit of rent when expressed in dollars, is $1/M$ times that value, and substituting for X from equation (3), there results

$$\frac{Y}{M} = \frac{1}{R\,\sigma\sqrt{2\pi}}\, e^{-\frac{(\log R/M)^2}{2\sigma^2}}. \tag{4}$$

If it is desired to make computations from this equation, it is best to use the base 10 for logarithms rather than the natural base. For this purpose the equation becomes approximately

$$\frac{Y}{M} = \frac{.1733}{R\,\sigma_{10}}\, 10^{-\frac{.2171}{\sigma_{10}^2}(\log_{10} R/M)^2}. \tag{5}$$

For convenience it may be set down that

$$\sigma_{10} = 0.4343\,\sigma_e,$$

and

$$\sigma_e = 2.3026\,\sigma_{10},$$

although it will very rarely be necessary to make such computations.

It has been stated above that the median value of X (or rents expressed in dollars) may be logically regarded as the origin of the logarithmic skew curve, although X is not equal to zero at this point, but is equal to 1. If some of the other forms of statistical averages are also known, the properties of the curve may be better understood. To determine the mode, the first derivative of equation (2) of the curve is equated to zero, and there results

$$\log X = -\sigma^2,$$

or

$$X = e^{-\sigma^2}, \tag{6}$$

$$= 10^{-2.3026\sigma_{10}^2}.$$

Equation 6 above defines the peak of the curve, or the mode of the variable.

The arithmetic mean for a distribution agreeing with the logarithmic skew curve probably can not be defined by any mathematical expression sufficiently simple for practical use. It is a function of σ, but its exact position on the curve has not been determined. As applied to rent data, the mean may be computed direct from a house count summary with an error of two or three per cent. Thus found, its position on the curve is in the neighborhood of the 65th to 70th percentile.

The geometric mean coincides with the median for a logarithmic skew distribution. This follows from the fact that the median value of X corresponds to the median, which is also the mean, value of $\log X$.

The measure of dispersion for a logarithmic skew curve is also a measure of skewness. Up to this point σ has been used as the measure of dispersion, in agreement with conventional usage. For practical purposes another measure may be substituted, which has a more readily understood meaning. This is the *quartile deviation*, known also by the misleading term *probable error*. The quartile deviation for a logarithmic skew curve is that deviation either above or below the median which includes one-fourth of all the items in the array. It may, like σ, be measured in logarithms, and

$$\text{Quartile Deviation} = 0.6745\,\sigma.$$

Perhaps the easiest mathematical conception of a measure of skewness and dispersion is that of the ratio of the upper quartile to the median. This is identical with the ratio of the median to the lower quartile, and is the number whose logarithm is the quartile deviation as defined above. We shall let this ratio be denoted by Q.

Either σ or the quartile deviation for a given set of data may be best determined from a straight line graph on logarithmic probability paper. The following table gives the positions in the array for certain convenient multiples of σ and the quartile deviation, when measured in logarithms.

Deviation from Median	*Percentile Position*
σ	84.13
2σ	97.72
3σ	99.865
Quartile Deviation	75.0
2 Qu. Dev.	91.13
3 Qu. Dev.	97.85
4 Qu. Dev.	99.65

To obtain the value of the ratio Q, take the antilogarithm of the quartile deviation determined in this manner. For ordinary purposes it is sufficiently accurate to obtain Q as the ratio of the upper quartile to the median, read from the 75th and 50th per cent lines on the graph.

Power Losses in Insulating Materials

By E. T. HOCH

SYNOPSIS: It is shown that a satisfactory measure of power loss in a dielectric is the product of phase angle and dielectric constant. Although the dielectric constant need not be explicitly considered in the design of condensers, it is important in such cases as the design of apparatus panels, and vacuum tube bases. The method used in measuring phase angle and dielectric constant is discussed.—*Editor.*

IN working with electrical circuits operating at very high frequencies and moderately high voltages, such as radio transmitting circuits, it is found that failure in the insulation is seldom due to puncture or flashover as is usually the case at power frequencies, but is generally due to excessive heating which, in turn, causes both mechanical and chemical disintegration. As this heating is due almost entirely to the energy losses occurring in the dielectric itself, it is essential that the factors involved in the calculation of these losses be well understood.

In the past, various indices have been used as a measure of power losses for the purpose of comparing different dielectrics. Of these, power-factor, phase difference and watts per cubic centimeter probably are the most common. None of these, however, is very satisfactory for this purpose since the first two give only part of the desired information, and the last is not in any sense a property of the material, as it is dependent on both the voltage gradient and the frequency.

However, it can be shown that the product of the phase difference and the dielectric constant of a material is to a sufficient approximation an index of its power losses. Let us consider for a moment the complete expression for dielectric loss. In any condenser the capacity

$$C = a\,K$$

where a is a constant depending on the geometrical dimensions, and K is the dielectric constant. If a voltage, E, is applied to the condenser the power loss

$$P = E\,I\,\sin\Psi,$$

where I is the current through the condenser and Ψ is the phase difference of the dielectric; $\sin\Psi$ being the power factor. (Plate

resistance assumed negligible.) For small angles this may be written

$$P = E\,I\,\Psi,^{1}$$
$$= 2\pi f\,E^2\,a\,K\,\Psi,$$

since $I = 2\,\pi f\,E\,C$, f being the frequency.

In the particular case of a condenser of two parallel plates

$$a = m\,\frac{A}{d},$$

where m is a constant depending on the units used, A the area of one plate, and d the thickness of the dielectric.

Hence $$P = 2\pi f\,E^2\,m\,\frac{A}{d}\,K\,\Psi.$$

But the volume of dielectric $V = A\,d$, and the voltage gradient $E_g = \frac{E}{d}$. Therefore the power loss per unit volume is

$$\frac{P}{V} = m'\,E_g^2\,f\,K\,\Psi, \qquad (1)$$

where $m' = 2\,\pi\,m$, and $m'\,K\,\Psi$ = loss per unit volume at unit frequency and potential gradiant.

Thus it is seen that while no single factor of the expression can be used to represent the losses, the product of phase difference and dielectric constant [2] can be used in this way. Furthermore, for most good insulators, this product remains fairly constant throughout a considerable range of voltage and frequency. For example, we have found that for such materials as wood, phenol fibre, and hard rubber, the change of this product with frequency is of the order of 20 per cent from 200,000 cycles to 1,000,000 cycles. Hence it is possible to compare directly the losses in different materials even though the measurements were not made at exactly the same frequency.

If Ψ is taken in degrees, E_g in volts per centimeter, and f in cycles per second, the constant m' reduces to $0.97\times10.^{-14}$. Hence, for a frequency of 1,000,000 cycles per second and a potential gradient of 10,000 volts per centimeter, the product of K and Ψ (in degrees) is within 3 per cent of being numerically equal to the dielectric loss in watts per cubic centimeter.

[1] The substitution of the angle for its sine is correct to better than 5 per cent for angles as large as 30°.

[2] This relation has been brought to the attention of the Committee on Electrical Insulating Materials of the American Society for Testing Materials and is included in their "Tentative Method of Test for Phase Difference (Power Factor) and Dielectric Constant of Molded Electrical Insulating Materials at Radio Frequencies."

Data showing the variations with frequency and temperature of the phase difference, dielectric constant, and their product, for several materials are given in Tables I. and II. below.

TABLE I.

Dielectric Constant, Phase Difference and Their Product for Several Commercial Insulating Materials [3]

Material	Frequency C. P. S.	Dielectric Constant	Phase Difference Degrees	Product
Phenol Fibre A	295,000	5.9	2.9	17.1
	500,000	5.8	2.9	16.8
	670,000	5.7	2.9	16.5
	1,040,000	5.6	3.3	18.5
Phenol Fibre B	190,000	5.8	2.2	12.7
	500,000	5.6	2.5	14.0
	675,000	5.6	2.6	14.6
	975,000	5.6	2.8	15.7
Phenol Fibre C	200,000	5.4	2.1	11.3
	395,000	5.4	2.2	11.8
	685,000	5.3	2.3	12.2
	975,000	5.2	2.4	12.5
Phenol Fibre D	194,000	5.4	4.2	22.7
	500,000	5.2	3.9	20.3
	695,000	5.2	3.9	20.3
	1,000,000	5.1	3.8	19.4
Wood (Oak)	300,000	3.2	2.1	6.7
	425,000	3.3	2.0	6.6
	635,000	3.3	2.2	7.3
	1,060,000	3.3	2.4	7.9
Wood (Maple)	500,000	4.4	1.9	8.4
Wood (Birch)	500,000	5.2	3.7	19.2
Hard Rubber	210,000	3.0	.5	1.5
	440,000	3.0	.5	1.5
	710,000	3.0	.5	1.5
	1,126,000	3.0	.6	1.8
Flint Glass	500,000	7.0	.24	1.68
	720,000	7.0	.24	1.68
	890,000	7.0	.23	1.61
Plate Glass	500,000	6.8	.4	2.7
Cobalt Glass	500,000	7.3	.4	2.9
Pyrex Glass	500,000	4.9	.24	1.18

[3] All of the samples had been in the laboratory for some time during summer weather without artificial drying or other special preparation.

TABLE II.

Variation with Temperature of Dielectric Constant, Phase Difference and Their Product for Some Commercial Insulating Materials (Frequency 500,000 C. P. S.) [4]

Material	Temperature Degrees C	Dielectric Constant	Phase Difference Degrees	Product
Molded Phenol Product A.....	21	5.6	3.1	17.4
	71	6.9	6.5	45.0
	120	10.4	22.0	230.0
	21	5.5	2.9	16.0
Molded Phenol Product B.....	21	5.2	2.3	12.0
	71	6.1	3.7	22.5
	120	7.6	8.9	68.0
	21	5.2	2.3	12.0
Molded Phenol Product C.....	21	5.3	2.8	14.8
	71	6.1	3.6	22.0
	120	6.7	9.6	64.0
	21	5.0	2.5	12.5
Phenol Fibre B.....	21	5.6	2.5	14.0
	71	6.6	3.1	20.5
	120	6.5	4.6	30.0
	21	5.4	2.4	13.0
Phenol Fibre C.....	21	5.4	2.3	12.4
	71	6.0	3.9	23.5
	120	5.3	4.9	26.5
	21	4.9	2.4	11.8
Phenol Fibre D.....	21	5.2	3.9	20.3
	71	6.6	6.9	46.0
	120	6.3	13.5	85.5
	21	5.1	3.1	15.8
Hard Rubber.....	21	3.	.5	1.5
	71	3.1	1.2	3.7
	120	3.2	3.7	11.8
Pyrex Glass.....	20	4.9	.24	1.18
	74	5.0	.4	2.0
	125	5.0	.7	3.5
	19	4.9	.25	1.22

The above data were obtained by the resistance variation method,[5] Fig. 1. Each value of phase difference and of dielectric constant represents the average of at least five readings on a single sample using not less than three different values of the known resistance R. A condenser, the dielectric of which consists of the material to be tested, is connected in series with a suitable inductance, a known re-

[4] The measurements on each sample were made in the order in which they are given in the table.

[5] Bureau of Standard Circular No. 74, p. 180.

sistance which can be varied, and a radio frequency ammeter. An oscillator is coupled loosely to the inductance and its frequency varied until resonance is obtained as indicated by maximum current through the meter. Without changing the tuning, the resistance is changed

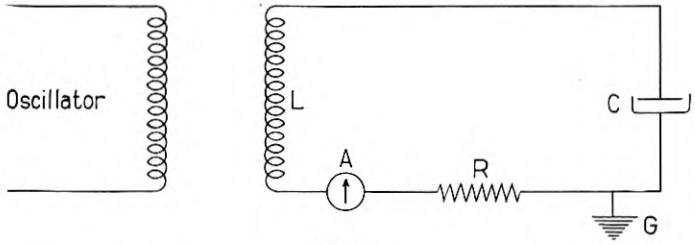


Fig. 1

and a second reading of current is obtained. Then, since the e.m.f. induced in the measuring circuit is the same in both cases, if R_1 and R_2 are the known resistances, I_1 and I_2 the corresponding currents, and r the resistance of the remainder of the circuit,

$$I_1 (r + R_1) = I_2 (r + R_2),$$

$$r = \frac{R_2 I_2 - R_1 I_1}{I_1 - I_2},$$

or, if R_1 be made zero

$$r = \frac{R_2}{\frac{I_1}{I_2} - 1}.$$

A standardized variable air condenser having negligible resistance is then substituted for the condenser under test and the process repeated except that the circuit is tuned to resonance by varying the capacity instead of the frequency. In this way the resistance of the circuit exclusive of the test condenser may be determined. The difference between these two circuit resistances is the resistance of the test condenser from which the phase difference may be computed. The capacity of the test condenser is equal to the capacity of the standard condenser which produces resonance. From this and the dimensions of the sample, its dielectric constant may be computed.

In addition to the general precautions mentioned in the Bureau of Standards Bulletin, two others should be observed in the measurement of dielectrics. First, the electrodes must be in intimate contact at all points with the surface of the sample, as a very small air space will cause a large error in the values of phase difference and di-

electric constant obtained for the sample. For this reason only mercury electrodes have been found suitable, the sample being floated on a pool of mercury forming the lower electrode and the upper electrode being formed by pouring a pool of mercury inside a metal ring on the upper surface of the sample. This introduces the second difficulty. The lower electrode being, of necessity, larger than the upper one, the electric field spreads out considerably beyond the edges of the upper electrode and increases the effective area by an unknown amount.

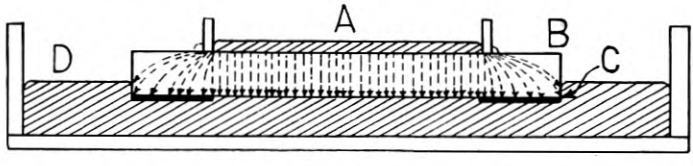


Fig. 2

To determine the magnitude of this error, measurements were made on samples prepared as shown in Fig. 2 A is the upper electrode, B is the sample under test and C is a tinfoil guard ring shellaced to the lower surface of the sample but separated from the lower electrode D by a sheet of paper shellaced over the guard ring. The guard ring covers all of the lower surface of the sample except the area equal to the upper electrode and directly under it. The direct capacity [6] between A and D is measured using the guard ring as a shield to intercept the flux which spreads out around the upper electrode, diverting it away from the measuring circuit. This measurement is made at audio frequency on a completely shielded substitution bridge. The difference between this capacity and the capacity of A to D without the guard ring is approximately the correction due to this edge effect. Using samples 6 inches square with an upper electrode 5 inches square, this correction was found to be about 7 per cent for samples $\frac{1}{8}$ inch thick and 14 per cent for samples $\frac{1}{4}$ inch thick. All values of dielectric constant given above have been corrected. The phase difference is not affected appreciably since it is dependent only on the ratio of resistance to reactance and does not involve the area of the sample.

The radio frequency generator used consists of a vacuum tube oscillator having a maximum output of 250 watts. The coupling between the generator and the measuring circuit was very loose and

[6] Direct Capacity Measurement, George A. Campbell, *The Bell System Technical Journal*, July, 1922.

care was taken to avoid capacity couplings. The measuring circuit was shielded from the observer by a metal screen. In spite of these precautions the results obtained on the same sample at different times do not agree as well as might be desired although the individual readings taken at the same time agree in most cases to within 5 per cent. Other observers have found that measurements made on the same sample at intervals of a few hours often differ by more than the apparent error of the measurements and have attributed it to actual changes in the properties of the material.[7] Hence it is possible that at least part of the apparent variation with frequency shown above is due to unknown errors in the measurements or unknown changes in the samples or both.

As an illustration of the error involved in taking only phase difference as a measure of power loss, suppose we wish to compare hard rubber having a phase difference of about 0.5 degree and a dielectric constant of 3., with a certain grade of glass having a phase difference of about 0.3 degree and a dielectric constant of 7. On the basis of phase difference alone the hard rubber appears very much worse than the glass, but when the dielectric constant is taken into account, the glass is found to give a 40 per cent higher power loss. Similarly, some untreated woods were found to have considerably lower losses than the phenol fibres although their phase differences are nearly the same.

Conclusion

In the case of ordinary insulation where the object is to provide a mechanical separator or support, the product of phase difference and dielectric constant is a true measure of the energy loss per unit volume as shown by the equation (1). In the case of a *condenser* where the object is to obtain a given *capacity*, the phase difference alone determines the power loss since in this case the effect of the increased dielectric constant is exactly balanced by the smaller volume of dielectric required.

[7] R. Mesing—*L'Onde Electrique*, April, 1922, p. 235 and Augustin Frigon—*Comptes Rendus*, May 22, 1922, p. 1339.

Application to Radio of Wire Transmission Engineering[1]

By LLOYD ESPENSCHIED

SYNOPSIS: This article points out that radio and wire communication systems are subject, fundamentally, to the same general requirements, and its purpose is to develop for radio, points of view which are familiar to wire transmission engineers. The transmission characteristics over a wide range of distances are compared. For short distances the comparison is favorable to wires. Although over great distances, the attenuation of electric waves, guided by wires, may be greater than the unguided waves of radio, it is pointed out that at the present time intermediate amplifiers can be more economically applied in wire transmission than in radio to boost the message energy. Economy of transmission requires the handling of messages at as low an energy level as possible and, as the author points out, wire transmission satisfies this requirement much better than radio. Referring to the transcontinental line with radio extensions, which was used recently to talk from Catalina Island in the Pacific Ocean to a ship in the Atlantic Ocean, it is stated that had all of the necessary energy been introduced at one end of the circuit, there being no intermediate amplification, the total power required would have been 1.8×10^{23} kilowatts, an amount unavailable in the world. In the actual system, distributing the amplification along the transmission line, the power required sums up to something less than 1 kilowatt.

Interference between messages and extraneous disturbances is discussed, and the requirements involved in keeping message energy well above the energy level of the disturbances in both systems are pointed out. The limitations on two-way operation resulting from "singing" of the entire system are considered for both cases and for combination wire and radio circuits as well. The method of improving the efficiency of transmission by suppressing the carrier and one side band is discussed. Finally the factors involved in obtaining high grade quality of transmission are enumerated.—*Editor.*

ONE of the most interesting aspects of the development of radio during the last few years, and particularly of radio telephony, is the obvious convergence of its technique with that of wire transmission. It is, of course, the advent into both of these arts of that remarkable device, the electron tube, which is responsible for the close technical relations which now exist between them.

This community of interest, however, altho thus greatly stimulated by a device of such range of utility as to find important applications in both arts, is not due primarily to any device *per se*, but rather to the fact that both type of systems are subject fundamentally, as communication systems, to the same general requirements and design considerations concerning their intelligence-carrying capabilities. These underlying communication requirements lead to similar considerations in both types of systems as to the efficiency and fidelity with which the transmission of intelligence is effected and give rise

[1] Presented before The Institute of Radio Engineers, New York, January 23, 1922. Received by the Editor April 17, 1922. Also printed in the *Procd. Radio Institute* for October, 1922.

to a transmission background, as it were, which is common to both arts.

The engineering handling of the transmission problems which arise from these fundamental communication requirements has been quite highly developed in the older of the two arts—wire transmission—in connection with telephone repeaters and carrier telephone and telegraph systems. It should be, therefore, interesting and profitable to apply some of the transmission technique thus developed in the wire art to several of the more important radio problems. In so doing, we obtain rather new viewpoints of radio transmission and a useful correlation of it with the better established wire methods. It is hoped, therefore, that the picture which is presented of radio and wire transmission, treated from a common standpoint, may contribute to a better appreciation of both arts by radio and wire engineers alike and may make clear the underlying transmission principles which are common to them.

Principal among the problems of electric communication is the one of delivering at the receiving end the required volume of signal with the necessary freedom from interference. The delivering of the required volume is a matter of overcoming the transmission losses of the system by amplification; while the obviating of interference is, of course, concerned with the reduction of the ratio of the interfering to the signaling energy.

Transmission Losses

In considering these factors we will take up first the primary one of the losses which are suffered by the carrier waves as they are propagated thru the transmission medium. In both wire and radio transmission, of course, the actual propagation of the electromagnetic wave energy occurs in the "ether," the difference being that in the wire case, the waves are bound to a guiding path, whereas in the radio case they are transmitted freely in all directions and bound merely to the earth's surface. This difference in the mechanism of transmission gives rise to an important difference in the transmission losses occurring in the two cases. In order to assist in visualizing the two cases they are indicated diagrammatically in Fig. 1.

Referring first to the wire case, the law in accordance with which the current and voltage strength decrease as the transmission wave travels along the wire, is the familiar one of attenuation.

$$\begin{aligned} I_1\epsilon^{-\alpha l} &= I_2, \\ E_1\epsilon^{-\alpha l} &= E_2, \end{aligned} \tag{1}$$

which simply expresses the fact that, as the wave proceeds along the wire, the losses in the resistance of the conductor and in the insulation, extract for each mile a certain definite proportion of the voltage and current which arrives at that point. After traveling (l) miles the original current I_1 is attenuated down to a value $I_1\epsilon^{-\alpha l}$ which represents the received current I_2. This is the same general law of damping as applies to the dying down of the voltage and current in an oscillation circuit, except that here the damping is with respect to distance along the line rather than time. We are assuming, of course, that the circuit is so terminated as to avoid reflection effects at the terminals—a condition readily met, by making the terminal impedance equal to the characteristic line impedance. This is indicated in the figure by the designations, Z (internal) equals Z (line). A similar relation is taken for the radio case. The " line " impedance is here the antenna radiation resistance while the " termi-

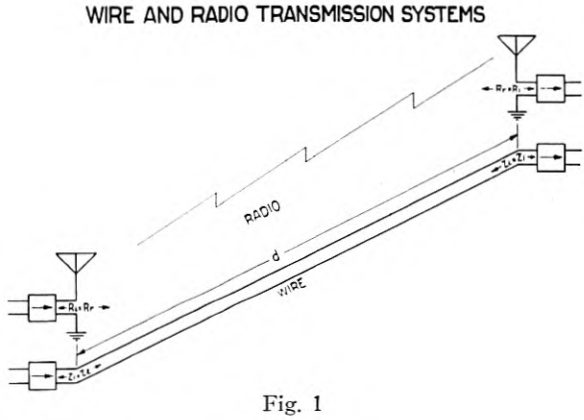


Fig. 1

nal " impedance is the resistance internal to the antenna and the apparatus, assuming resonance; thus R (internal) equals R (radiation).

We know that in radio there are two distinct causes of the transmission loss: (1) that, due to the spreading out of the waves, which is characteristic of non-guided wave transmission; and (2) that due to absorption in the air and earth's surface, which extracts a definite percentage loss for each mile of the radio circuit and which conforms, therefore, to an exponential law similar in its general nature to that of wire attenuation.

This transmission law, as expressed by the familiar Austin-Cohen

formula, is given in the Appendix.[2] In order to express the radio transmission loss in some general manner which will be comparable to the expression of wire transmission loss, these conditions have been taken for the radio case:

(1) That we will use as the measure of the transmission loss the ratio of the square root of the power radiated from the sending antenna to the square root of the power delivered in the receiving antenna. This is, of course, the same criterion as is used for wire transmission.

(2) The radiation resistance of the two antennas, sending and receiving, are made equal, analogous to the equality of line impedance at the two ends of the wire system.

(3) Also the internal antenna resistance, which corresponds to the terminating impedance in the wire case, is made equal to the radiation resistance for both ends. This is the condition of maximum power transfer between the "line" and the terminal.

These assumptions set up the two cases, radio and wire, on a comparable basis and facilitate a comparison of them. They are favorable to radio in that they do not take account of practical limitations which obtain in antennas. The radio curves should be read, therefore, as giving the minimum possible losses for daylight transmission over water.

These curves show the manner in which the transmission loss varies with distance, for various frequencies, for both radio and wire. The ordinates are plotted in terms of the logarithm of the ratio of the sent to the received currents, or voltages, in circuits of equal impedances. In so doing we are plotting the losses on the straight attenuation basis upon which they are usually plotted in the wire art; that is, the ordinates represent the exponent (αl) of the wire attenuation law, and may be directly interpreted in terms of miles of standard cable[3] by multiplying by 21. approximately. The advantage of dealing with the exponent rather than the current ratio

[2] Measurements on ship-shore transmission made since the above was written, indicate that the Austin-Cohen law holds quite well for frequencies as high as about 1,000,000 cycles.

[3] For the mile of standard cable the attenuation α (at 800 cycles) equals 0.109. Therefore the equation for current ratio, in terms of miles of standard cable, becomes

$$\frac{I_1}{I_2} = \epsilon^{\alpha l} = \epsilon^{0.109 l}$$

from which

$$l = \frac{1}{0.109} \log_\epsilon \frac{I_1}{I_2} = 21.13 \log_{10} \frac{I_1}{I_2}.$$

itself is the very considerable one which is characteristic of logarithms, namely, that when thus expressed the individual losses and gains thruout a system may be summed up algebraically, and the over-all transmission equivalent of the system thus readily determined.

It should be noted that the transmission loss given in the radio curves is that obtaining between the point at which power is delivered to the ether at the sending end and that at which it is delivered to the dissipative load of the receiving antenna circuit. In Fig. 1 these points are represented by R_r at the transmitter and R_i at the receiver. If at the sending end, we start with the power developed within the generator, meaning in R_i instead of R_r, then the power ratio is simply doubled, for the conditions assumed, and the attenuation is 0.15 units or about 3 miles greater than given in the curves. The curves can be used for obtaining the loss in any practical case simply by taking the minimum loss as given by the curves and adding thereto the additional loss obtaining in the actual antenna.

Referring now to Fig. 2—the transmission losses in the two cases are given for distances up to 200 miles (320 km.). The straight lines represent the wire losses, the bending-over curves the radio losses. Of the radio curves, the dash lines give the spreading-out losses alone, while the full lines give the total losses, including absorption.

The first thing one observes is the difference in the nature of the two sets of curves—the wire losses being represented by straight lines, because of their exponential law and the fact that it is the logarithm or the exponent itself which is being plotted, while the radio curves jump up rapidly at first and then straighten out, in accordance with the " inverse-with-distance " law.

The second thing one notes is the fact that as a result of the large initial (or " jump off ") loss, the radio values run on the whole higher than do the wire for the more usable wire frequencies, and very much greater than the wire losses at telephone frequencies (1 k.c.).[4] For the wire case the number 8 Birmingham wire gauge open wire circuit is taken.[5] This is the standard long distance telephone circuit of the United States. The constants are given in the appendix.

A third characteristic which one notes in the radio curves is that the losses are greater for the higher frequencies or, conversely, lower for the lower frequencies. This is because the efficiency of the antenna has been kept constant for all frequencies. In practice the

[4] 1 k.c. is 1 kilocycle per second or 1,000 cycles per second.

[5] Diameter of number 8 Birmingham wire gauge wire = 0.165 in. = 0.42 cm.

transmission losses at the lower frequencies are higher than here indicated because of limitations in antenna heights.

Were we to take the ideal condition *as regards the transmission medium itself*, where for wires there is no conductor or dielectric loss, and for radio there is, likewise, no earth or air absorption loss, we would note: (1) that, for wires, there would be no attenuation what-

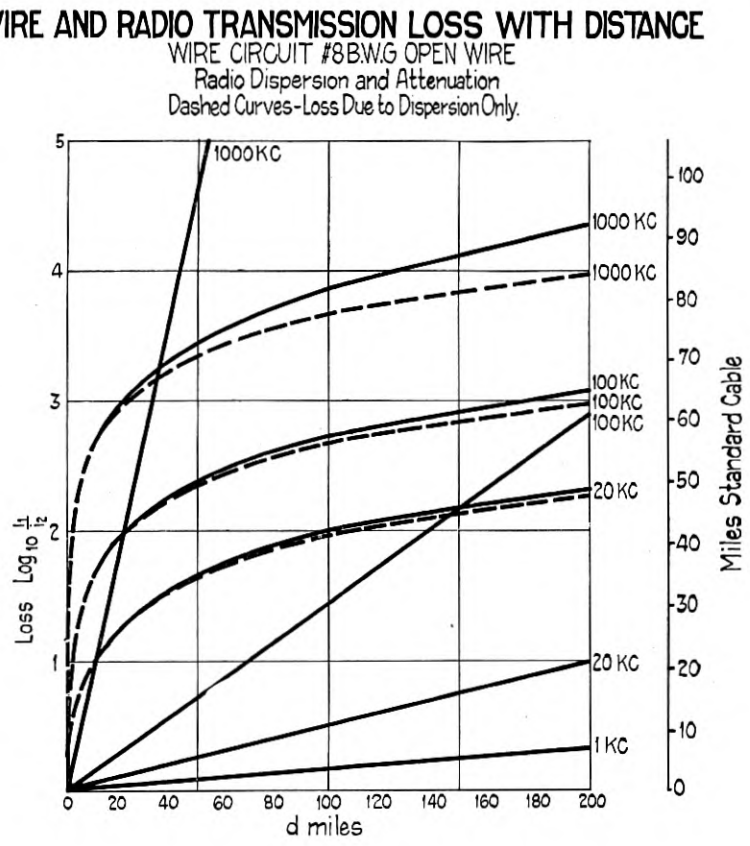


Fig. 2

ever, the curve following along the X axis; (2) for radio, there would remain the loss due to dispersion, inherent in the unguided method of transmission, the magnitude of which loss is, of course, very substantial. The dash-line radio curves show the radio losses without attenuation, the full line curves with attenuation.

Considering the actual condition, where there is dissipative loss

in the transmitting medium, we find that for moderate distances, up to 200 miles (320 km.), as plotted in Fig. 2, the wire losses are in general less, and at telephone frequencies very much less, than the radio losses. The low wire attenuation at telephone frequencies is, of course, in keeping with experience and accounts for the economical terminal apparatus which is employed in telephone practice. Likewise the relatively high losses for radio accounts for the large amplification at either the sending or receiving end or both, which experience has proven to be necessary. This brings in an interesting side-

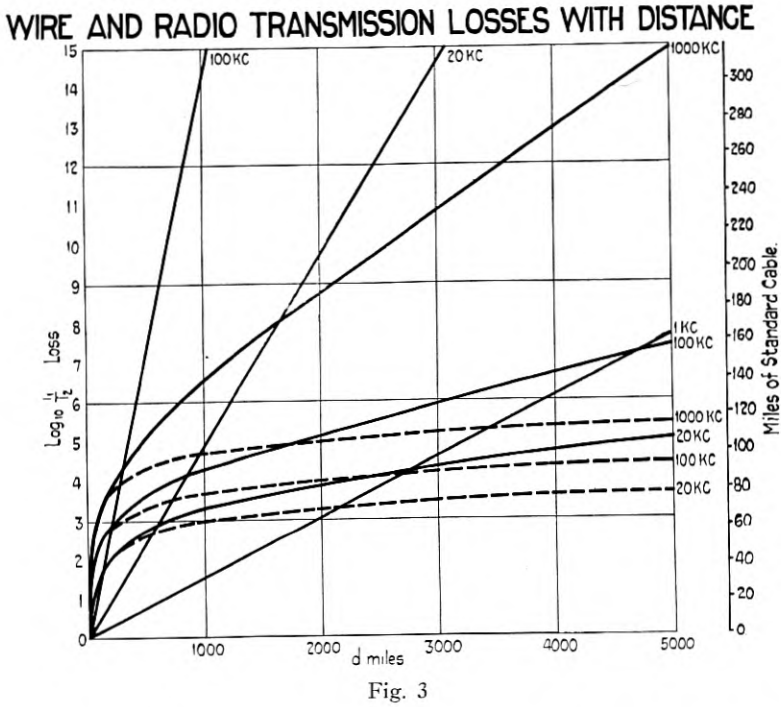


Fig. 3

light, namely, that altho in radio the transmission medium is provided by nature, the effective *use* of this medium is not as economical as might be expected because it requires considerable equipment, amplifiers at both ends for overcoming the large attenuations, selective means for dividing-up the frequency range and thereby "multiplexing" the ether, and antennas for getting into the medium and out again.

For the higher frequencies, the wire attenuations increase relatively more rapidly than the radio, thus limiting the frequency range

which can be employed on wires without likewise running into large amplification requirements. For example, the loss at 100,000 cycles for a distance of 200 miles (320 km.) is about as great over wires as the minimum loss which it is theoretically possible to obtain over radio.

Referring now to the attenuations for longer distances, as given in Fig. 3, it is of interest to note that for distances of the order of 2,000 to 3,000 miles (3,200 to 4,800 km.) the lower radio frequency curves cross the 1,000 cycle wire curve, meaning that for these distances it is possible for radio transmission to be as efficient as straight telephone transmission. The wires present to carrier frequencies for these long distances losses which are generally greater than prevail for radio.

These attenuation relations cannot be directly converted into an economic comparison, however, for the economies depend not upon the attenuation itself but upon, among other factors, the cost involved in *overcoming* the losses by means of amplification; and this cost in turn depends largely upon the extent to which the amplification can be applied at weak powers, as by the frequent application of telephone repeaters. By applying repeaters every few hundred miles in the wire case, the attenuation is prevented from piling up and the amplification is handled at relatively weak and therefore economical power levels. This brings us to the point of requiring that the attenuation values given above, be considered in reference to the amplification and power required to overcome them and yield the necessary volume of transmission at the receiving end over and above interference.

Interference and Its Effect Upon the Transmitting Power Required

In both the radio and wire cases there is always present in the transmission medium a certain amount of stray wave energy which tends to interfere with the proper reception of the message-carrying waves. It is necessary that the communication waves arrive at the receiving end of the system with such power as to be large compared with the interfering waves—by a factor determined by the type and grade of communication involved. Inasmuch as the stray energy always has some finite value, this requirement of freedom from interference will determine in the radio case as well as in some types of wire transmission the minimum wave power required at the receiving end of the transmission system.

In the wire system the minumum power requirement may be expressed directly in terms of a power, or—as it is usually—a current, in the receiving apparatus, the "transfer" between the line and the receiver being a constant and efficient relation. In radio the power delivered out of the ether or "line" into the receiving antenna is so

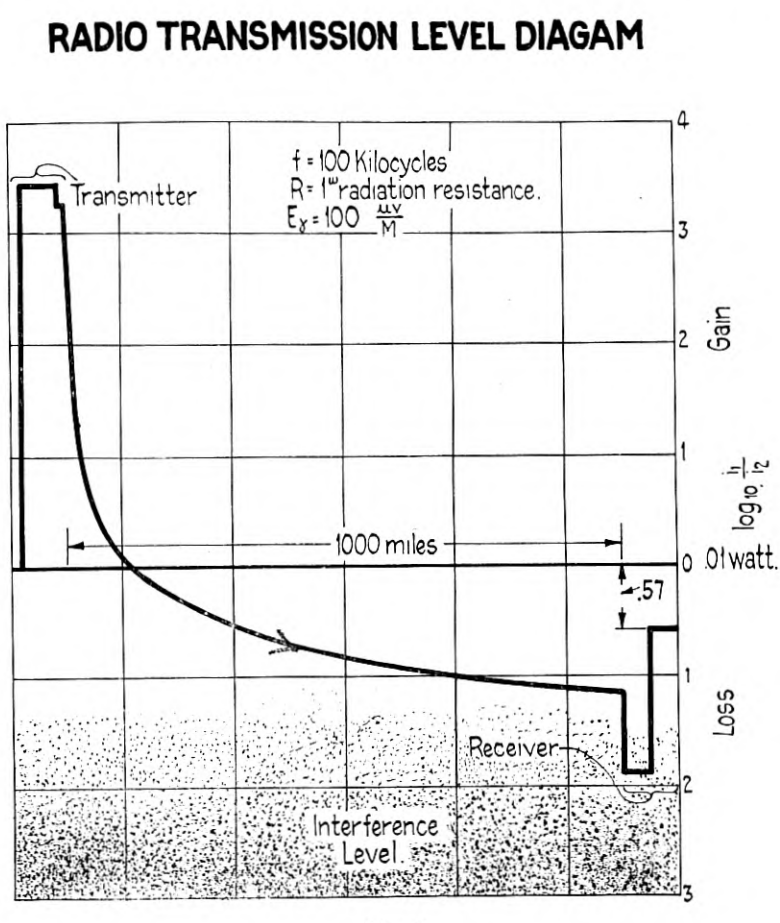


Fig. 4

largely a function of the antenna dimensions that it is necessary to express the necessary received power in terms of the received field, usually, simply as a field intensity—and not as a certain current in the terminal apparatus. Of course, the transmission loss occasioned by getting from the "ether" into the receiving circuit does not affect

the interference relation but merely the absolute amount of terminal amplification required. On the other hand the transmission loss occasioned by getting into the "ether" at the sending end affects the interference relation vitally, as we shall see.

Transmission Levels

This necessity of having to keep the power of the received waves above the interference level may be visualized by reference to Fig. 4. Here we have what in wire practice is called a "transmission level" diagram. Such a diagram is useful in showing what goes on in the system from the power and interference standpoints. The vertical scale is plotted in terms of the transmission level expressed as the logarithm of the current or field intensity ratios, and the horizontal scale represents progression along the system. For illustration purposes, the presence of interference is indicated at the bottom of the transmission-level scale by the shading.

Tracing thru the diagram we proceed as follows:

The point of "zero"level is taken roughly as that corresponding to the power delivered into a telephone circuit by a certain telephone transmitter when spoken into by the average talker, and is here taken to be 0.01 watt. As the voice currents are amplified to power proportions in the transmitting station, at the left, the transmission level is greatly increased, as illustrated by the vertical jump in the curve. The amplified voice currents are assumed to be converted by modulation into high frequency currents at this high power level and put into the antenna. The high frequency loss in the antenna system is indicated by the perpendicular jog in the curve. The drooping-off curve then commences, starting with a point which represents the power usefully applied to the ether in accordance with the expression I^2R, where I is the antenna current and R is the radiation resistance. The level curve falls off in accordance with the transmission loss curves previously discussed, as it extends across the transmitting medium to the receiving station. It will not do to permit the transmission level to fall as low as that of the interference, so I have shown that the transmission reaches the receiving station before dropping down very far into the interference level. At the receiving point a further transmission loss occurs in getting into the receiving antenna circuit, shown by the drop in the curve, but this loss obtains for the interference as well as the desired signals and does not affect the interference ratio. The terminal amplification brings the level up to that required for suitable audition and the difference between

where this level leaves off and the original zero level, measures the over-all transmission equivalent of the circuit, shown in this case as about $\log_{10} \frac{I_1}{I_2} = 0.57$ or about 12 miles of standard cable. This corresponds to a current ratio of about 4, a value ample for good " talk." Of course, in a one-way circuit the terminal amplification can be raised to any value desired. In a two-way circuit, however, a limit in the terminal amplification is imposed by interference between the two transmissions, as will be understood subsequently.

We may make the following useful observations from this curve:

1. The net transmission equivalent represents the difference between the over-all loss and the over-all gain.
2. The over-all gain is divided between the transmitting and receiving ends. We should like to throw as much of this amplification as possible to the receiving end because of the economy with which amplification can be provided at low powers.
3. The extent to which we can do this, however, is distinctly limited by the fact that the transmission level obtaining at the receiving end in the transmission medium must be held above a certain amount in order to overcome interference.
4. It is, then, the absolute intensity of the interference which determines the receiving power level required, and in turn this together with the attenuation back to the transmitting station which determines the transmitting power required.

Thus the two transmission features most fundamentally important in a radio communication system are (1) the interference level and (2) the transmission loss thru the medium. These once given, the other engineering considerations follow naturally. There are analogously fundamental factors in wire communication systems. In the latter case, however, the art has advanced to a point where means of controlling the interference level are available, so that the ratio of interference to transmitted power may be made small by decreasing the former rather than increasing the latter.

Minimum Transmission Levels Obtaining in Practice

The working value which should be assumed for the ratio between the transmission level of the received signals and the interference, depend upon the type of communication involved, whether it is telephone or telegraph, for example, and upon the grade of service to be given. There is a wide difference between the transmission level which will enable telephonic signals to be barely discerned by an

expert ear and that which is required for a public service communication system which must provide sufficient operating margin to enable the average person to converse with ease and certainty under all ordinary conditions. Under favorable static conditions, the transmission level can be permitted to fall to extraordinarily low values. When this condition is accompanied by a substantial reduction in the effective attenuation, which sometimes occurs at short wave lengths especially at night, apparently due to the effective absence of either absorption, then it becomes possible to "get thru" over relatively long distances with powers diminutive as compared with those required for giving a regular service. With these exceptional transmission conditions we are, of course, familiar. They are exemplified by the long distances reached at night by the amateurs, as across the Atlantic, and by the hearing of the normally 30-mile (48 km.) Catalina Island system in Australian waters. The transmission curves of Fig. 3 account for these unusual long distance transmissions if we assume that the attenuation due to absorption is eliminated on these occasions by some natural cause. Thus, at 3,000 miles (4,800 km.) the curves for 1,000 kilocycles (300 meters), for example, show that were the absorption eliminated, the transmission equivalent would be improved by the difference between about 10.8 and 5.2 for $\log_{10} \frac{I_1}{I_2}$, or 5.6, an improvement equivalent to a little over 100 miles of standard cable. The remaining or purely spreading-out loss of about 5 units, or 100 miles of standard cable, is then taken care of by the sending and receiving amplification.

Interference may occur in either or both of two ways—by the interference level rising to a point comparable with the normal transmission level at the receiving end of the ether circuit, or by the transmission level of the waves themselves dropping so low, due to excessive atmospheric absorption, as to fall below that of the atmospheric disturbances. For reliable transmission it is necessary, therefore, to deliver normally at the receiving end, a wave intensity sufficient to allow for the fluctuations which occur in atmospheric absorption and in the intensity level of the atmospherics. The importance of working to transmission level standards which give an adequate operating margin against interference, for the types of service required, will be appreciated from the foregoing. The following values of minimum transmission levels will be of value to know:

(a) For carrier wire telephone transmission at frequencies in the tens of thousandths of cycles, the limiting interference may be

our old friend "static" or some interference is experienced from high frequency transients in power systems. Unless the lines are especially well transposed for these frequencies, the interference requires that the transmission level be kept above a minimum value of the order of $\log_{10} \frac{I_1}{I_2} = 1.2$ (about -25 miles of standard cable below zero level).

(b) While for radio telephone transmission the available data are as yet very meagre, we have obtained a few order-of-magnitude figures which should be of interest. For the Catalina Island radiophone system, for example, the minimum field intensity is estimated at roughly 1,000 microvolts per meter. The circuit is sometimes quite noisy during the summer months altho not prohibitively so. In our ship-to-shore radio telephone experiments along the Atlantic coast, we have on occasions worked with lower field intensities, as low as 100 microvolts per meter. The latter figure, however, gives a grade of service far below wire standards.

(c) The best data on the minimum permissible transmission level for radio telegraphy are those obtained from the experience in trans-Atlantic telegraph operation. The figures prevailing for present trans-oceanic radio-telegraph operation are understood to lie in the order of 10 to 100 microvolts per meter, depending upon individual cases and the time of the year.

The Net Transmission Equivalent

The net over-all transmission equivalent of the system is measured by the ratio of the transmitted to the received signaling power, and is shown in Fig. 4 as the difference between the transmission levels at the two ends. This relatively small loss represents the difference between two large values, the transmission loss and the transmission gain thruout the system. Relatively small changes in either the attenuation or amplification may, therefore, cause large changes in the net equivalent of the circuit, thus tending to give rise to instability in the transmission performance of the circuit.

This problem of fluctuation becomes very serious with the use of very high frequencies, whether transmitted by wires or by radio. Were we to attempt to employ, for example, a million cycles for wire carrier transmission over considerable distances, as has been proposed, not only would the losses be very large, but they would be unstable, changing with weather conditions, so that the maintenance of a

constant volume of transmission would become extremely difficult. Similarly in radio transmission, the fluctuations in the ether attenuation, particularly at short wave lengths where over long distances we experience the well known "swinging" or fading effects, render the maintenance of a satisfactory volume of transmission a difficult problem. As noted above these fluctuations, particularly as between day and night transmission with very high frequencies, may be enormous.

It is of value to the radio engineer to have some idea of the over-all circuit transmission equivalents which are necessary for satisfactory telephone communication. In the wire telephone art, the maximum equivalent between subscribers is ordinarily taken as about 30 miles of standard cable or $\log_{10} \frac{I_1}{I_2} = 1.4$. Under quiet conditions, considerably larger transmission equivalents can be talked over. The long distance toll lines themselves are usually designed for transmission equivalents of 0.5 to 0.75 or 10 to 15 miles of standard cable. These figures will serve as a general guide for the transmission equivalents which radio telephone circuits should provide. Where a radio circuit forms a link in a direct wire circuit as, for example, in the case of Catalina Island, it is desirable to work the radio link as close to a zero equivalent as possible, that is, to give out at the receiving end a volume nearly equal to that fed in at the transmitting end.

Two-Way Operation

When the two one-way radio channels are merged at their two ends into a regular telephone circuit for connection to the wire network, as illustrated in Fig. 5, then there is a limit in the transmission equivalent which can be given over the radio part of the circuit.

This limit will be appreciated by reference to Fig. 5. It is imposed by the tendency of the two one-way channels to form a round-trip circuit by "feeding-back" from one to the other via the voice frequency connecting circuit. If the total amplification around the circuit including the voice-frequency line, exceeds the total losses in the circuit, "singing" will result. Were no line balance provided at the voice frequency terminals, then it would be impossible to operate the circuit at a zero equivalent. By setting up a balancing circuit at each end in the manner illustrated, a transmission loss is, in effect, inserted between the sending and receiving sides of the voice circuit which tends to prevent this sing-around action. Actually, there is a limitation in the degree of balance which can be realized between the

telephone line and the balancing network, especially if the telephone line is to be switched at a nearby central office, and this factor, together with the margin of safety which is required between the operating condition and the singing condition, prevents the radio channels from being operated much better than the zero equivalent. This whole matter of realizing in practice an adequate transmission equivalent, will be appreciated to be an especially difficult problem in the case of marine radio telephony, where the connection is switched from one vessel to another at varying distances.

It should be noted further, with reference to two-way operation, that the difficulty of effecting simultaneous sending and receiving at a station arises primarily from the large attenuation which must be

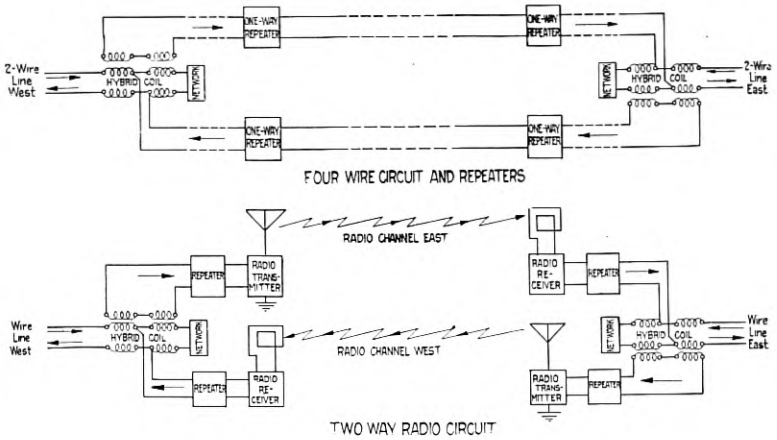


Fig. 5

overcome and the resulting large ratio between the energies transmitted into and received from the ether. The receiver must be prevented from being overloaded by the home transmitter and this, in general, requires that there be provided between the high frequency side of the transmitter and that of the receiver, a transmission loss comparable in size to that obtaining over the radio circuit itself. This "separating" transmission loss is ordinarily provided (a) by frequency-selecting circuits (tuned circuits and filters), the sending and receiving transmissions being placed on different frequencies; (b) by balance, as when using the blind spot of a loop-antenna receiver, and (c) by spatial separation between sending and receiving points, where the large step-off loss is used to advantage.

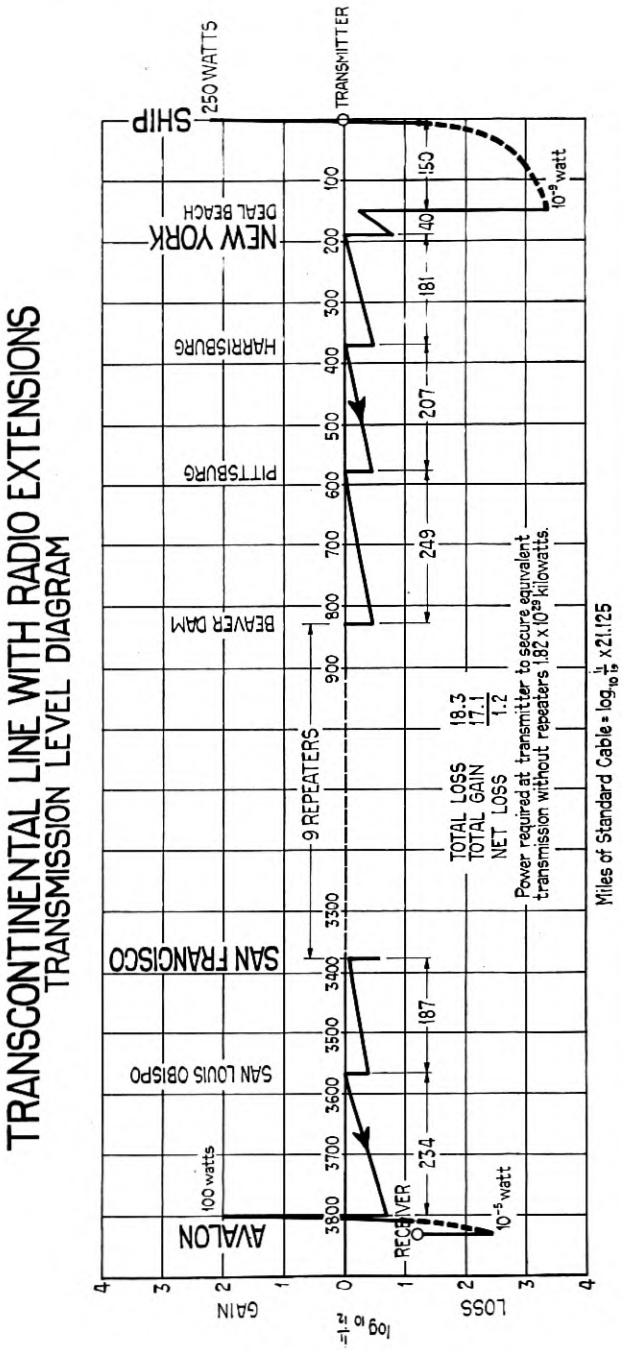

TRANSCONTINENTAL LINE WITH RADIO EXTENSIONS
TRANSMISSION LEVEL DIAGRAM
SHIP
250 WATTS
TRANSMITTER
NEW YORK
DEAL BEACH
HARRISBURG
PITTSBURG
BEAVER DAM
SAN FRANCISCO
SAN LOUIS OBISPO
AVALON
100 watts
RECEIVER
9 REPEATERS
100
200
300
400
500
600
700
800
900
3300
3400
3500
3600
3700
3800
150
40
181
207
249
187
234
10^{-3} watt
10^{-5} watt
TOTAL LOSS 18.3
TOTAL GAIN 17.1
NET LOSS 1.2
Power required at transmitter to secure equivalent transmission without repeaters 1.82×10^{29} kilowatts.
Miles of Standard Cable = $\log_{10} \frac{1}{r} \times 21.125$
GAIN
LOSS
$\log_{10} r =$
4
3
2
1
0
1
2
3
4

Fig. 6

Transmission Levels on Combination Wire and Radio Telephone Systems

It will be of interest to trace thru the approximate transmission levels which obtain for a radio system linked up with a long repeatered land line system.

In Fig. 6, there is taken a rather striking example of this case in the transcontinental telephone line as connected up to radio extensions at its termini—to Catalina Island on the Pacific and to a vessel at sea on the Atlantic. The transmission illustrated is that occurring from east to west. The voice currents start out from the vessel at zero level, are amplified to a relatively high level and upon being transmitted to the shore 150 miles (240 km.) away, drop to a very low level. At the shore radio station they are boosted up, at New York amplified again, and put upon the transcontinental circuit. Regularly at about 300 miles (480 km.) the telephone repeaters pull back the transmission level to about its original value. In the radio link at the western end the currents are again amplified to a high level at the transmitting station, drop down to a very low level at the receiver and are brought back to a level at which they can be heard. Actually in the receiving telephone the transmission is about $\log_{10} \frac{I_1}{I_2} = 1.2$ below zero level, or roughly 25 miles of standard cable "down." The total loss and the total gain in the circuit is enormous, as is shown by the figures given in the diagram. This is a rather striking illustration of the extent to which amplification properly distributed and maintained can be used to overcome attenuations enormous in the aggregate. Just to give a better idea of what these values of attenuation and amplification mean, it may be noted that were it necessary to supply at the transmitting end all of the amplification required for delivering this volume of transmission to the receiving end thru the combination circuit, the kilowatts required would be measured by a twenty-nine place figure, an amount of power unavailable in the world. The importance of correctly distributing the amplification along the system is well illustrated by this figure by comparing it with the signaling power actually represented in the system, which sums up to something less than 1 kilowatt. The difference is simply a question of the transmission level at which the amplification is worked.

Fig. 7 gives a view of the interior of one of these radio telephone stations of the American Telephone and Telegraph Company and Western Electric Company. It is located at Deal Beach, New Jersey.

In the foreground is the switchboard for enabling the operator to control the radio-wire circuit at the connecting point. In the background are the transmitter units—four of them. These, together with the four antennas with which the station is equipped, "multiplex" the ether, in effect, and permit four channels to be established to as many distant stations. It is intended that three of these be telephone talking channels and the fourth a signaling or a reserve talking channel. The receiving station is located at another point. It is not desired to describe this station in any detail but merely to illustrate it as an example of a radio repeating station functioning to

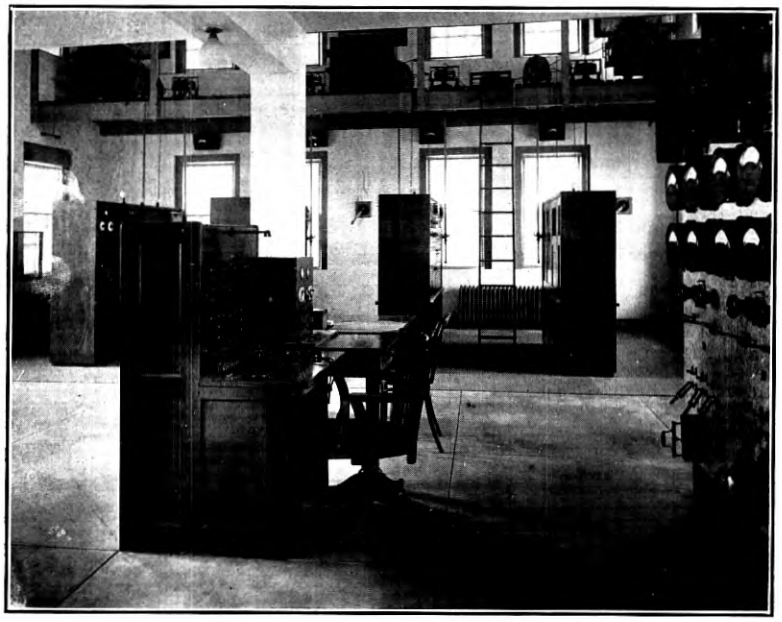


Fig. 7

connect the wire system with ships at sea and capable of effecting simultaneously three different connections. It is hoped that this ship-to-shore development may be itself the subject of an *Institute* paper.

Intermediate Repeaters

The transcontinental line with radio extensions as shown in Fig. 6 is a good illustration of the use of intermediate repeaters generally. Two types of repeaters are represented, the straight wire telephone

repeaters and the shore radio stations which are in effect huge repeaters relaying between the land line and the radio circuits.

Because of the moderate attenuation obtaining in the wire transmission system, we can work with fairly long repeater spacings, about 300 miles in this case, and with moderate amounts of power and yet keep the transmission levels at the receiving end relatively much

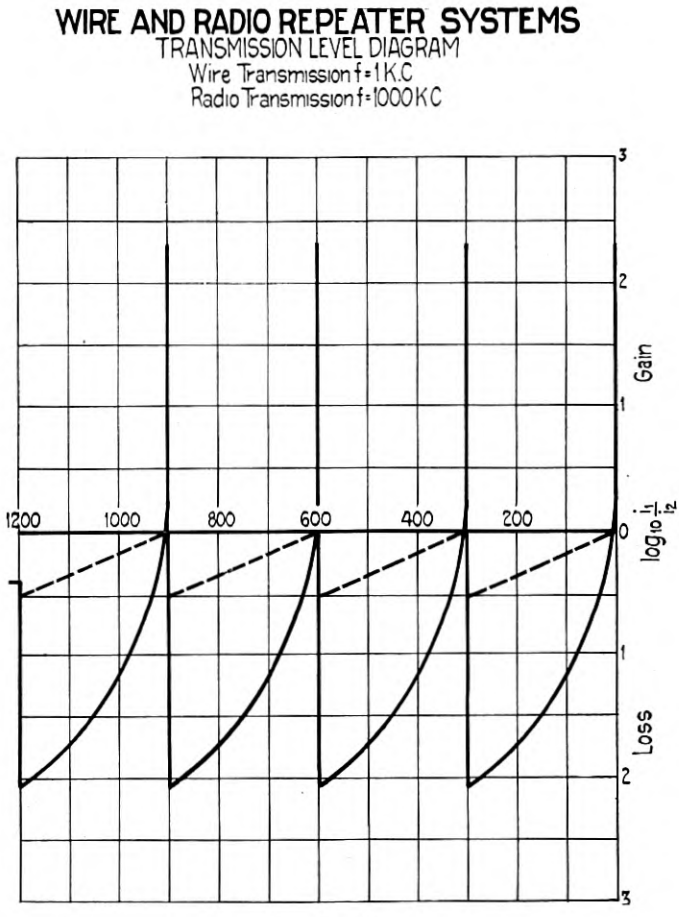


Fig. 8

higher than is usual in radio systems. Due to the large attenuations obtaining over the radio extensions, the radio repeaters must put out a high transmission level, making them costly, and even with this relatively high output, the level drops to very low values at the receiving

end. This falling off occurs largely in the get-away loss at the transmitting end of the radio circuit, as the diagram indicates.

Fig. 8 depicts an all-radio system provided with intermediate repeaters and compares for illustration purposes the transmission levels obtaining therein with those for a wire system. The solid lines are for radio and the dash for wire. It will be seen that the radio system courses thru wide transmission level variations as compared to the ordinary wire system due to the large attenuations obtaining and particularly to the large step-off loss near the sending station.

The figure illustrates the same spacing for both radio and wire repeating and gives a measure of the difference in amplification required in the two cases. Altho in the radio repeaters the level can be permitted to drop to low values, nevertheless a large part of the total amplification has to be supplied at relatively high power levels and it is this fact, together with the antenna structures required at each point to "get into" the ether transmission medium anew, that militates against the economics of radio repeaters as compared with straight-away radio transmission. The tendency will be to "stretch out" the straight-away transmission due to the fact that for the longer distances the transmission loss increases relatively slowly. While we may look for some important uses of radio repeaters in special cases, we should not, in general, expect them to be as important to the radio art as are wire repeaters in wire operation.

Transmission of Side Band Without Carrier

In dealing with the subject of power levels in radio transmission, it is important to recognize that a modulated radio telephone wave consists of two components, one, the carrier frequency itself and the other, the so-called side bands, which are the actual modulated components. This resolution of the modulated carrier into two or, rather, three components, the carrier and two side-bands, has been given mathematically a number of times and need not be repeated. It is physically analogous to the resolution of the unidirectional current of a microphone transmitter into direct current and alternating current components, the direct current corresponding to the carrier and the alternating current to the modulated components.

Now, the important thing about this matter of side bands and the unmodulated carrier component, with reference to transmission considerations, is this, that it is the side bands alone, and not the carrier, which convey the actual intelligence. The function of the

carrier comes in merely at the receiving end, in the detector, as a means for translating the side band from radio frequency back to audio frequency.

This will be made clearer by reference to Fig. 9. At the bottom of the figure is shown schematically a one-way radio system. Above it is depicted the voice-frequency band, showing the manner in which it is shifted by modulation up to the carrier frequency range, and at the receiving end, by detection, back to the voice frequency range. The voice frequency band, as it comes out of the ordinary telephone transmitter, is shown at (b_1) at its normal telephone-frequency posi-

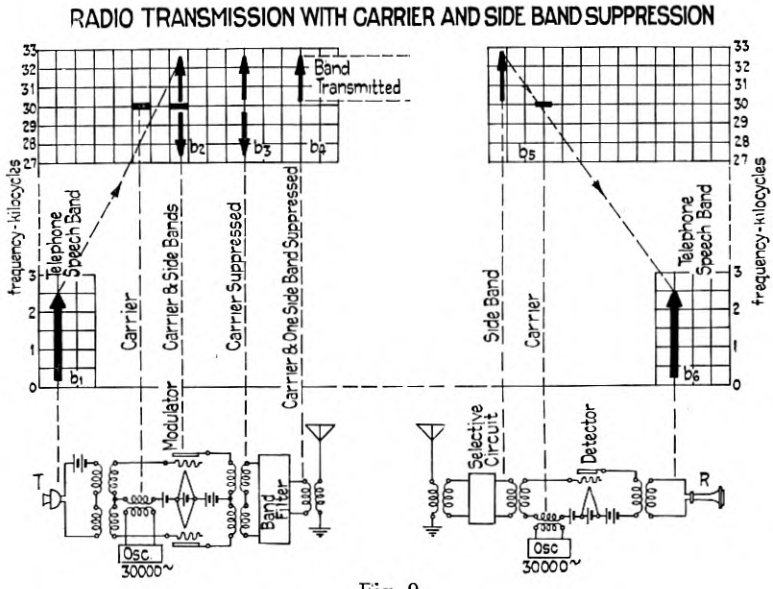


Fig. 9

tion. Upon modulation with the carrier the reference point of the voice frequency band is shifted from zero frequency (direct current) up to the carrier frequency as shown at b_2 where the two side bands appear. The effect of modulation is, therefore, simply to shift the band of signaling frequencies upward in the frequency range and refer it in a double relation to the carrier frequency.

Located between the upper and lower side band in the figure, there is indicated the unmodulated component of the carrier. The fact that this component is unnecessary so far as the actual intelligence-carrying energy is concerned, is proven by the fact that it need not be transmitted to the receiver. The carrier may be sup-

pressed as shown at b_3. A means for doing this is the Carson balanced tube modulator circuit illustrated below in the figure.

A reduction of the total band can be effected by filtering out one of them as shown at b_4. The remaining single band is the simplest component of the modulation process with which intelligence can be transmitted to the distant end. Upon arriving at the receiving end at b_5, this side band is fed into the detector along with a carrier of the same frequency as that employed at the sending end; these two components demodulate one another, with the result that the side band is shifted down to its original audio-frequency position in the scale, as indicated at b_6.

Actually this general method of transmission, involving both carrier suppression and side band elimination, is being employed in wire carrier systems in the Bell Telephone Plant.[6] It is briefly explained here because it represents a valuable improvement in wire transmission which should have important application in radio.

From the standpoint of transmission levels its application is in showing that the real intelligence-carrying component of a radio wave is the side-band and not the carrier itself. In considering transmission levels accurately care should be taken, therefore, to deal in terms of the level or wave-intensity of the side-band component and not the carrier. It is because of this that as nearly complete modulation as possible is desired at the transmitting station.

It follows that the power resident in the carrier is a pure waste in so far as overcoming interference is concerned. An important power saving can be effected in the transmitting station by providing some such means as is illustrated whereby the carrier power is held back in the circuit. The two side-bands together can never be greater in current and voltage value than the carrier, and each side-band alone cannot be greater than half the carrier. The power of the carrier is therefore always at least four times the power of one side-band or twice that of both together. Thus by "holding-back" the carrier at the transmitter we can transmit with but one-third the power ordinarily required. Actually the power saving is much greater than this because of the necessity of normally working with larger ratios between carrier and side-band in order to accommodate the peaks of the telephone waves and thereby preserve the quality of transmission. The power saving is, of course, especially important in long distance work.

The suppression of one of the two side-bands halves the frequency

[6] "Carrier Current Telephony and Telegraphy," by Colpitts and Blackwell, *Journal of the American Institute of Electrical Engineers*, February 16–18, 1921.

band required for transmission and would double the message-carrying capacity of the ether were no frequency range required to space the channels apart. This advantage of the present method is likewise of special importance in long-distance-long-wave transmission.

Quality of Transmission

We have spoken above of factors concerned with the volume of transmission and only incidentally of that other requisite of transmission, namely, good quality. Without going into this matter in much detail, it will be well to make note of the several factors involved in obtaining good quality, as follows:

1. It is important that a substantial *band* of voice frequencies be transmitted. Of course, distorted talk can be transmitted on a relatively narrow band, but commercial transmission has been found to require a single side-band width of the order of 2,000 or more cycles, the band width increasing with the quality desired, up to about 5,000 cycles.

2. It is necessary that the distortion which is due to non-linearity of transmission with respect to amplitude, be avoided. This is equivalent to saying that there should not be permitted to take place self-modulation between the components of the side-band, nor the too close cutting-off of the peaks of the telephone waves due to saturation effects.

3. The transmission must be kept free from interfering noises. The ratio between interfering noise current and voice currents of the order of 0.1 is regarded as large in wire practice. While this amount of interfering current will not prohibit service, it does seriously impair the effectiveness of transmission and annoys the listener. In radio the ratio of static noise to signal strength is very often much greater than this value. As the radio art progresses it will be necessary to work toward standards more nearly in keeping with those which have been found necessary for wire service.

The writer wishes to express his indebtedness to the following of his associates for helpful suggestions and assistance—Messrs. J. R. Carson, Ralph Bown, and D. K. Martin.

APPENDIX

The curves of Figs. 2 and 3 are based upon the following equations and data:

The radio curves are based on the familiar Austin-Cohen formula:

$$I = \frac{7.8 \times 10^{-10}\, h_r\, h_s\, f\, I_s}{R\, d}\, \epsilon^{-4.4 \times 10^{-6} d\sqrt{f}} \tag{1}$$

where I = amperes
R = ohms
h = meters
f = cycles
d = miles

Taking equal antenna heights at two ends $h_s = h_r$.

As regards antenna resistance we assume symmetry as between the two ends, and that the external (radiation) resistance of the antenna equals the resistance within the antenna (which resistance would be internal apparatus resistance in the case of a perfect antenna). This makes R_r (radiation resistance) $= R_i$ (ohmic resistance); and R of (1) becomes $= R_r + R_i$ where:

$$R_r = 17.8 \times 10^{-15}\, h^2 f^2. \tag{2}$$

Expressing in terms of current ratio and substituting values of R, equation (1) becomes,

$$\frac{I_s}{I_r} = 45.5 \times 10^{-6} f\, d.\quad \epsilon^{4.4 \times 10^{-6} d\sqrt{f}}. \tag{3}$$

In order to plot this equation on the same basis as we usually plot wire attenuation, the logarithm of the ratio is used, thus;

$$\log_{10} \frac{I_s}{I_r} = \log_{10} 45.5 \times 10^{-6} f\, d + \frac{4.4 \times 10^{-6}}{2.303}\, d\sqrt{f} \tag{4}$$

which is the equation of the curves plotted.

The ratio of the currents in the two antennas is in this case a true measure of the transmission because they are in circuits of equal impedances, by the assumption of antenna symmetry.

Data for the Wire Curves

$$\alpha = \frac{R}{2}\sqrt{\frac{C}{L}} + \frac{G}{2}\sqrt{\frac{L}{C}}.$$

For number 8 Birmingham wire gauge open wire (diam. = 0.165 in. = 4.19mm. wire spacing = 12 in. = 30.5 cm. 40 poles per mile) dry weather, the constants per mile are;

$$L = 3{,}370\ \mu h.$$
$$C = 9{,}140\ \mu\mu f.$$

	Frequency, Kilocycles				
	1	20	100	1000	
R =	0.14	10.0	21.5	65.7	ohms per loop mile
G =	0.55	10.0	*50.0	*500.0	μ ohms per loop mile
α =	0.003488	0.0112	0.03289	0.2059	

* Estimated.

A Low Voltage Cathode Ray Oscillograph[1]

By J. B. JOHNSON

SYNOPSIS: A sensitive cathode ray oscillograph is described which operates at the low potential of from 300 to 400 volts. The electron stream comes from a thermionic cathode and is focused by the action of ionized gas in the tube. This gas, at a pressure of a few thousandths of a millimeter, serves to reduce to 1mm. diameter a spot which would be 1 cm. across in a high vacuum tube. The sensitivity of the tube is such that the deflection of the spot is about 1 mm. per volt applied between deflector plates. When using magnetic deflection, a pair of coils 4 cm. in diameter, placed on the sides of the tube produces a deflection of approximately 1 mm. per ampere-turn of the coils.—*Editor.*

A CATHODE ray oscillograph tube operating at a comparatively low voltage was described by the writer some time ago before the American Physical Society.[2] Since then, the tube has been further improved and its operation studied so that now both the structure of the tube and the principles which have made the construction possible can be described in greater detail.

In the older types of Braun tubes the electron stream is produced by a high voltage discharge through the residual gas in the tube. This requires a source of steady potential of from 10,000 to 50,000 volts, an installation which is expensive, non-portable, and dangerous. In the new type of tube the low voltage operation has been obtained by the use of a Wehnelt cathode as the source of electrons, so that the lower limit of voltage is set by the effect of the electrons on the fluorescent screen and not by the voltage needed to obtain the electrons. At 300 volts the electrons produce quite bright fluorescence on the screen and the tubes are therefore designed to operate at 300 to 400 volts.

The external appearance of the tube is shown in Fig. 1. The electrodes are located at one end of the pear-shaped bulb, and the fluorescent material is deposited on the inside of the larger, flattened end. The tube is provided with a base which fits into a bayonet socket such as is used for vacuum tubes, and all the connections are made through the base. There are two orthogonal pairs of deflector plates inside the tube for electrostatic deflection, while magnetic deflection is produced by applying a field from the outside.

The internal structure differs considerably from that of previous forms of Braun tube and it will therefore be described somewhat fully.

[1] Also published in the *Journal of the Optical Society of America and Review of Scientific Instruments*, September, 1922.

[2] *Phys. Rev.* (2), Vol. 17, p. 420, 1921.

The Focusing

In some forms of Braun tube a sharp spot has been secured by using a very high voltage, and therefore high electron velocity, so that after the electrons have passed through one or two fine apertures to make the beam parallel there is not time enough for the mutual repulsion to spread the beam again appreciably before the electrons strike the screen. With other tubes an external "striction" coil has been used which maintains a strong longitudinal magnetic field in the region between the anode and the cathode and which brings the electrons to a focus on the screen. In the low voltage tube the spreading of the electron stream is greater than in high voltage tubes because of the greater time during which the mutual repulsion of the electrons

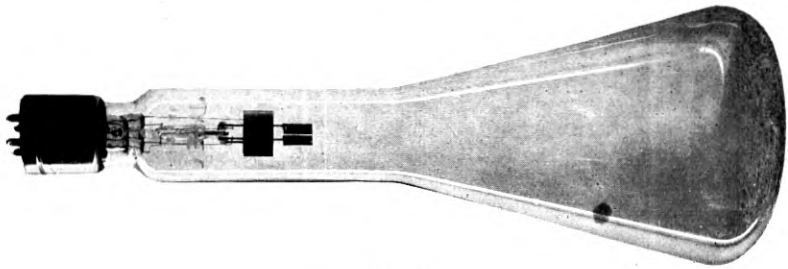


Fig. 1

acts, so that some means of focusing must be used. The electrons can be brought to a focus by a longitudinal magnetic field so adjusted that each divergent electron makes very nearly one complete turn of a spiral and in travelling the length of the tube returns to the axis at the screen. In this way a very sharp spot can be produced, but the sensitivity of the beam to deflection is reduced very much by the directing magnetic field.

The method of focusing that is used in the present tubes grew out of the suggestion by Dr. H. J. van der Bijl, that a small amount of gas be introduced into the tube. This gas, at a pressure of a few thousandths of a millimeter of mercury, serves to reduce to 1 mm. diameter a spot which would be 1 cm. across in a high vacuum tube. The sharpness of the spot depends also upon the current in the electron stream so that the focus may be controlled by the cathode temperature. The mechanism of this focusing action will be explained later.

The presence of this slightly ionized gas also serves the purpose of preventing the accumulation of charges on the glass, and it provides

for the discharging of the fluorescent screen so that the electrons can drift back to the metallic circuit.

The Electrode Unit

With gas present in the tube, steps have to be taken to guard against arcing and the injurious effects of positive ion bombardment on the cathode. This is done by making the volume of gas surrounding the electrodes very small. For this purpose the cathode and anode, themselves small, are sealed into a short and narrow glass tube so that the volume exposed to both electrodes in common is less than 1 cu. cm. All paths between the electrodes are then so short that at this low pressure there is not sufficient ionization to build up an arc.

The structure of this unit, or "electron gun" is shown in Fig. 2. The cathode, *f*, is an oxide coated platinum ribbon of the same kind

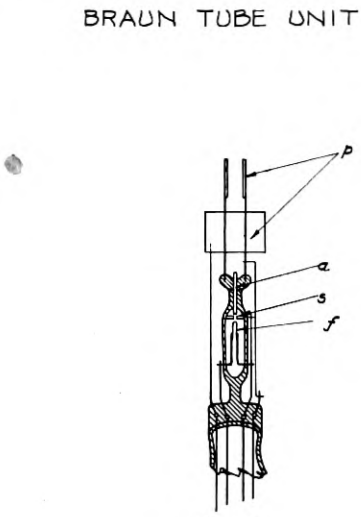


Fig. 2

as the filament in our audion tubes. The anode, *a*, is a thin platinum tube 1 cm. long and 1 mm. in diameter, one end of which is about 1 mm. from the top of the filament loop, the other end opening into the main tube towards the fluorescent screen. Between the cathode and the anode and connected to the cathode is a metal shield, *s*, with a small aperture through which the electrons must pass in going to the anode. Nearly all of the electrons must then go to the inside of the tubular anode, and a small fraction of them pass through the whole

length of the anode and form the beam in the main part of the tube.

The deflector plates, *p*, are also mounted rigidly on this unit. In order to avoid large differences of potential in the tube, one plate from each pair is permanently connected to the anode, the variable potentials being applied to the other plates. The complete unit is mounted at the small end of the tube with the anode and deflector plates toward the fluorescent screen.

The Filament

In some early forms the filament was bent into a simple hair pin loop which was placed close to the aperture in the shield. It was then found that the positive ions striking the filament from the direction of the anode soon destroyed the oxide coating and left the filament inactive. This trouble was largely overcome by placing the filament out of the direct path of the positive ions. The flat filament is now shaped into a ring as shown in Fig. 3, slightly larger in diameter than

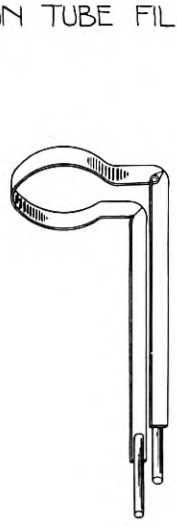


Fig. 3

the aperture in the shield and is placed coaxial with the anode. The momentum of the positive ions then carries them past the active part of the filament and they strike where little damage can be done. The length of service of the tube is still limited by the filament life, but this has been increased by the above artifice so that the tube now gives around 200 hours of actual operation.

The Deflector Elements

The deflector plates are made of German silver, which is non-magnetic and which has a high specific resistance that diminishes the effect of eddy currents when magnetic deflection is used. The plates are 13.7 mm. long in the direction of the tube axis and the separation between them is 4.7 mm.

The sensitivity of the tube is such that the deflection of the spot is about one mm. per volt applied between the deflector plates. When

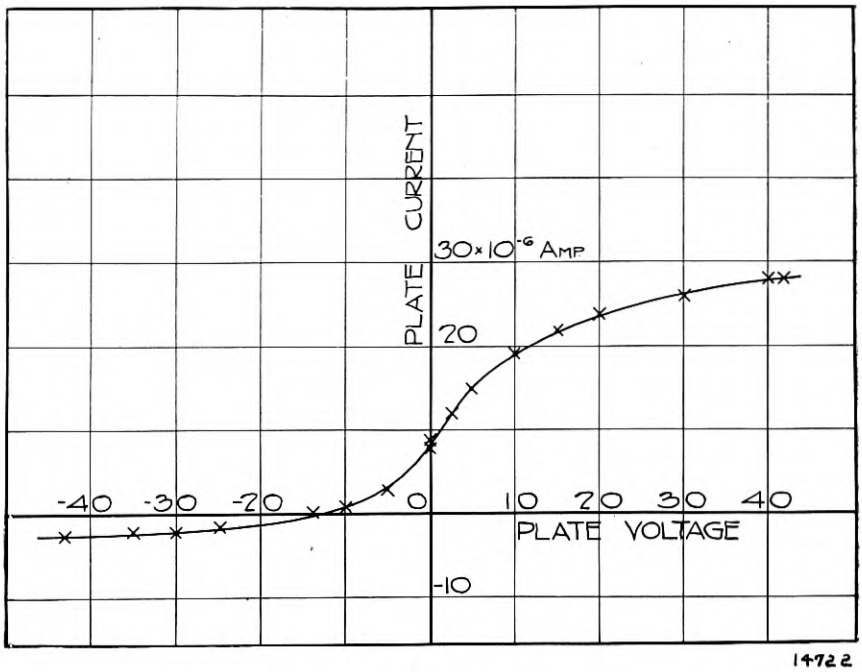


Fig. 4

using magnetic deflection, a pair of coils 4 cm. in diameter placed on the sides of the tube at the level of the deflector plates produces a deflection of approximately 1 mm. per ampere-turn flowing in the coils.

The electrons striking the screen drift back to the anode structure, and most of them are collected by the deflector plates. There is also a small ionization current flowing to the plates. The tube is therefore not strictly an electrostatic device, and this must be kept in mind when using it. Fig. 4 shows the current flowing to the two free plates at various voltages with respect to the anode. With the large posi-

tive values of plate voltage the current to the plates is practically equal to the current in the electron stream and consists largely of the returning electrons. The small current in the other direction when the plate voltage is negative is a measure of the ionization in the tube.

The Fluorescent Screen

The screen is spread on the inner surface of the large end of the tube, using pure water glass for binder. The active material consists of equal parts of calcium tungstate and zinc silicate, both specially prepared for fluorescence. This mixture produces a generally more useful screen than either constituent alone. The pure tungstate gives a deep blue light which is about 30 times as active on the photographic plate as the yellow-green light of the silicate, while the silicate gives a light which is many times brighter visually than that from the tungstate. By mixing the two materials in equal parts a screen is produced which is more than half as bright visually as pure zinc silicate and more than half as active photographically as pure calcium tungstate.

For mechanical strength the end of the bulb which carries the screen is rounded outwards so that the screen is not a plane surface. This introduces a distortion of the fluorescent pattern which in most instances is negligible. If the pattern is recorded by a camera whose lens is D cm. from the end of the tube, then the apparent reduction of the deflection produced by the curvature of the bulb is given in terms of the deflection y approximately by

$$\Delta y = \frac{20 + D}{400\, D}\, y^3 \ cm.$$

The Function of the Gas

The part which the gas plays in focusing the beam of electrons is an interesting phenomenon which depends upon the difference in the mobilities of electrons and positive ions. The electrons of the beam are pulled toward the common axis by a radial electric field produced by an excess of positive electricity in the electron stream and an excess of negative electricity in the space outside the beam. This distribution is produced as follows: Some of the electrons of the stream, in passing through the gas, collide with gas molecules and ionize them. Both the colliding electrons and the secondary electrons leave the beam but the heavy positive ions receive very little velocity from the impact and drift out of the beam with only their comparatively low thermal velocity. Positive ions, therefore,

accumulate down the length of the stream and may exceed in number the negative charges passing along. At the same time, electrons are moving at random outside the stream, producing negative electrification. There is then a field surrounding the stream which tends to pull the electrons inward. If there were only the mutual repulsion between the electrons to compensate for, this would be done when the number of positive ions in the beam equals the number of electrons. There is in addition an original divergence of the beam which must be overcome. If this divergence is assumed to be one degree from the axis and the electron current 2×10^{-5} amp., then a simple calculation shows that the radial field required to pull the beam to a focus at the usual distance is about one volt per cm. This field strength is produced, with beams of the ordinary intensity, if there are four positive ions for each electron in the stream, a condition which seems not unreasonable.

The number of ions per electron in the stream is probably constant as the current in the stream is varied, since the conditions of collision and recombination are not altered. When the current is increased, therefore, the total positive ionization of the beam increases, the field around the beam becomes stronger, and the electrons are brought to a focus in a shorter distance.

These deductions have been confirmed experimentally. That the focusing of the stream depends upon the current flowing was one of the earliest observations made in developing the tube and this method has been used ever since to obtain a sharp spot. The point of convergence can be seen moving in the manner expected when the current is changed, and the effect has been further verified by using a tube with a movable fluorescent screen so that the length of the electron beam could be varied. The presence of the electric field around the beam was shown by the effect of two beams on each other, in a tube in which there were two electron streams crossing each other at right angles at their mid-points, each falling on a fluorescent screen. When one beam was moved away from the other by a field between the deflector plates, the second beam moved as if attracted by the first. The directed electrons in each beam were attracted toward the positive ionization in the other, and for one particular adjustment of the tube the displacement was such as would have been caused by a field of about 3 volts per cm., a result not far different from that previously calculated.

Since the beam must produce its own positive ionization some time must elapse before it can produce by collisions the required number of positive ions. Calculation shows this time to be of the order of

10^{-6} second. When the beam moves it has to build up the ionization as it goes along, and we should expect that when deflected very rapidly it might no longer be focused, due to lack of positive ions in its path. A test was made of this by applying a high frequency potential on the deflector plates so that the spot described an elliptic pattern. At a frequency of 10^5 cycles per second the line was still sharp, but at 10^6 cycles there was a noticeable widening of the line which is probably to be ascribed to imperfect focusing at this high speed.

In these experiments the evidence all points to the view that the focusing of the electrons is caused by an excess of positive charge in the beam itself, produced by ionizing collisions of the electrons with the gas molecules. Further confirmation is found in the fact that a focus is much more readily obtained in the heavier gases having slow molecules, such as nitrogen, argon or mercury vapor, than in hydrogen and helium where the mean velocity of the molecules is greater. The tubes are therefore filled with argon, the heaviest available permanent gas which does not attack the electrodes. The best pressure for the length of tube adopted and for the current which can be obtained in the beam is 5 to 10 microns, and this leaves considerable latitude for the adjustment of the electron current to get a sharp focus.

Examples of the Use of the Tube

Because of the small amount of auxiliary apparatus required with this form of Braun tube it has proved to be a very convenient laboratory instrument. It has found application in studying the behavior of vacuum tubes and amplifier and oscillator circuits, of gas discharge tubes, of relays, and of numerous other kinds of apparatus, both at low and at high frequencies. Some reproductions of photographs of various types of curves are given below to illustrate the kind of results which are possible with this oscillograph.

Fig. 5 shows the hysteresis curve of a sample of iron wire. The wire was placed in a small solenoid with one end toward the side of the tube. The magnetizing current passed through a resistance, the voltage drop of which was applied to one pair of deflector plates so as to give a deflection proportional to the magnetizing field. The stray magnetic field from the iron itself produced the deflection proportional to the induction. Alternating current was used, and the exposure was 20 seconds with lens opening f 6.3 and speed roll film.

In Figs. 6a and 6b are shown the current-voltage relations of an

oscillating vacuum tube. The axes were obtained by grounding one or the other deflector element.

The measurement of modulation in a radio transmitting set has

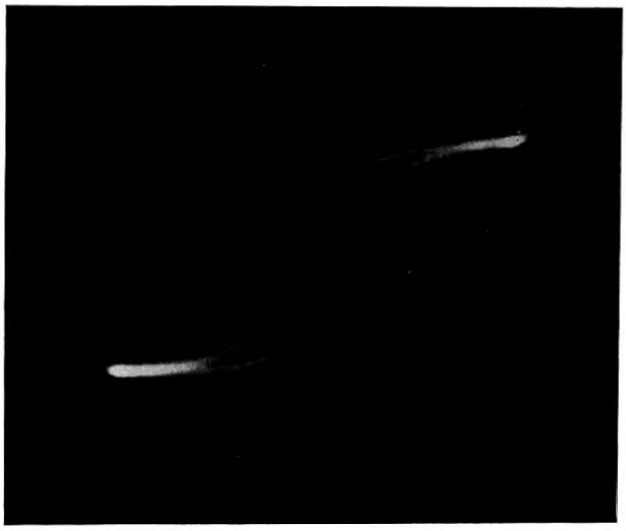


Fig. 5

been reduced to a fairly simple process by means of the cathode ray tube. The low frequency modulating voltage, controlled by the

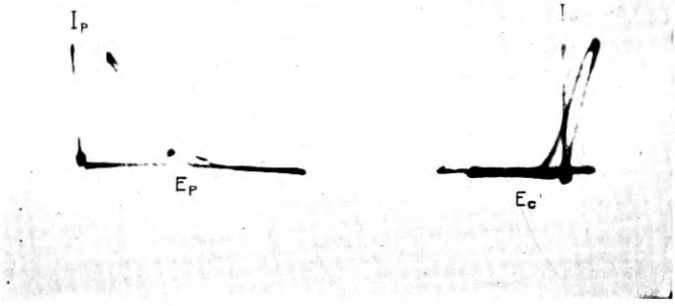


Fig. 6

voice, is applied to one pair of deflector plates, while the radio frequency output, with amplitude varying according to the low frequency voltage, is applied to the other pair of deflector plates. The resulting pattern on the screen is a quadrilateral of solid fluorescence, since the

Fig. 7

two frequencies are not commensurate. The two vertical sides indicate the greatest and the least amplitude of the high frequency, while the other two sides show the current-voltage characteristic of the transmitter. Fig. 7 shows such a pattern (retouched), the edges being much brighter than the centre. The exposure was two minutes using a Seed 23 plate and *f* 6.8 lens opening.

The Contributors to this Issue

GEORGE A. CAMPBELL, B.S., Massachusetts Institute of Technology, 1891; A.B., Harvard, 1892; Ph.D., 1901; Göttingen, Vienna and Paris, 1893–96. Mechanical Department, American Bell Telephone Company, 1897; Engineering Department, American Telephone and Telegraph Company, 1903–1919; Department of Development and Research, 1919—; Research Engineer, 1908—. Dr. Campbell has published papers on loading and the theory of electric circuits and is also well-known to telephone engineers for his contributions to repeater and substation circuits. The electric filter which is one of his inventions plays a fundamental rôle in telephone repeater, carrier current and radio systems.

RALPH V. L. HARTLEY, A.B., Utah, 1909; B.A., Oxford, 1912; B.Sc., 1913; instructor in physics, Nevada, 1909–10; Engineering Department, Western Electric Company, 1913—. For some time Mr. Hartley has been closely connected with the development of carrier current, telephone repeater, and telegraph systems.

THORNTON C. FRY, A.B., Findlay, 1912; A.M., University of Wisconsin, 1913; Ph.D., 1920; instructor of mathematics, Wisconsin, 1912–16; Engineering Department, Western Electric Company, 1916—. Mr. Fry has written several papers on the theory of electric circuits and other subjects allied to telephony.

JOHN R. CARSON, B.S., Princeton, 1907; E.E., 1909; M.S., 1912; Research Department, Westinghouse Electric and Manufacturing Company; 1910–12; instructor of physics and electrical engineering, Princeton, 1912–14; American Telephone and Telegraph Company, Engineering Department, 1914–15; Patent Department, 1916–17; Engineering Department, 1918; Department of Development and Research, 1919—. Mr. Carson's work has been along theoretical lines and he has published several papers on theory of electric circuits and electric wave propagation.

R. L. WEGEL, A.B., Ripon College, 1910; assistant in physics, University of Wisconsin, 1910–12; physicist with T. A. Edison, 1912–13; Engineering Department of Western Electric Company, 1914—. Mr. Wegel has been closely associated with the development of telephone transmitters and receivers, and has made important contributions to the theory of receivers.

EDWARD C. MOLINA, Engineering Department of the American Telephone and Telegraph Company, 1901–19, as engineering assistant; transferred to the Circuits Design Department to work on machine switching systems, 1905; Department of Development and Research, 1919—. Mr. Molina has been closely associated with the application of the mathematical theory of probabilities to trunking problems and has taken out several important patents relating to machine switching.

WILLIAM C. HELMLE, B. S., University of Wisconsin, 1917; University of Chicago, 1919-20; Commercial Engineer's Office, American Telephone and Telegraph Company 1920—.

E. T. HOCH, B.S., in Electrical Engineering, Case School of Applied Science, 1914; Western Electric Company, Manufacturing and Installation Departments, 1914–15; Engineering Department, 1915—.

LLOYD ESPENSCHIED, Pratt Institute, 1909; United Wireless Telegraph Company as radio operator, summers, 1907–08; Telefunken Wireless Telegraph Company of America assistant engineer, 1909–10; American Telephone and Telegraph Company, Engineering Department and Department of Development and Research, 1910—. Took part in long distance radio telephone experiments from Washington to Hawaii and Paris, 1915; since then his work has been connected with the development of radio and carrier systems.

J. B. JOHNSON, B.S., University of North Dakota, 1913; M.S., 1914; Ph.D., Yale, 1917; Engineering Department, Western Electric Company, 1917—. Since coming to the Western Electric Company, Mr. Johnson has devoted much time to high vacua and ionization in gases.